Prefix Language Models are Unified Modal Learners

Shizhe Diao*
The Hong Kong University of Science and Technology
sdiaoa@connect.ust.hk

Wangchunshu Zhou
ByteDance AI Lab
wcszhou@outlook.com

Xinsong Zhang†
ByteDance AI Lab
zhangxinsong.0320@bytedance.com

Jiawei Wang
Shanghai Jiao Tong University
wjw_sjt@sjtu.edu.cn

Abstract

With the success of vision-language pre-training, we have witnessed the state-of-the-art has been pushed on multi-modal understanding and generation. However, the current pre-training paradigm is either incapable of targeting all modalities at once (e.g., text generation and image generation), or requires multi-fold well-designed tasks which significantly limits the scalability. We demonstrate that a unified modal model could be learned with a prefix language modeling objective upon text and image sequences. Thanks to the simple but powerful pre-training paradigm, our proposed model, DaVinci, is simple to train, scalable to huge data, and adaptable to a variety of downstream tasks across modalities (language / vision / vision+language), types (understanding / generation) and settings (e.g., zero-shot, fine-tuning, linear evaluation) with a single unified architecture. DaVinci achieves the competitive performance on a wide range of 26 understanding / generation tasks, and outperforms previous unified vision-language models on most tasks, including ImageNet classification (+1.6%), VQAv2 (+1.4%), COCO caption generation (BLEU@4 +1.1%, CIDEr +1.5%) and COCO image generation (IS +0.9%, FID -1.0%), at the comparable model and data scale. Furthermore, we offer a well-defined benchmark for future research by reporting the performance on different scales of pre-training dataset on a heterogeneous and wide distribution coverage. Our results establish new, stronger baselines for future comparisons at different data scales and shed light on the difficulties of comparing VLP models more generally.

1 Introduction

Self-supervised language model pre-training [1–17] 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) [18, 19], achieving state-of-the-art results on various vision-language tasks [20]. Most existing VLMs are BERT-like Transformer [21] encoders pre-trained with a combination of different vision-language pre-training (VLP) objectives: masked multi-modal modeling [18, 22–24], multi-modal alignment prediction [18, 22–24], region of interest feature regression [22], image-text matching [25, 26], to name a few. However, the roadmap towards large language models reveals a transition pattern from encoder-only models like BERT [2] / RoBERTa [4] to sequence-to-sequence models like T5 [7] / BART [5] and autoregressive models like GPT-3 [8] / PaLM [27] to tackle more tasks in a unified way, and from complicated objectives like masked

* Work done during the internship at ByteDance AI Lab.
† Corresponding author

The code and pre-trained models will be released at https://github.com/shizhediao/DaVinci
language modeling / next sentence prediction / replace token detection to a simple language modeling objective to improve the scalability of pre-training. This suggests that while achieving competitive results, the encoder-only architecture and complicated pre-training objectives of most current VLMs inevitably limit the potential towards pre-training more scalable and general VLMs.

To this end, a number of recent studies [28–31] 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 [28] and OFA [31] formulate various vision-and-language problems into seq2seq tasks and pre-train a seq2seq VLM by multi-tasking on these tasks. This approach is also hard to scale because it is non-trivial to collect a large number of vision-language datasets for pre-training. On the other hand, ERNIE-ViLG [29] and SimVLM [30] 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.

Motivated by the success of large-scale generative pre-training of (prefix) language models and the goal of unifying modalities and task formats, we introduce prefix multi-modal modeling, a unified generative pre-training framework that extends prefix language modeling to the multi-modal context. 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 [32] and convert images into discrete sequences of image tokens [33]. 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 and is easily scalable for large-scale pre-training. We pre-train DAVINCI\(^4\), 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 vision-language understanding, image-to-text generation, text-to-image generation, and single-modal language / vision tasks. DAVINCI consistently outperforms FLAVA [34], 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.

