WRITE AND PAINT: GENERATIVE VISION-LANGUAGE MODELS ARE UNIFIED MODAL LEARNERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in vision-language pre-training have pushed the state-of-the-art on various vision-language tasks, making machines more capable of multi-modal writing (image-to-text generation) and painting (text-to-image generation). However, few studies investigate if these two essential capabilities can be learned together and boost each other, making a versatile and powerful multi-modal foundation model. In this work, we disclose the potential of symmetric generative vision-language pre-training in learning to write and paint concurrently, and propose a new unified modal model, named DAVINCI, trained with prefix language modeling and prefix image modeling, a simple generative self-supervised objective on image-text pairs. Thanks to the proposed prefix multi-modal modeling framework, DAVINCI is simple to train, scalable to huge data, adaptable to both writing and painting tasks, and also strong on other vision, text, and multi-modal understanding tasks. DAVINCI achieves competitive performance on a wide range of 27 generation/understanding tasks and demonstrates the superiority of combining vision/language generative pre-training. Furthermore, we carefully benchmark the performance of different vision-language pre-training objectives on different scales of pre-training datasets on a heterogeneous and broad distribution coverage. Our results demonstrate the potential of exploiting self-supervision in both language and vision inputs, and establish new, stronger baselines for future comparisons at different data scales.[1]

1 INTRODUCTION

Self-supervised language model pre-training (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Brown et al., 2020) has reshaped the landscape of modern natural language processing (NLP) research, pushing the state-of-the-art of a wide range of NLP tasks. Recently, this success has been transferred to the multi-modal context and resulted in a number of vision-language pre-trained models (VLMs) (Lu et al., 2019; Tan & Bansal, 2019a), achieving state-of-the-art results on various vision-language tasks. Most existing VLMs are BERT-like Transformer (Vaswani et al., 2017) encoders pre-trained with a combination of different vision-language pre-training (VLP) objectives: masked multi-modal modeling (Lu et al., 2019; Tan & Bansal, 2019b; Chen et al., 2020; Li et al., 2020), multi-modal alignment prediction (Lu et al., 2019; Tan & Bansal, 2019b; Chen et al., 2020; Li et al., 2020), region of interest feature regression (Tan & Bansal, 2019b), image-text matching (Li et al., 2021; Zeng et al., 2021), to name a few. However, the roadmap towards large language models reveals a transition pattern from encoder-only models like BERT (Devlin et al., 2019) / RoBERTa (Liu et al., 2019) to sequence-to-sequence models like T5 (Raffel et al., 2020) / BART (Lewis et al., 2020) and autoregressive models like GPT-3 (Brown et al., 2020) / PaLM (Chowdhery et al., 2022) to tackle more tasks in a unified way, and from complicated objectives like masked language modeling / next sentence prediction / replace token detection to a simple language modeling objective to improve the scalability of pre-training. This suggests that the generative pre-training paradigm with simple targets shows great potential for pre-training more scalable and general VLMs.

To this end, several recent studies (Cho et al., 2021; Zhang et al., 2021a; Wang et al., 2021b; 2022) investigated sequence-to-sequence (seq2seq) vision-language pre-training and achieved state-of-the-art results on a range of vision-language understanding and generation tasks. For example, VL-T5 (Cho et al., 2021) and OFA (Wang et al., 2022) formulate various vision-and-language

[1]The code and pre-trained models will be released in the final version.

problems into seq2seq tasks and pre-train a seq2seq VLM by multi-tasking on these tasks. This approach is hard to scale because it is non-trivial to collect a large number of vision-language datasets for pre-training. In addition, ERNIE-ViLG (Zhang et al., 2021a) and SimVLM (Wang et al., 2021b) pre-train seq2seq VLMs with a simple language modeling or prefix language modeling objective on a large number of image-caption pairs. While achieving promising results, these objectives are not versatile enough, resulting in VLMs that are only capable of a subset of tasks in image-text modalities. On the other hand, the recent success of generative language pre-training (Brown et al., 2020) and generative vision pre-training (He et al., 2022; Bao et al., 2021) motivate us to explore generative vision-language pre-training to learn more versatile and scalable vision-language models.

In this work, we introduce prefix multi-modal modeling, a unified generative pre-training framework that extends prefix language modeling to the multi-modal context and learns a multi-modal foundation model by learning to write and paint simultaneously. As illustrated in Figure 1, given an image-caption pair, we split the image and caption into two parts denoted as prefix and suffix. To make prefix image modeling compatible with the seq2seq formulation of conventional prefix language modeling, we follow DALLE (Ramesh et al., 2021) and convert images into discrete sequences of image tokens (van den Oord et al., 2017). We then train the model to generate the suffix in one modality based on the prefix in the same modality and the complete input in the other modality. In this way, prefix multi-modal modeling can fully exploit self-supervision from image-caption pairs by learning to write and paint simultaneously and is easily scalable for large-scale pre-training. We pre-train DaVinci [2], a vision-language foundation model, with the proposed prefix multi-modal modeling framework on large-scale image-text pairs. DaVinci is the first self-supervised vision-language foundation model that is versatile for all kinds of tasks in vision-and-language modalities, including image-to-text generation, text-to-image generation, vision-language understanding, and single-modal language / vision tasks. DaVinci consistently outperforms FLAVA (Singh et al., 2021), an existing vision-language foundation model, on both language, vision, and multi-modal tasks, and performs competitively with state-of-the-art models across a wide range of tasks and modalities. Moreover, DaVinci also shows strong few-shot and zero-shot image/text generation capability.

In addition, most existing VLMs are pre-trained with mixed pre-training objectives and different data sources varying in size, making it difficult to disentangle the impact of pre-training objectives and data sources on the downstream tasks. To this end, we conduct a systematic analysis of the performance of generative vision-language pre-training by carefully ablating different pre-training objectives, such as prefix language modeling and prefix image modeling, and the amount of pre-training data with different qualities, revealing the impact of different objectives and data sources to facilitating future research.

To summarize, our contribution is three-fold: (1) We introduce prefix multi-modal modeling, a simple unified generative vision-language pre-training framework that is scalable for large-scale pre-training and versatile for image-to-text generation, text-to-image generation and various multi-modal / single-modal understanding tasks. (2) We pre-train DaVinci, a vision-language foundation model, with the proposed approach, demonstrating competitive performance on a wide range of 27 downstream tasks and the superiority of combining vision/language generative pre-training. (3) We conduct an analysis about the impact of different pre-training data sources and pre-training objectives on the performance of seq2seq VLMs.

2 Related Work

Inspired by the success of language model pre-training, several studies investigated vision-language pre-training on large-scale image-caption pairs. ViLBERT (Lu et al., 2019) and LXMERT (Tan & Bansal, 2019b) first propose to extract visual object features with an external object detection model like Fast-RCNN (Girshick, 2015), feed the image features together with texts into Transformer models, and train the model to align vision and language representations with masked multi-modal modeling and multi-modal alignment prediction objectives. Many following works (Li et al., 2020; Zhang et al., 2021b; Chen et al., 2020; Li et al., 2022a; 2021; Zeng et al., 2021; Wang et al., 2021a) propose several new objectives to improve object detection based VLP and explored using vision Transformer (Dosovitskiy et al., 2021; Touvron et al., 2021) as visual feature extractor.

[2] Named after the Italian polymath *Leonardo da Vinci*, who displayed infinite grace in everything. We noticed that this name is used in GPT-3 versioning. However, we think there is no conflict because it is only a suffix for a specific checkpoint of the GPT-3 family.

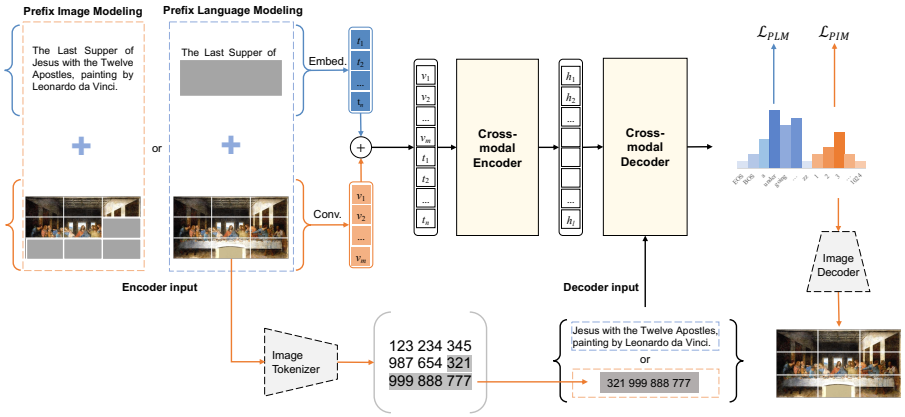

Figure 1: Illustration of the overall architecture and pre-training procedures of DAVINCI, a Transformer-based sequence-to-sequence model. Given an image-text pair, DAVINCI first splits either the word sequence or image token sequence into prefix and suffix. It then concatenates the prefix with the complete sequence in the other modality as input. DAVINCI is trained to recover the suffix with maximum likelihood estimation.

More recently, FLAVA (Singh et al., 2021), a new vision-language foundation model, is pre-trained with a masked multi-modal modeling objective. Performing competitively on language, vision, and vision-language understanding tasks, FLAVA is designed for understanding tasks without text and image generation abilities.

While achieving promising results on **multi-modal understanding** tasks, most VLMs are based on encoder-only architectures with bidirectional attention, making them non-trivial to adapt to **multi-modal generation** tasks such as image captioning and text-to-image generation. Inspired by the success of seq2seq pre-trained language models such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), VL-T5 (Cho et al., 2021) and OFA (Wang et al., 2022) propose to formulate both vision-language pre-training objectives and various downstream vision-language tasks as seq2seq tasks and pre-train a seq2seq VLM by multi-tasking on these tasks. However, the scalability and the zero-shot transfer capability of this approach are limited by the availability of large-scale and diverse vision-language tasks. To this end, SimVLM (Wang et al., 2021b), the most related work to our approach, instead pre-trains a seq2seq VLM with a simple prefix language modeling objective on text generation. It easily scales to very large and potentially noisy pre-training data and achieves competitive results. However, SimVLM only exploits language self-supervision, and thus it does not perform well on image understanding tasks and is unable to tackle image generation tasks.

As for the text-to-image generation task, Ramesh et al. (2021); Ding et al. (2021) achieved promising performance by learning an auto-regressive target with Transformer and VQ-VAE tokenizer. Most recently, Ramesh et al. (2022); Saharia et al. (2022) advanced the image generation capability by using diffusion models and high-quality text embeddings (e.g., CLIP, T5). Therefore, it is natural to explore boosting image generation via stronger multi-modal understanding.

Previous studies are good at either image-to-text or text-to-image generation, but few studies investigate whether these two important capabilities can be learned together and boost each other. In this paper, we explore making a versatile and powerful multi-modal foundation model that is good at multi-modal understanding, text-to-image generation and image-to-text generation tasks.

3 DAVINCI

Given the superior performance of auto-regressive language models (LM) (Brown et al., 2020; Chowdhery et al., 2022; Rae et al., 2021) on zero-shot and few-shot transfer abilities, we decided to adopt a decoder optimized by language modeling loss to retain the generalization capabilities, and an encoder to represent the prefix input. Unlike using a causal mask in the decoder, the encoder employs fully-visible attention for the prefix input. This architecture resembles prefix language modeling, which shows effectiveness on a wide range of language tasks (Dong et al., 2019; Raffel et al., 2020) and enables zero-shot generalization abilities. Contrary to the previous multi-stage approaches (Wang et al., 2021a; Singh et al., 2021), our model is trained from scratch in an end-to-end manner thanks to the model's simplicity. In this section, we introduce the proposed prefix multi-modal modeling framework and the DAVINCI model. The overall architecture of DAVINCI is depicted in Figure 1.

We first explain our model architecture in detail in §3.1 and then introduce pre-training objectives and procedures in §3.2.

3.1 MODEL ARCHITECTURE

Textual Feature Embedding Given an input sentence S, we first use WordPiece (Wu et al., 2016) to tokenize it to a sequence of tokens $W = \{w_1, w_2, ..., w_n\}$. To obtain text features T, for each token w_i, a token embedding e_i and position embedding p_i are computed by two separate embedding matrices. Finally, the textual feature embedding $T = \{t_1, t_2, ..., t_i, ..., t_n\}$ is calculated by

$$t_i = LayerNorm(e_i + p_i), \tag{1}$$

where i indicates the i-th position, and $LayerNorm$ (Ba et al., 2016) is a layer normalization function.

Visual Feature Embedding Given an input image I, we first use a CNN backbone to extract and learn the image features. Following (Dai et al., 2021; Wang et al., 2021b), we use the first three blocks of ResNet (He et al., 2016) to obtain the feature maps. The feature maps are then flattened to $F = \{f_1, f_2, ..., f_m\}$ along the spatial dimension, where m denotes the number of features. To keep the position information of visual features, we inject absolute learned positional embeddings p and the final visual embeddings $V = \{v_1, v_2, ..., v_i, ..., v_m\}$ are calculated by

$$v_i = f_i + p_i, \tag{2}$$

where i indicates the i-th position.

Cross-Modal Transformer To fuse the textual and visual feature embeddings into a common space, we adopt a simple canonical Transformer architecture as the fusion module. The input is the combination of visual embedding V and textual embedding T, namely $X = \{x_1, x_2, ..., x_l\} = [V, T] = \{v_1, v_2, ..., v_m, t_1, t_2, ..., t_n\}$. The input embedding vectors X are then fed into a cross-modal Transformer encoder to obtain hidden state vectors $H = \{h_1, h_2, ..., h_l\}$. Finally, a Transformer decoder is applied to generate visual or textual tokens with H and decoder input as illustrated in Figure 1.

Image Tokenizer and Decoder Because Transformer is modeling on discrete tokens, to unify the text tokens and image tokens, we discretize an image into tokens by an image tokenizer and reconstruct the raw image by an image decoder. The image tokenizer and decoder are implemented with a discrete variational autoencoder (dVAE) (Ramesh et al., 2021). After training of the image tokenizer, it could tokenize an image I into a sequence of discrete visual tokens $Z = \{z_1, z_2, ..., z_m\}$ according to a learned vocabulary. Visual tokens Z serve as the ground-truth labels for the prefix image modeling objective. In our work, we directly use an off-the-shelf image tokenizer and decoder from VQGAN (Esser et al., 2021), with a vocabulary size of 1024 and a compression rate of 16, which means a 256×256 image will be tokenized into 16×16 grid of tokens and then flattened to a sequence of 256 tokens.

3.2 PRE-TRAINING OBJECTIVES

Our major motivation is to conduct language modeling with image supervision and image modeling with natural language supervision simultaneously, which only requires image and text pairs that are easy to collect, making our approach easy to scale. The interaction would force the vision-language model to have a deeper understanding of both text and image. Learning from this interaction connects the visual representation with textual representation, enabling zero-shot transfer.

Prefix Language Modeling (PLM) The core idea of prefix language modeling is "given a full image X_{image} and a prefix caption $\tilde{X}_{text}$, recover the masked textual tokens (i.e., suffix caption Y_{text})". Given an input caption, we first randomly mask some continuous words at the end (we call it suffix caption hereafter) and recover the masked textual tokens with full image by optimizing the negative log likelihood,

$$\mathcal{L}_{\text{PLM}} = -\sum_{(I,S)\in D} \log p(\mathbf{Y}_{\text{text}} \,|\mathbf{X}_{\text{image}}, \tilde{\mathbf{X}}_{\text{text}}), \tag{3}$$

where I and S are images and captions from the pre-training corpus D.