In addition, most existing VLMs are pre-trained with different data sources varying in sizes and sources, making it difficult to disentangle the impact of pre-training methods and data sources on the downstream tasks. In our experiments, we conduct a systematic analysis of the performance of SimVLM (prefix language model) and DAVINCI (prefix multi-modal model) with different amounts of pre-training data, revealing the impact of different data sources and 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 multiple modalities (vision, language, multi-modal) and tasks (understanding or generation). (2) We pre-train DAVINCI, a vision-language foundation model, with the proposed approach and show that it performs competitively across tasks and modalities. (3) We conduct an analysis about the impact of different pre-training data sources on the performance of seq2seq VLMs.

2 Related Work

Inspired by the success of language model pre-training, a number of studies investigate vision-language pre-training on large-scale image-caption pairs. VILBERT [18] and LXMERT [22] first propose to extract visual object features with an external object detection model like Fast-RCNN [35], 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. Several following works [24, 36, 23] propose several new objectives to improve object detection based VLP.

Recently, with the development of vision Transformer [37, 38], a number of works [25, 26, 39] explored taking raw image pixels as vision input and extracting overall image features with vision Transformers. This makes VLMs more efficient by alleviating the object detection process while also

\(^4\)Named after the Italian polymath Leonardo da Vinci, who displayed infinite grace in everything.
enables VLMs to benefit from powerful pre-trained vision models such as Swin Transformer [40] and BEiT [41], thus achieving state-of-the-art performance on a wide range of vision-language tasks.

While achieving promising results, most VLMs are based on encoder-only architectures, making them not directly applicable for generative tasks such as image captioning and generative question answering. Inspired by the success of seq2seq pre-trained language models such as T5 [7] and BART [5], VL-T5 [28] and OFA [31] 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. While achieving promising results, the scalability of this approach is limited by the availability of large-scale and diverse vision-language tasks. To this end, SimVLM [30], the most related work to our approach, instead pre-trains a seq2seq VLM with a simple prefix language modeling objective on text generation. As such, it easily scales to very large and potentially noisy pre-training data and achieves competitive results. Our approach differs from SimVLM with a novel prefix image modeling objective which enables our model to better align vision and text modalities and be applicable to image generation tasks. More recently, FLAVA [34], a new vision-language foundation model, is pre-trained with a masked multi-modal modeling objective. While performing competitively on both language, vision, and vision-language understanding tasks, the encoder-only architecture of FLAVA limits its versatility for generation tasks.

3 DAVINCI

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 [42] 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

\[
t_i = \text{LayerNorm}(e_i + p_i),
\]

where \( i \) indicates the \( i \)-th position, and \( \text{LayerNorm} \) is a layer normalization function [43].

Visual Feature Embedding Given an input image \( I \), we first use a CNN backbone to extract and learn the image features. Following [44, 30], we use the first three blocks of ResNet [45] 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 feature embedding \( V = \{v_1, v_2, ..., v_i, ..., v_m\} \) is calculated by
\[
v_i = f_i + p_i,
\]
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) [32]. 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 [46], with a vocabulary size of 1024 and a compression rate of 16, which means a \( 256 \times 256 \) image will be tokenized into \( 16 \times 16 \) 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 at the same time, 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,
\[
\mathcal{L}_{PLM} = - \sum_{(I, S) \in D} \log p(Y_{text} | X_{image}, \tilde{X}_{text}),
\]
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 when it is zero, the prefix caption is nonexistent and this task will degenerate to “image captioning” task, which forces the model to generate caption with the input image.
\[
\mathcal{L}'_{PLM} = - \sum_{(I, S) \in D} \log p(Y_{text} | X_{image})
\]

**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 image tokenizer.
\[
\mathcal{L}_{PIM} = - \sum_{(I, S) \in D} \log p(Y_{image} | X_{text}, \tilde{X}_{image})
\]
Similar to PLM, when the prefix length is zero, this task will degenerate to “text-to-image generation” task, forcing the model to generate an image with the input caption:

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

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

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

### 4 Experiments

#### 4.1 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 [47]), 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 1. 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      |          |
|                      | SBU           | Web          | 859.7K    | 859.7K      | 1.7M     |
|                      | CC-3M         | Web          | 2.9M      | 2.9M        | 14.9M    |
|                      | CC-12M        | Web          | 11.1M     | 11.1M       |          |
| **Small-scale Web Data (SWD)** | VG regions    | COCO         | 108.2K    | 3.6M        |          |
|                      | VG objects    | COCO         | 108.2K    | 925.6K      | 17.0M    |
|                      | COCO objects  | COCO         | 110.3K    | 736.6K      |          |
|                      | Refcoco       | COCO         | 27.9K     | 589.9K      | 17.0M    |
|                      | Open Image    | Flickr       | 1.7M      | 3.6M        |          |
|                      | Obj365        | Flickr       | 577.6K    | 577.6K      |          |
| **Object-Region Data (ORD)** | VG regions    | COCO         | 108.2K    | 3.6M        |          |
|                      | 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      | 3.6M        |          |
|                      | Obj365        | Flickr       | 577.6K    | 577.6K      |          |
| **Vision Data (VD)** | ImageNet-21K  | ImageNet     | 13.2M     | 13.2M       | 13.2M    |
|                      | DAVINCI-200M  | Web          | 205.6M    | 205.6M      | 601.3M   |
|                      | LAION-400M    | Web          | 395.7M    | 395.7M      |          |
| **Text Data (TD)**   | C4            | Web          | –         | –           | 800GB    |

**Table 1:** 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.

#### 4.2 Downstream Tasks

**Language Understanding** We conduct experiments on GLUE benchmark including MNLI [48], CoLA [49], MRPC [50], QQP [51], SST-2 [52], QNLI [53], RTE [54–57], and STS-B [58]. We follow the practice of BART [5] 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). Linear evaluation follows a common practice [59, 60, 34] 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 [61], Food101 [62], CIFAR10 [63], CIFAR100 [63], Cars [64], Aircraft [65], DTD [66], Pets [67], Flowers102 [68], MNIST [69], STL10 [70], and Country211 [71].

Multi-modal Understanding We consider three popular multi-modal tasks: VQAv2 [72], SNLI-VE [73] and NLVR2 [74] to evaluate our model’s multi-modal understanding ability. For VQAv2, following ALBEF [25], 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 [30] to feed the image to encoder and the text to 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 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 [30].

Text-to-Image Generation 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 [46]. Following [32], we directly evaluate our pre-trained model on 30,000 images randomly sampled from COCO [75] validation split. Both Fréchet Inception Distance (FID) [76] and Inception Score (IS) [77] are reported.

Image-to-Text Generation For image-to-text generation (also called image captioning), the image is given to the encoder, and the decoder will generate the corresponding caption. Our experiments are conducted on COCO dataset [75] with cross-entropy optimization. Other task-specific techniques such as CIDEr optimization [78] are not introduced.

4.3 Implementation Details

Pre-training Our model is of base size, with similar parameters to BERTbase. 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 [45] 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. 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 $256 \times 256$, which is the only data augmentation used during training. All pre-training experiments are conducted on 32GB NVIDIA V100 GPUs. We adopt mixed-precision [79] 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 setup 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, 2 text-to-image tasks and 1 image-to-text task. We compare our model
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 on vision understanding and text-to-image generation. For example, on text-to-image generation, our model first outperforms previous GAN-based models in terms of FID. Although GAN-based models have higher IS, we argue FID is more reliable due to the well-known manipulation tricks and over-fitting issues of

---

3Since 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 own data with the same size as DAVinci. Experiments on SimVLMsmall ensure our successful reproduction (see Appendix A.3).
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, thus they are not comparable and labeled in gray. bold denotes the best across unified models.