Because of the lack of textual information, recovering the suffix caption requires the model to understand both the image and prefix caption. The full image is rich in semantic information that would help language modeling. The prefix length is randomly decided during training, and especially

when prefix caption is none, this task will degenerate into "image captioning" task, which forces the model to generate a caption with the input image.

$$\mathcal{L}'_{\text{PLM}} = -\sum_{(I,S)\in D} \log p(\mathbf{Y}_{\text{text}} \,|\mathbf{X}_{\text{image}}) \tag{4}$$

Prefix Image Modeling (PIM) The core idea of prefix image modeling is "given a full caption and a corrupted image (we call it prefix image hereafter), recover the masked visual tokens". Given an input image, we first randomly mask some continuous image patches at the end (we call it suffix image hereafter). The prefix image and full caption will be fed into the model and try to recover the original visual tokens obtained by the image tokenizer.

$$\mathcal{L}_{\text{PIM}} = -\sum_{(I,S)\in D} \log p(\mathbf{Y}_{\text{image}} \,|\mathbf{X}_{\text{text}}, \tilde{\mathbf{X}}_{\text{image}}) \tag{5}$$

Similar to PLM, when prefix image is none, this task will degenerate into "text-to-image generation" task, forcing the model to generate an image with the input caption:

$$\mathcal{L}'_{\text{PIM}} = -\sum_{(I,S)\in D} \log p(\mathbf{Y}_{\text{image}} \,|\mathbf{X}_{\text{text}}) \tag{6}$$

Unified Learning Objective Our model is learned by optimizing the combination of PLM and PIM.

$$\mathcal{L} = \mathcal{L}_{\text{PLM}} + \mathcal{L}_{\text{PIM}} \tag{7}$$

4 EXPERIMENTS

4.1 PRE-TRAINING DATASETS

Since existing studies pre-trained their models on different corpora, making the fair comparison difficult. Considering results only on the state-of-the-art performance would underestimate the potential of this line of research. Therefore, we propose several practical settings including small-scale and large-scale, and then conduct detailed comparisons on them in section 5.1. More details about the datasets are shown in Appendix A.6.

Data Type	Dataset	Image Domain	#Total
In-Domain Data (ID)	COCO, Visual Genome	COCO	1.3M
Small-scale Web Data (SWD)	SBU, CC-3M, CC-12M	Web	14.9M
Object-Region Data (ORD)	VG regions, VG objects, COCO objects, Refcoco Open Image, Obj365	COCO Flickr	17.0M
Vision Data (VD)	ImageNet-21K	ImageNet	13.2M
Large-scale Web Data (LWD)	LAION-400M, DAVINCI-200M	Web	601.3M
Text Data (TD)	C4	Web	800GB

Table 1: Statistics of the pre-training datasets. #Total denotes the total number of image-text pairs.

4.2 DOWNSTREAM TASKS

We test our models' ability and versatility on five dimension: **language understanding** on 8 GLUE tasks (Wang et al., 2019), **vision understanding** on ImageNet fine-tuning and 12 popular vision datasets for linear evaluation, **multi-modal understanding** on VQAv2 (Goyal et al., 2017b), SNLI-VE (Xie et al., 2019) and NLVR2 (Suhr et al., 2019), **text-to-image generation** on COCO (Chen et al., 2015), and **image-to-text generation** on COCO, NoCaps (Agrawal et al., 2019), and VLUE (Zhou et al., 2022). Details of downstream tasks and the fine-tuning process are described in Appendix A.2.

4.3 IMPLEMENTATION DETAILS

Pre-training Our model is of base size, with similar parameters to BERT$_{base}$. The Transformer is implemented with a 6-layer encoder and a 6-layer decoder, 768 dimensions for hidden states, 512 for maximum input length and 3072 for intermediate size. We train our model from scratch

without initializing the Transformer encoder and decoder. However, the image encoder is initialized from ResNet-101 (He et al., 2016) with ImageNet weights since we find a warm start provides a reliable visual representation and helps the convergence. For models pre-training on large-scale data, we optimize 10 epochs while for other small-scale datasets, we optimize 40 epochs with AdamW optimizer. The weight decay is set to 0.01 with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is 2e-4 with a warm-up period for the first 2% steps and linearly decayed to 0 after 2% of the total training steps. All pre-training experiments are conducted on 32GB NVIDIA V100 GPUs. We adopt mixed-precision (Micikevicius et al., 2018) to accelerate training and save memory. The model trained on the largest data takes around 10 days on 1024 V100 GPUs.

Fine-tuning The learning rate is $\in$ {1e-5, 5e-5} and our model is optimized by AdamW. Because the image resolution is different between pre-training and fine-tuning, the position parameters are adapted using linear interpolation. For all downstream tasks, we apply random resize crops and horizontal flips augmentation during training. All fine-tuning experiments are conducted on 32GB NVIDIA V100 GPUs. More details of the network architectures and hyper-parameters setups are given in Appendix A.1.

4.4 Experiment Results

We extensively compare the performance of DaVinci with state-of-the-art unified foundation models and vision-language models across vision, language, and multi-modal tasks, accessing five different abilities: (1) text understanding, (2) image understanding, (3) text-to-image generation, (4) image-to-text generation, (5) multi-modal understanding.

Overall Performance We report the overall performance on 8 language tasks from GLUE, 12 vision tasks, 3 multi-modal tasks, 3 image-to-text tasks and 1 text-to-image task. We compare our model with FLAVA and SimVLM [3], two of the most recent and best performing vision-language foundation models. We also include comparisons with some baseline models (e.g., MIM, MLM, CLIP). There are several observations. First, DaVinci (column 8) outperforms FLAVA (column 3) and SimVLM (column 7) across almost all tasks, providing a new and stronger unified foundation model. Compared with FLAVA, DaVinci improves an average of 2.04%, 1.68% on language and vision tasks, respectively. Compared with SimVLM, DaVinci achieves comparable results on language tasks (+0.13%) while performing much better on vision tasks (+2.46%). To make a fair comparison in terms of similar data size, we compare FLAVA (70M data, column 3) with DaVinci (46.4M data, column 6). It is observed that DaVinci still outperforms FLAVA even with much fewer data. Considering the multi-modal tasks, DaVinci consistently outperforms FLAVA and SimVLM on VQA and VE. Note that FLAVA is incapable of generation and SimVLM cannot generate images; only DaVinci is competent to all tasks and demonstrates a stronger capability of unifying vision and language tasks.

Zero-shot and Few-shot Transfer One of the critical benefits of generative pre-trained vision-language models is the good generalization ability on zero-shot and few-shot tasks. For zero-shot transfer, two out-of-domain distribution datasets are considered (NoCaps and VLUE), with results shown in Table 2. First, DaVinci outperforms SimVLM on both zero-shot and few-shot settings, demonstrating its better transfer capabilities. It also shows the effectiveness and robustness of the synergy of our proposed language supervision and image supervision. Second, it is observed that the performance improvement is bigger on 647.7M data (column 7 v.s. column 8) than 46.4M data (column 5 v.s. column 6). This shows DaVinci generalizes well with the increase of large-scale data. We even observe some performance drops on small data (46.4M) but excellent performance improvements on large data (647.7M). It is consistent with the recent observation that zero-shot ability could only be triggered with large pre-training data (Wei et al., 2022) and scaling to large data and keeping simple training objectives benefit generalization performance (Wang et al., 2021b).

Comparison with state-of-the-art vision-language models In addition to unified vision-language foundation models, we compare DaVinci with state-of-the-art vision-language models as well. The results are shown in Table 2. DaVinci demonstrates its superiority in vision understanding and text-to-image generation. Compared with current popular auto-regressive image generation models

[3] Since SimVLM is not open-sourced and uses 1.8B in-house data without telling the exact size of its $base$ model, we replicate it on our data with the same size as DaVinci. Experiments on SimVLM$_{small}$ ensure our successful reproduction (see Appendix A.3).

		BERT	RoBERTa	ViT	CLIP	MLM 1	MIM 2	FLAVA 3	CLIP 4	SimVLM 5	DAVINCI 6	SimVLM 7	DAVINCI 8
Task	Eval.	16GB	160GB	13.2M	400M	70M	70M	70M	70M	46.4M	46.4M	647.7M	647.7M
MNLI	FT	84.20	87.60	–	33.52	73.23	–	80.33	32.85	82.13	82.25	**83.27**	83.13
CoLA	FT	54.60	63.60	–	25.37	39.55	–	50.65	11.02	52.47	52.10	54.22	**54.75**
MRPC	FT	84.75	90.20	–	69.91	73.24	–	84.16	68.74	82.70	83.14	84.26	**84.54**
QQP	FT	89.00	91.90	–	65.33	86.68	–	88.74	59.17	88.39	88.15	**89.05**	88.92
SST-2	FT	92.50	94.80	–	88.19	87.96	–	90.94	83.49	90.65	90.48	91.12	**91.37**
QNLI	FT	91.00	92.80	–	50.54	82.32	–	87.31	49.46	87.55	87.21	**88.28**	87.90
RTE	FT	62.50	78.70	–	55.23	50.54	–	57.76	53.07	59.80	60.72	63.34	**64.22**
STS-B	FT	88.20	91.20	–	15.98	78.89	–	85.67	13.70	86.62	86.27	**87.24**	87.05
NLP Avg.		80.84	86.35	–	50.51	71.55	–	78.19	46.44	78.79	78.79	80.10	**80.23**
ImageNet	LE	–	–	80.90	80.20	–	41.79	75.54	72.95	74.31	75.87	76.04	**77.65**
Food101	LE	–	–	86.70	91.56	–	53.30	88.51	85.49	83.41	89.33	85.52	**90.12**
CIFAR10	LE	–	–	96.90	94.93	–	76.20	92.87	91.25	91.56	93.01	92.41	**93.96**
CIFAR100	LE	–	–	86.40	81.10	–	55.57	77.68	74.40	72.51	78.98	75.23	**80.11**
Cars	LE	–	–	54.70	85.92	–	14.71	70.87	62.84	61.44	72.69	68.83	**74.57**
Aircraft	LE	–	–	46.00	51.40	–	13.83	47.31	40.02	41.28	47.42	47.75	**49.55**
DTD	LE	–	–	74.30	78.46	–	55.53	77.29	73.40	72.55	77.12	76.59	**78.33**
Pets	LE	–	–	92.70	91.66	–	34.48	84.82	79.61	78.77	85.52	86.13	**88.21**
Flowers102	LE	–	–	99.20	97.12	–	67.23	96.37	94.94	93.24	96.12	95.41	**96.88**
MNIST	LE	–	–	97.40	99.01	–	96.40	98.42	97.38	96.66	98.67	98.45	**99.01**
STL10	LE	–	–	99.50	99.09	–	80.12	98.89	97.29	97.51	99.03	98.02	**99.21**
Country211	LE	–	–	17.50	30.10	–	8.87	28.92	25.12	26.45	28.99	27.81	**29.94**
Vision Avg.		–	–	77.68	81.71	–	49.84	78.12	74.56	74.14	78.56	77.34	**79.80**
VQAv2	FT	–	–	–	54.83	–	–	72.49	59.81	72.12	73.89	75.03	**76.44**
SNLI-VE	FT	–	–	–	74.27	–	–	78.89	73.53	78.74	79.11	79.63	**80.01**
NLVR2	FT	–	–	–	–	–	–	–	–	77.45	77.91	79.72	**80.25**
I2T@B4	FT	–	–	–	–	–	–	–	–	38.00	38.50	38.10	**39.20**
I2T@C	FT	–	–	–	–	–	–	–	–	126.96	128.66	128.91	**130.44**
T2I@IS ↑	FT	–	–	–	–	–	–	–	–	–	17.55	–	**22.41**
T2I@FID ↓	FT	–	–	–	–	–	–	–	–	–	23.58	–	**19.82**
VQAv2	FS	–	–	–	–	–	–	–	–	54.69	54.85	51.88	**54.90**
SNLI-VE	FS	–	–	–	–	–	–	–	–	67.45	67.57	67.96	**68.04**
NLVR2	FS	–	–	–	–	–	–	–	–	51.46	51.19	51.49	**51.52**
I2T@B4	FS	–	–	–	–	–	–	–	–	35.90	36.40	32.70	**37.00**
I2T@C	FS	–	–	–	–	–	–	–	–	117.75	120.43	112.20	**122.56**
I2T@B4	ZS	–	–	–	–	–	–	–	–	11.40	10.80	13.80	**18.70**
I2T@C	ZS	–	–	–	–	–	–	–	–	45.30	45.55	56.69	**68.44**
VLUE@B4	ZS	–	–	–	–	–	–	–	–	9.20	9.40	10.40	**10.60**
VLUE@C	ZS	–	–	–	–	–	–	–	–	33.92	34.80	39.75	**40.83**
NoCaps@C	ZS	–	–	–	–	–	–	–	–	48.05	45.51	48.64	**58.58**
T2I@IS ↑	ZS	–	–	–	–	–	–	–	–	–	14.91	–	**17.44**
T2I@FID ↓	ZS	–	–	–	–	–	–	–	–	–	29.83	–	**24.21**
Multi-modal Avg.		–	–	–	–	–	–	–	–	57.89	58.30	59.13	**62.50**

Table 2: Experimental results on vision, language and multi-modal downstream tasks. MNLI results are average of MNLI-m and MNLI-mm. MRPC and QQP results are average of accuracy and F1. Matthews correlation coefficient (MCC) is reported for CoLA and Pearson correlation coefficient (PCC) is reported for STS-B. @B4, @C denote BLEU@4, CIDEr, respectively. I2T and T2I denote image-to-text and text-to-image tasks. Multi-modal Avg. is the average score of all multi-modal tasks. FT: fine-tuning, LE: linear evaluation, FS: few-shot, ZS: zero-shot. Under few-shot setting, we fine-tune a pre-trained model for 3 epochs on 1% training data. Results for BERT are obtained from Iki & Aizawa (2021). Results for RoBERTa are from its corresponding paper (Liu et al., 2019) and they use the use mid-training (Phang et al., 2018) on MNLI for RTE, MRPC and STS-B while other models (e.g., BERT, SimVLM, DAVINCI) do not apply this trick. Results for ViT are from ViT-Base/16 model (Radford et al., 2021). CLIP-400M is the CLIP ViT-Base/16 model, and its linear evaluation results are from Singh et al. (2021).

like DALLE and CogView, our model achieves comparable IS and better FID scores with significantly fewer model parameters than DALLE and CogView. Note that the original DALLE is implemented based on VQVAE, so here, we compare our model with reproduced VQGAN-based DALLE with similar model sizes, and find DAVINCI still achieves a significant improvement over it. Generated images are presented in Appendix A.4 for further qualitative comparison.

On multi-modal tasks such as VQA, DAVINCI not only outperforms unified models (e.g., SimVLM (640M)) and other encoder-decoder multi-modal models (e.g., E2E-VLP, VL-T5), but also achieves competitive performance with many conventional encoder-only multi-model models (e.g., VinVL, ALBEF, VLMO). Note that SimVLM (1.8B) and OFA are not directly comparable because SimVLM uses 1.8B in-house image-text pairs, and OFA uses human-labeled data of many downstream tasks during pre-training. Even though, we still report their results for reference and observe a better performance on ImageNet fine-tuning and text-to-image generation than OFA.