IS [92, 93]. Compared with more advanced auto-regressive image generation models like DALLE andCogView, our model still achieves comparable IS and better FID scores with significantly less model parameters than DALLE and CogView. Note that the original DALLE is implemented based on VQVAE, here we compare our model with reproduced VQGAN-based DALLE with similar model size, 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-modal 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, text-to-image generation and VQA 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).

5 Analyses

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
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. Results are reported on the development set.

training from scratch and the scalability of extending to 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 useful in vision language pre-training (refer to settings 3 and 4). We surprisingly observe that models pre-trained with object-region data which has 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 for 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 data and vision data, which are 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 downstream tasks: COCO Captions, SNLI-VE and NLVR2. 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. All three modules bring improvement and the combination confirms a synergistic effect. In addition, it is observed that without PLM, the performance decreases significantly, indicating the importance of language supervision. More ablation studies on other tasks (e.g., vision understanding) are presented in Appendix A.5.

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. To enhance both vision and language understanding, we propose a novel, simple, and unified pre-training seq2seq model DAVINCI, to leverage the language supervision and image supervision through two objectives under a unified framework: prefix language modeling and prefix image modeling. Our method is easy to implement, simple and effective, especially it is scalable well without extra efforts. Experimental
results imply that explicitly generating suffix caption and suffix image offer large gains on all benchmark settings.

**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 [94, 95], dataset distillation [96], and progressive training [97]. 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 [98, 99]. 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.), and 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 [100], which are common potential risks in image generation research. Viable solutions are watermarking [101] and introducing a strict user license.

**Acknowledgments and Disclosure of Funding**

We would like to acknowledge Yan Zeng, Wenguan Huang and Zhi Zhang at ByteDance, Zhiling Zhang at Shanghai Jiao Tong University for their generous assistance in data collection and helpful discussions. We also wish to thank Hang Li at ByteDance, Tong Zhang at HKUST for inspiring feedback, valuable comments and great support to this work.

**References**

[1] 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), pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

[2] 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), pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

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

[4] 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.

[5] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Gharavininejad, 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, pages 7871–7880, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[6] 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, pages 13042–13054, 2019.

[7] 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.
[8] 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, editors, *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/1457c0dd6bfcb4967418bfb8ac50142f64a-Abstract.html.

[9] Zhiyi Fu, Wangchunshu Zhou, Jingjing Xu, Hao Zhou, and Lei Li. Contextual representation learning beyond masked language modeling. In *ACL (1)*, pages 2701–2714. Association for Computational Linguistics, 2022.

[10] Wangchunshu Zhou, Dong-Ho Lee, Ravi Kiran Selvam, Seyeon Lee, and Xiang Ren. Pre-training text-to-text transformers for concept-centric common sense. 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=3k20LAiHYL2.

[11] Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. Zen: Pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740, 2020.

[12] Shizhe Diao, Ruijia Xu, Hongjin Su, Yilei Jiang, Yan Song, and Tong Zhang. Taming pre-trained language models with n-gram representations for low-resource domain adaptation. 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)*, pages 3336–3349, 2021.

[13] Wangchunshu Zhou, Tao Ge, Canwen Xu, Ke Xu, and Furu Wei. Improving sequence-to-sequence pre-training via sequence span rewriting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 571–582, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.45. URL https://aclanthology.org/2021.emnlp-main.45.

[14] Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. BERT-of-theus: Compressing BERT by progressive module replacing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7859–7869, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.633. URL https://aclanthology.org/2020.emnlp-main.633.

[15] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian J. McAuley, Ke Xu, and Furu Wei. BERT loses patience: Fast and robust inference with early exit. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *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/d4dd111a4fd973394238aca50c05bebe3-Abstract.html.

[16] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. BERT learns to teach: Knowledge distillation with meta learning. In *ACL (1)*, pages 7037–7049. Association for Computational Linguistics, 2022.

[17] Shizhe Diao, Xuechun Li, Yong Lin, Zhichao Huang, and Tong Zhang. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*, 2022.