The advantages of image generation over DALLE / CogView, the superiority of image-to-text over SimVLM, and the competitive performance with conventional multi-modal models demonstrate the synergistic effect of our proposed PLM (language supervision) and PIM (image supervision).

Model	#Params.	Text MNLI Acc	Vision ImageNet LE / FT	Image2Text COCO B@4 / C	Text2Image COCO IS↑ / FID↓	Multi-modal VQA test-dev / test-std	 NLVR2 dev / test-P
Encoder-only Multi-modal Models							
VisualBERT (Li et al., 2019)	170M	81.60	–	–	–	70.80 / 71.00	67.40 / 67.00
ViLBERT (Lu et al., 2019)	274M	79.90	–	–	–	70.55 / 70.92	–
VL-BERT (Su et al., 2020)	170M	81.20	–	–	–	71.16 / –	–
LXMERT (Tan & Bansal, 2019a)	240M	80.40	–	–	–	72.42 / 72.54	74.90 / 74.50
OSCAR (Li et al., 2020)	155M	–	–	36.5 / 123.7	–	73.16 / 73.44	78.07 / 78.36
VinVL (Zhang et al., 2021b)	157M	–	–	38.2 / 129.3	–	75.95 / 76.12	82.05 / 83.08
ViLT (Kim et al., 2021)	88M	–	–	–	–	70.85 / –	74.91 / 75.57
ALBEF (Li et al., 2021)	210M	–	–	–	–	75.84 / 76.04	82.55 / 83.14
X-VLM (Zeng et al., 2021)	240M	–	–	39.6 / 132.6	–	78.22 / 78.37	84.41 / 84.76
VLMO (Wang et al., 2021a)		–	–	–	–	76.64 / 76.89	82.77 / 83.34
Encoder-Decoder Multi-modal Models							
UNICORN (Yang et al., 2021)		–	–	35.8 / 119.1	–	69.20 / 69.40	– / –
Uni-ENDN (Li et al., 2022b)	110M	–	–	–	–	72.20 / 72.50	– / –
Pixel-BERT (Huang et al., 2020)	144M	–	–	–	–	74.45 / 74.55	76.50 / 77.20
E2E-VLP (Xu et al., 2021a)	94M	–	–	36.2 / 117.3	–	73.25 / 73.67	77.25 / 77.96
VL-T5 (Cho et al., 2021)	220M	–	–	34.5 / 116.5	–	– / 70.30	74.60 / 73.60
VL-BART (Cho et al., 2021)	220M	–	–	35.1 / 116.6	–	– / 71.30	71.70 / 70.30
Text2Image Models							
AttnGAN (Xu et al., 2018)		–	–	–	23.30 / 35.20	– / –	– / –
DM-GAN (Zhu et al., 2019)		–	–	–	32.20 / 26.50	– / –	– / –
DALLE (Ramesh et al., 2021) (250M)	12B	–	–	–	17.90 / 27.50	– / –	– / –
DALLE (Ramesh et al., 2021) (640M)†	82M	–	–	–	15.79 / 29.22	– / –	– / –
CogView (Ding et al., 2021)	4B	–	–	–	18.20 / 27.10	– / –	– / –
Unified Models							
Unifying (Huang et al., 2021)	228M	–	–	37.3 / 122.6	– / 29.90	– / –	– / –
FLAVA (Singh et al., 2021)	240M	80.33	75.54 / –	–	–	72.80 / 72.49	– / –
SimVLM (Wang et al., 2021b) (640M)†	153M	**83.27**	76.04 / –	38.5 / 128.7	–	75.04 / 75.03	78.82 / 79.72
SimVLM (Wang et al., 2021b) (1.8B)		83.40	80.60 / –	39.0 / 134.8	–	77.87 / 78.14	81.72 / 81.77
OFA (Wang et al., 2022)	182M	84.30	– / 82.20	41.0 / 138.2	21.50* / 20.80*	78.00 / 78.10	– / –
Florence (Yuan et al., 2021)	637M	–	– / 90.05	– / –	– / –	80.16 / 80.36	– / –
DAVINCI	154M	83.13	**78.81 / 83.92**	**39.2 / 130.4**	**17.44 (22.41*) / 24.21 (19.82*)**	**76.32 / 76.44**	**80.03 / 80.25**

Table 3: Comparison with state-of-the-art vision-language models on vision, language and multi-modal downstream tasks. All results are from *base*-size models. LE and FT denote linear evaluation and fine-tuning performance, respectively. Image2Text results are reported without CIDEr optimization. † are our reproduced models. * are the results after fine-tuning. SimVLM (1.8B) and OFA are pre-trained with much larger corpus or human-labeled data of many downstream tasks, and thus they are not comparable and labeled in gray. Florence (Yuan et al., 2021) is pre-trained with much larger model size (Florence-CoSwin-H, 637M) and more pre-training data (900M), so the numbers are in grey. **bold** denotes the best across unified models.

5 ANALYSIS

5.1 IMPACT OF PRE-TRAINING DATASETS

In this section, we disclose the impact of various multi-modal data sources for VLMs. We choose SimVLM and DAVINCI as our baseline models for their competitive performance, the capability of training from scratch, and the scalability of extending to the noisy large-scale corpus. We use the same text corpus, $C4$, for all the variations. The results are shown in Table 4. In general, the performance is increased along with the data size, and DAVINCI consistently outperforms SimVLM on almost all the data settings and all the downstream tasks. Both object-region data and vision data are clearly helpful in vision language pre-training (refer to settings 3 and 4). We surprisingly observe that models pre-trained on object-region data with much fewer images, performs even better than models pre-trained with small-scale web data on the COCO Caption task (refer to settings 2 and 3). Although large-scale web data is usually noisier than small datasets (e.g., ID, ORD, VD, and SWD), it is powerful for multi-modal pre-training (refer to settings 5 and 8).

We believe our analysis has broader impacts on the research of VLMs in the community. First, this enables fair comparisons for pre-trained models in the same data settings. Second, one can focus on the model designs at part or all of the data settings according to available computation resources. Third, we reveal that object-region and vision data, normally overlooked in VLM pre-training, also play a significant role.

5.2 ABLATION STUDY

To verify the contributions of different modules in our framework, we ablate them and evaluate the DAVINCI model on three kinds of downstream tasks: language understanding (MNLI, SST-2), vision understanding (ImageNet, Food101, CIFAR10), multi-modal understanding (VQAv2, SNLI-VE, NLVR2), image-to-text generation (COCO Captions), and text-to-image generation. Our model adopts dynamic masking ratios as described in Section 3.2. We also conduct experiments with static masking ratios to compare these two masking methods. Under static masking setting, the prefix

Settings	Pre-training Data					#Image	#Caption	Models	COCO Captions	VQA	SNLI-VE	NLVR2
	ID	SWD	ORD	VD	LWD				B@4 / C	Acc	Acc	Acc
1	✓					0.2M	1.3M	SimVLM	35.2 / 115.06	68.89	76.10	71.21
								DaVinci	**35.8 / 117.30**	**69.25**	**76.22**	**72.55**
2	✓	✓				15.1M	16.2M	SimVLM	37.0 / 122.63	71.54	78.36	75.50
								DaVinci	**37.4 / 123.11**	**71.88**	**78.62**	**77.46**
3	✓		✓			2.7M	18.3M	SimVLM	**38.2** / 123.85	69.57	76.65	70.50
								DaVinci	38.0 / **124.20**	**70.02**	**76.92**	**72.01**
4	✓			✓		13.4M	14.5M	SimVLM	36.2 / 119.73	70.53	76.90	73.25
								DaVinci	**36.6 / 121.27**	**71.23**	**77.40**	**74.62**
5	✓	✓	✓	✓		30.5M	46.4M	SimVLM	38.5 / 128.12	71.84	78.81	76.75
								DaVinci	**38.6 / 128.73**	**73.53**	**79.24**	**77.55**
6					✓	601.3M	601.3M	SimVLM	37.3 / 123.81	73.73	78.79	77.69
								DaVinci	**37.6 / 124.42**	**73.95**	**79.29**	**78.54**
7	✓				✓	601.5M	602.6M	SimVLM	37.9 / 125.50	74.64	79.05	77.68
								DaVinci	**38.1 / 125.91**	**74.91**	**79.22**	**78.12**
8	✓	✓	✓	✓	✓	631.8M	647.7M	SimVLM	38.5 / 128.25	75.04	79.32	78.82
								DaVinci	**39.1 / 130.21**	**76.32**	**80.04**	**80.03**

Table 4: Evaluation on downstream tasks using COCO Captions, VQA, SNLI-VE, and NLVR2. #Image and #Caption denote the numbers of images and image-text pairs that are used in the pre-training.

Method	COCO B@4 / C	VQA Acc	SNLI-VE Acc	NLVR2 Acc	ImageNet Acc	Food101 Acc	CIFAR10 Acc	MNLI Acc	SST-2 Acc	T2I IS / FID
No Pre-training	32.1 / 96.71	52.73	54.23	51.08	–*	–*	–*	66.32	79.84	–
Ours	35.8 / 117.30	69.25	76.22	72.55	48.88	75.32	73.82	81.76	90.25	12.35 / 53.14
– PLM	33.6 / 111.17	65.15	73.91	53.28	48.05	74.17	72.98	81.42	89.97	10.26 / 59.64
– PIM	34.3 / 116.58	68.89	75.79	69.78	45.54	71.18	70.11	81.94	90.53	–
– Text2Text	34.1 / 115.21	68.14	75.38	70.34	48.67	74.26	73.23	76.48	88.14	12.07 / 54.77
PL=0	35.4 / 117.00	66.90	75.52	71.05	48.45	68.18	73.73	78.69	89.00	11.76 / 55.38
PL=15%	35.7 / 116.53	69.16	75.09	70.44	41.58	52.15	68.55	79.02	89.46	–
PL=50%	35.1 / 115.53	68.55	74.54	56.92	37.69	49.16	70.15	78.59	89.69	–

Table 5: Ablation study on COCO Captions, VQA, SNLI-VE, NLVR2, ImageNet, Food101 and CIFAR10, MNLI, SST-2, text-to-image generation. "–" denotes removing the corresponding objective. Because linear probe requires a pre-trained model to be frozen, the "No Pre-training" results on ImageNet, Food101 and CIFAR10 are not reported and labeled by *. PL denotes the prefix length under the fixed masking ratio settings. For text-to-image (T2I), we report the zero-shot results. Note that the following four variants cannot perform zero-shot text-to-image generation: (1) No Pre-training, (2) Ours without PIM, (3) PL=15%, and (4) PL=50%.

length is fixed to 0, 15% and 50%. Note that when prefix length is 0, it becomes a standard language model. Experiments are conducted with the same model architecture (6-layer encoder + 6-layer decoder with 768 hidden dimensions) on in-domain data (ID). The results are shown in Table 5. First, all three modules (PLM, PIM, and Text2Text) bring improvement and the combination confirms a synergistic effect. Second, it is observed that without PLM, the performance decreases significantly on multi-modal understanding and image-to-text generation, indicating the importance of language supervision. Third, PIM brings more gains than PLM and text2text on vision understanding, which is expected because it enhances the vision encoding ability with image supervision. In addition, text2text objective is important to text understanding. The comparison between dynamic masking ratios and static masking ratios (PL=0, 15% and 50%) reveals that dynamic masking is better. We attribute this improvement to the smoothing effects of dynamic masking ratios. We also find that the standard language model (PL=0) performs worse on VQA, Food101 and text-to-image generation, which is consistent with the observation in SimVLM. Last, on the text-to-image generation task, it is observed that PLM is also helpful, confirming the synergistic effect of PIM and PLM again. Intuitively, PIM and PLM can help each other learn the alignments of visual and textual features, which will benefit both image generation and other multi-modal tasks.

6 Conclusion and Discussion

In this work, we first benchmark several settings on sequence-to-sequence vision-language pre-training in terms of pre-training dataset size, aligning SimVLM and our model on them. We propose a simple and unified generative pre-training model, DaVinci, to simultaneously leverage the language supervision and image supervision through two objectives under a unified framework: prefix language modeling and prefix image modeling. DaVinci is simple yet effective, demonstrating strong capabilities in both multi-modal writing and painting tasks. Experimental results explicitly imply that combining suffix caption generation and suffix image generation offer large gains on all benchmark settings. We also discussed limitations and the future work in Appendix A.5.

REFERENCES

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski (eds.). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1000.

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 8947–8956. IEEE, 2019. doi: 10.1109/ICCV.2019.00904. URL https://doi.org/10.1109/ICCV.2019.00904.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *ArXiv preprint*, abs/1607.06450, 2016. URL https://arxiv.org/abs/1607.06450.

Hangbo Bao, Li Dong, and Furu Wei. BEiT: Bert pre-training of image transformers. *arXiv preprint*, 2021.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *TAC*, 2009.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

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*, 2015.

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 *European Conference on Computer Vision (ECCV)*, 2020.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1931–1942. PMLR, 2021. URL http://proceedings.mlr.press/v139/cho21a.html.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *ArXiv preprint*, abs/2204.02311, 2022. URL https://arxiv.org/abs/2204.02311.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pp. 177–190. Springer, 2005.

Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *Advances in Neural Information Processing Systems*, 34:3965–3977, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Ming Ding, Zhuoyi Yang, Wenyi 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. *Advances in Neural Information Processing Systems*, 34, 2021.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002.

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*, pp. 13042–13054, 2019.

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

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*, pp. 12873–12883, 2021.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9, Prague, 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-1401.

Ross B. Girshick. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1440–1448. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.169. URL https://doi.org/10.1109/ICCV.2015.169.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *ArXiv preprint*, abs/1706.02677, 2017a. URL https://arxiv.org/abs/1706.02677.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. IEEE Computer Society, 2017b. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, volume 7, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

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. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Kashmir Hill and Jeremy White. Designed to deceive: Do these people look real to you. *The New York Times*, 11, 2020.

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*, pp. 1138–1147, 2021.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint*, 2020.

Taichi Iki and Akiko Aizawa. Effect of visual extensions on natural language understanding in vision-and-language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2189–2196, 2021.

Shankar Iyer, Nikhil Dandekar, Kornél Csernai, et al. First quora dataset release: Question pairs. *data. quora. com*, 2017.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21k.html.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

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 *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

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 preprint*, 2019.

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

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

Yehao Li, Jiahao Fan, Yingwei Pan, Ting Yao, Weiyao Lin, and Tao Mei. Uni-eden: Universal encoder-decoder network by multi-granular vision-language pre-training. *ArXiv preprint*, abs/2201.04026, 2022b. URL https://arxiv.org/abs/2201.04026.

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*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13–23, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1gs9JgRZ.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 3498–3505. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *ArXiv*, abs/1811.01088, 2018.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

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 (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

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 (JMLR)*, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 2021. URL http://proceedings.mlr.press/v139/ramesh21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1179–1195. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.131. URL https://doi.org/10.1109/CVPR.2017.131.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 2226–2234, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

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*, abs/2112.04482, 2021. URL https://arxiv.org/abs/2112.04482.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

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

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

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)*, pp. 5100–5111, Hong Kong, China, 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

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)*, pp. 5100–5111, Hong Kong, China, 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021. URL http://proceedings.mlr.press/v139/touvron21a.html.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

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. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *ArXiv preprint*, abs/2202.03052, 2022. URL https://arxiv.org/abs/2202.03052.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv preprint*, abs/2111.02358, 2021a. URL https://arxiv.org/abs/2111.02358.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint*, 2021b.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290. URL https://aclanthology.org/Q19-1040.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/1609.08144, 2016. URL https://arxiv.org/abs/1609.08144.