[18] 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’Alchê-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural
[19] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

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

[21] 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, S.V.N. Vishwanathan, and Roman Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[22] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, 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, pages 1931–1942. PMLR, 2021. URL http://proceedings.mlr.press/v139/cho21a.html.

[29] 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, 2021. URL https://arxiv.org/abs/2112.15283.
[30] 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*, 2021.

[31] 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.

[32] 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, editors, *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*, pages 8821–8831. PMLR, 2021. URL http://proceedings.mlr.press/v139/ramesh21a.html.

[33] 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, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html.

[34] 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.

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

[36] 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)*, 2021.

[37] 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.

[38] 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, editors, *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*, pages 10347–10357. PMLR, 2021. URL http://proceedings.mlr.press/v139/touvron21a.html.

[39] 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, 2021. URL https://arxiv.org/abs/2111.02358.

[40] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, 2021.

[41] Hangbo Bao, Li Dong, and Furu Wei. BEiT: Bert pre-training of image transformers. *arXiv preprint*, 2021.
[42] 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.

[43] 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.

[44] 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.

[45] 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*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

[46] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021.

[47] 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, editors, *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*, pages 4904–4916. PMLR, 2021. URL http://proceedings.mlr.press/v139/jia21b.html.

[48] 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)*, pages 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.

[49] 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.

[50] 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.

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

[52] 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*, pages 1631–1642, Seattle, Washington, USA, 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

[53] 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*, pages 2383–2393, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

[54] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.
[55] R Bar Haim, Idan Szpektor. The second pascal recognising textual entailment challenge. In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, volume 7, 2006.

[56] 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, pages 1–9, Prague, 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-1401.

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

[58] Eneko Agirre, Lluís Márquez, and Richard Wicentowski, editors. 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.

[59] 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, pages 9650–9660, 2021.

[60] 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, pages 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL https://doi.org/10.1109/CVPR42600.2020.00975.

[61] 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.

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

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

[64] 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, pages 554–561, 2013.

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

[66] 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, pages 3606–3613. IEEE Computer Society, 2014. doi: 10.1109/CVPR.2014.461. URL https://doi.org/10.1109/CVPR.2014.461.

[67] 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, pages 3498–3505. IEEE Computer Society, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://doi.org/10.1109/CVPR.2012.6248092.

[68] 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, pages 722–729. IEEE, 2008.

[69] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.
[70] 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, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, 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, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[72] 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, pages 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

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

[74] 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, pages 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

[75] 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.

[76] 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, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[77] 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, editors, Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pages 2226–2234, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.

[78] 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, pages 1179–1195. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.131. URL https://doi.org/10.1109/CVPR.2017.131.

[79] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David Garcia, 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.

[80] 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.
[81] 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.

[82] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, 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, pages 5583–5594. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21k.html.

[83] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In 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), pages 2592–2607, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.202. URL https://aclanthology.org/2021.acl-long.202.

[84] 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.

[85] 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, 2022. URL https://arxiv.org/abs/2201.04026.

[86] 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.

[87] Haiyang Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiao lei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attention generative adversarial networks. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 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.

[88] 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, pages 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.

[89] 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.

[90] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In Proceedings of the 29th ACM International Conference on Multimedia, pages 1138–1147, 2021.

[91] Shane Barratt and Rishi Sharma. A note on the inception score. ArXiv preprint, abs/1801.01973, 2018. URL https://arxiv.org/abs/1801.01973.
[93] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[94] Xiao Zhou, Weizhong Zhang, Hang Xu, and Tong Zhang. Effective sparsification of neural networks with global sparsity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3599–3608, 2021.

[95] Xiao Zhou, Weizhong Zhang, Zonghao Chen, Shizhe Diao, and Tong Zhang. Efficient neural network training via forward and backward propagation sparsification. *Advances in Neural Information Processing Systems*, 34, 2021.