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

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 *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 503–513, Online, 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.42. URL https://aclanthology.org/2021.acl-long.42.

Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. A survey on green deep learning. *ArXiv preprint*, abs/2111.05193, 2021b. URL https://arxiv.org/abs/2111.05193.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1316–1324. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00143. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Xu_AttnGAN_Fine-Grained_Text_CVPR_2018_paper.html.

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.

Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *ArXiv preprint*, abs/1708.03888, 2017. URL https://arxiv.org/abs/1708.03888.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14448–14457, 2021.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel 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 preprint*, 2021.

Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ArXiv preprint*, abs/2111.08276, 2021. URL https://arxiv.org/abs/2111.08276.

Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *ArXiv preprint*, abs/2112.15283, 2021a. URL https://arxiv.org/abs/2112.15283.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021b.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. Vlue: A multi-task benchmark for evaluating vision-language models. *CoRR*, abs/2205.15237, 2022.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5802–5810. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00595. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Zhu_DM-GAN_Dynamic_Memory_Generative_Adversarial_Networks_for_Text-To-Image_Synthesis_CVPR_2019_paper.html.

A APPENDIX

A.1 DETAILS OF HYPER-PARAMETERS

Pre-training Our model is a base-size Transformer implemented with a 6-layer encoder and a 6-layer decoder, 768 dimensions for hidden states, 512 for maximum input length, and 3072 for intermediate size. We train our model from scratch without initializing the Transformer encoder and decoder. The image encoder is initialized from ResNet-101 (He et al., 2016) with ImageNet weights since we find a warm start provides a reliable visual representation and helps the convergence. For models pre-training on large-scale data, we optimize 10 epochs while for other small-scale datasets, we optimize 40 epochs with the AdamW optimizer. The weight decay is set to 0.01 with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is 2e-4 with a warm-up period for the first 2% steps and linearly decayed to 0 after 2% of the total training steps. In each batch, there are 8,192 image-text pairs for text-to-image generation and image-to-text generation with 8,192 text-only documents for text-to-text generation. We use center-crop to resize each image to the size of 256×256, which is the only data augmentation used during training. All pre-training experiments are conducted on 32GB NVIDIA V100 GPUs. We adopt mixed-precision (Micikevicius et al., 2018) to accelerate training and save memory. The model trained on the largest data takes around 10 days on 1024 V100 GPUs. The default settings are shown in Table 6.

Fine-tuning The learning rate is $\in$ {1e-5, 5e-5} and our model is optimized by AdamW. Because the image resolution differs between pre-training and fine-tuning, the position parameters are adapted using linear interpolation. For all downstream tasks, we apply random resize crops and horizontal flips augmentation during training. All fine-tuning experiments are conducted on 32GB NVIDIA V100 GPUs. The default settings for text classification, image classification, multi-modal understanding and image-to-text generation are shown in Tables 7, 8, and 9, respectively.

config	value
optimizer	AdamW (Loshchilov & Hutter, 2019)
learning rate	2e-4
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	8192
learning rate schedule	linear decay
warmup ratio (Goyal et al., 2017a)	0.02
training epochs	{10, 40}
augmentation	RandomResizedCrop

Table 6: Pre-training setting.

config	value
optimizer	AdamW
learning rate	{1e-5, 2e-5, 5e-5}
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	{16, 32, 64}
learning rate schedule	linear decay
warmup ratio	0.1
training epochs	{5, 10}

Table 7: Text classification: GLUE setting.

A.2 DETAILS OF DOWNSTREAM TASKS

Language Understanding We conduct experiments on GLUE benchmark including MNLI (Williams et al., 2018), CoLA (Warstadt et al., 2019), MRPC (Dolan & Brockett, 2005), QQP (Iyer et al., 2017), SST-2 (Socher et al., 2013), QNLI (Rajpurkar et al., 2016), RTE (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), and

config	value
optimizer	LARS (You et al., 2017)
base learning rate	0.1
weight decay	0
optimizer momentum	0.9
batch size	16384
learning rate schedule	cosine decay
warmup epochs	10
training epochs	90
augmentation	RandomResizedCrop

Table 8: Image classification: Linear probing setting.

config	value
optimizer	AdamW
learning rate	{1e-5, 5e-5}
weight decay	0.02
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	1024
learning rate schedule	linear decay
warmup epochs	[2, 5]
training epochs	[5, 15]
label smoothing (Szegedy et al., 2016)	0.1
augmentation	RandomResizedCrop, HorizontalFlips

Table 9: Multi-modal understanding and image-to-text generation: fine-tuning setting.

STS-B (Agirre et al., 2007). We follow the practice of BART (Lewis et al., 2020) and feed the same input into the encoder and decoder, and the hidden state of the final decoder token is fed into a new multi-class linear classifier or regression head.

Vision Understanding We conduct vision experiments on both fine-tuning and linear evaluation (linear eval). The linear evaluation follows a common practice (Caron et al., 2021; He et al., 2020; Singh et al., 2021) in self-supervised learning to evaluate the representation quality, where the pre-trained backbone model is frozen, and a new linear classifier is appended on top of it. We choose 12 popular datasets: ImageNet (Russakovsky et al., 2015), Food101 (Bossard et al., 2014), CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), Cars (Krause et al., 2013), Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014), Pets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), MNIST (LeCun & Cortes, 2010), STL10 (Coates et al., 2011), and Country211 (Radford et al., 2021).

Multi-modal Understanding We consider three popular multi-modal tasks: VQAv2 (Goyal et al., 2017b), SNLI-VE (Xie et al., 2019) and NLVR2 (Suhr et al., 2019) to evaluate our model's multi-modal understanding ability. For VQAv2, following ALBEF (Li et al., 2021), the image and question are fed to the encoder and the decoder generates answers based on the multi-modal embeddings. For SNLI-VE, we follow SimVLM (Wang et al., 2021b) to feed the image to the encoder and the text to the decoder. A classifier is appended on top of our pre-trained model, and it is trained to predict the result based on the last hidden states of the decoder. For NLVR2, two input pairs are constructed, each of them including one image and the textual description. The prediction is made based on the concatenation of these two embeddings following SimVLM (Wang et al., 2021b). The resolutions for VQAv2, SNLI-VE, NLVR2 are 480, 384, 384, respectively.

Text-to-Image Generation The text-to-image task requires the model to understand the textual instruction first and then draw the image according to the input's intention. The input text is fed to our encoder and our decoder will generate visual tokens one by one. After obtaining visual tokens, they are decoded to a raw image by an image decoder. We directly use an off-the-shelf image decoder from VQGAN (Esser et al., 2021). Following (Ramesh et al., 2021) we directly evaluate our pre-trained model on 30, 000 images randomly sampled from COCO (Chen et al., 2015) validation split. Both Fréchet Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016) are reported. The image resolution is 256.

Image-to-Text Generation For image-to-text generation (also called image captioning), the image is given to encoder and the decoder will generate the corresponding caption. Our experiments are conducted on COCO dataset (Chen et al., 2015) with cross-entropy optimization. Other task-specific techniques such as CIDEr optimization (Rennie et al., 2017) are not introduced. The image resolution is 480. We also conduct zero-shot captioning experiments on NoCaps (Agrawal et al., 2019) and VLUE (Zhou et al., 2022).

A.3 Reproduction of SimVLM

Since SimVLM is not open-sourced, we need to reproduce it by ourselves. There are two main difficulties on the reproduction: 1. it uses 1.8 billion in-house data 2. the configurations (e.g., parameter size, number of layers) of its $base$ model are not clearly stated. However, there are still some clues in Section 4.4 of SimVLM paper, where they propose a SimVLM$_{small}$ model with 8 layers, 512 embedding dimension, and trained on about 200M web data. To demonstrate the success of our replication, we train a SimVLM$_{small}$ model with the exact same configurations on about 200M web data. We obtain a VQA score of 68.50, surpassing the reported score of 67.43 in the original paper. We argue this result verifies our successful replication.

A.4 Visualization of Image Generation

In this section, we conduct qualitative analysis by visualising the generation samples. Figure 2 shows the comparison with DALLE and OFA with the same query. More generated samples are shown in Figures 3.

Figure 2: Comparison with DALLE and OFA on text-to-image generation.

A.5 Limitation and Societal Impacts

Limitation. Like most of the previous pre-training studies, the entire project consumed 40 V100 GPU years on an in-house computing cluster with large electricity costs. We tried to keep our model size small enough, but there is still potential for efficiency improvement such as sparse training and progressive training. We will explore those techniques to improve the training efficiency and reduce the carbon footprint so that it can adhere to proposals on "green" deep learning (Schwartz et al., 2020; Xu et al., 2021b). Furthermore, although we have tried our best to include as many tasks as we can to demonstrate the versatility of DaVinci, we believe our method can be expanded to more tasks (e.g., machine translation, summarization, object detection, etc.), modalities (e.g., video and speech). We leave these investigations to future work.

Potential Societal Impacts. Our model has image generation ability with risk of abuse, like fake portraits on social media (Hill & White, 2020), which is common potential risk in image generation research. Viable solutions are watermarking (Yu et al., 2021) and introducing a strict user license.

A.6 PRE-TRAINING DATASETS

Since existing studies pre-trained their models on different corpora, some of which are publicly available (e.g., CC-3M, CC-12M) while some are in-house dataset (e.g., ALIGN (Jia et al., 2021)), making the fair comparison difficult. Considering results only on the state-of-the-art performance would underestimate the potential of this line of research. Therefore, we propose several practical settings, including small-scale and large-scale, and then conduct detailed comparisons on them in section 5.1.

We collect a large set of dataset with diverse distributions for pre-training. According to its source, we divide them into in-domain, small-scale web data, object-region data, vision data, and large-scale web data. The statistics and details are shown in Table 10. Most of them are naturally image-text pairs while to enrich our corpus, we leverage object descriptions, region descriptions, and vision data (i.e., ImageNet). For objects and regions, we crop them from the original image according to its bounding box. For vision data, because they are usually labeled with a single word or short phrase, we compose a description with prompt templates such as "A picture of [LABEL]" or "The image contains [LABEL]". For example, "A picture of cat" or "The image contains cat".

Data Type	Dataset	Image Domain	#Images	#Captions	#Total
In-Domain Data (ID)	COCO	COCO	110.3K	551.7K	1.3M
	Visual Genome	COCO	108.2K	759.0K	
Small-scale Web Data (SWD)	SBU	Web	859.7K	859.7K	14.9M
	CC-3M	Web	2.9M	2.9M	
	CC-12M	Web	11.1M	11.1M	
Object-Region Data (ORD)	VG regions	COCO	108.2K	3.6M	17.0M
	VG objects	COCO	108.2K	925.6K	
	COCO objects	COCO	110.3K	736.6K	
	Refcoco	COCO	27.9K	589.9K	
	Open Image	Flickr	1.7M	7.5M	
	Obj365	Flickr	577.6K	3.6M	
Vision Data (VD)	ImageNet-21K	ImageNet	13.2M	13.2M	13.2M
Large-scale Web Data (LWD)	DAVINCI-200M	Web	205.6M	205.6M	601.3M
	LAION-400M	Web	395.7M	395.7M	
Text Data (TD)	C4	Web	–	–	800GB

Table 10: Statistics of the pre-training datasets. #Images, #Captions and #Total denote number of images, number of image-text pairs and the total number of image-text pairs, respectively.

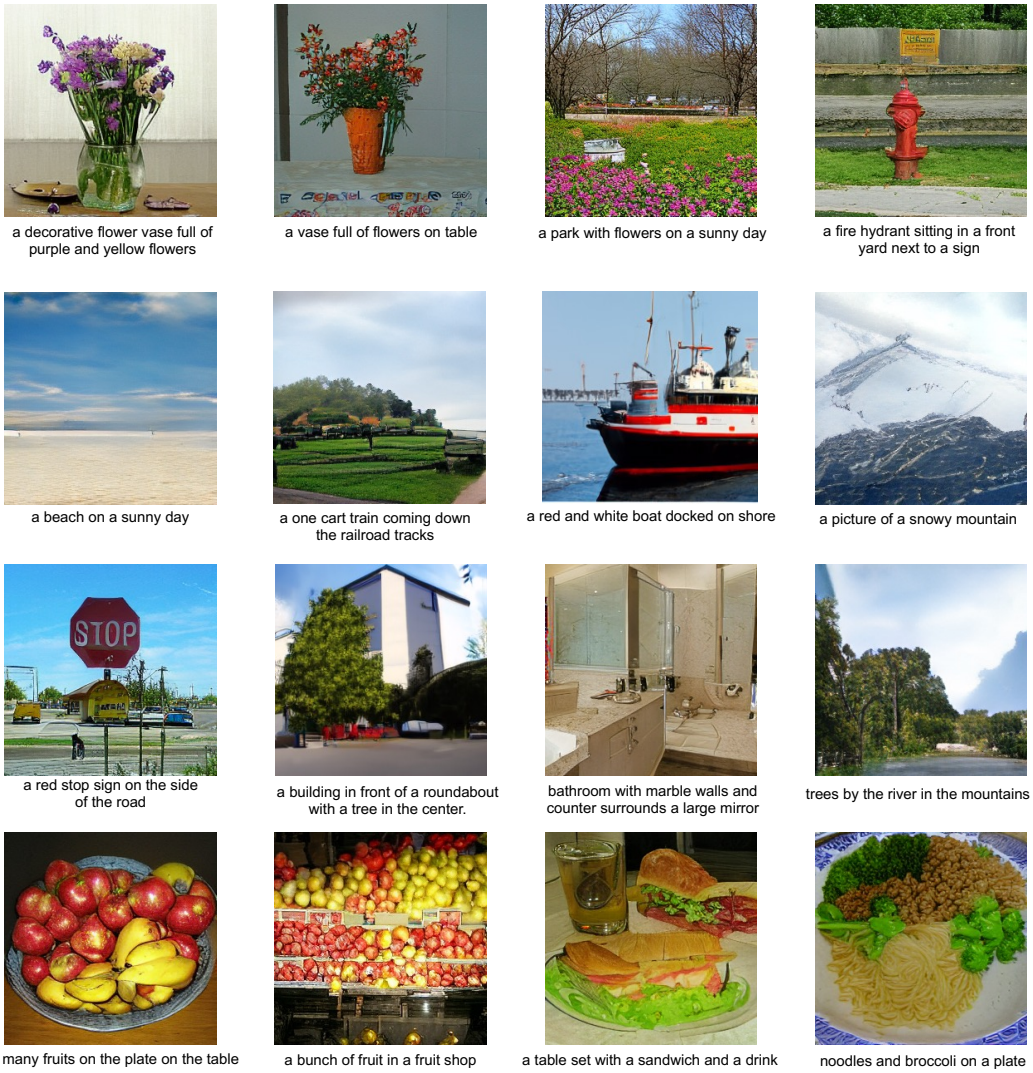

Figure 3: Generation samples by DAVINCI.