[96] Xiao Zhou, Renjie Pi, Weizhong Zhang, Yong Lin, and Tong Zhang. Probabilistic bilevel coreset selection. In *International Conference on Machine Learning*, PMLR, 2022.

[97] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

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

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

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

[101] 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*, pages 14448–14457, 2021.

[102] 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.

[103] 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, 2017. URL https://arxiv.org/abs/1706.02677.

[104] 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.

[105] 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*, pages 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.
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 [45] 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. 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 $256 \times 256$, which is the only data augmentation used during training. All pre-training experiments are conducted on 32GB NVIDIA V100 GPUs. We adopt mixed-precision [79] 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 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. 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 [102]            |
| learning rate           | 2e-4                   |
| weight decay            | 0.01                   |
| optimizer momentum      | $\beta_1, \beta_2 = 0.9, 0.999$ |
| batch size              | 8192                   |
| learning rate schedule  | linear decay           |
| warmup ratio [103]      | 0.02                   |
| training epochs         | (10, 40)               |
| augmentation            | RandomResizedCrop      |

Table 6: Pre-training setting.

| config                  | value                  |
|-------------------------|------------------------|
| optimizer               | AdamW                  |
| learning rate           | $\{1e-5, 5e-5, 5e-5\}$ |
| weight decay            | 0.01                   |
| optimizer momentum      | $\beta_1, \beta_2 = 0.9, 0.999$ |
| 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 [48], CoLA [49], MRPC [50], QQP [51], SST-2 [52], QNLI [53], RTE [54–57], and STS-B [58]. We follow the practice of BART [5] and fed 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. The image resolution is 256.
### Vision Understanding

We conduct vision experiments on both fine-tuning and linear evaluation (linear eval). Linear evaluation follows a common practice [59, 60, 34] 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 [61], Food101 [62], CIFAR10 [63], CIFAR100 [63], Cars [64], Aircraft [65], DTD [66], Pets [67], Flowers102 [68], MNIST [69], STL10 [70], and Country211 [71]. The image resolution is 256.

### Multi-modal Understanding

We consider three popular multi-modal tasks: VQAv2 [72], SNLI-VE [73] and NLVR2 [74] to evaluate our model’s multi-modal understanding ability. For VQAv2, following ALBEF [25], the image and question are fed to encoder and the decoder generates answers based on the multi-modal embeddings. For SNLI-VE, we follow SimVLM [30] to feed the image to encoder and the text to 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 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 [30]. The resolutions for VQAv2, SNLI-VE, NLVR2 are 480, 384, 384, respectively.

### Text-to-Image Generation

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 [46]. Following [32] we directly evaluate our pre-trained model on 30,000 images randomly sampled from COCO [75] validation split. Both Fréchet Inception Distance (FID) [76] and Inception Score (IS) [77] 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 [75] with cross-entropy optimization. Other task-specific techniques such as CIDEr optimization [78] are not introduced. The image resolution is 480.

---

| config          | value                           |
|-----------------|---------------------------------|
| optimizer       | LARS [104]                      |
| 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 | $\beta_1, \beta_2=0.9, 0.999$   |
| batch size      | 1024                            |
| learning rate schedule | linear decay              |
| warmup epochs   | $[2, 5]$                        |
| training epochs | $[5, 15]$                       |
| label smoothing | 0.1                             |
| augmentation    | RandomResizedCrop, HorizontalFlips |

Table 9: Multi-modal understanding and image-to-text generation: fine-tuning setting.
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\textsubscript{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\textsubscript{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
Figure 3.

![Comparison with DALLE and OFA on text-to-image generation.](image)

A.5 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) and image-to-text generation (COCO Captions). 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 10. First, all three modules 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. In addition, PIM brings more gains than PLM and text2text on
vision understanding, which is expected because it enhances the vision encoding ability with image
supervision. Last, text2text objective is important to text understanding.
Table 10: Ablation study on COCO Captions, VQA, SNLI-VE, NLVR2, ImageNet, Food101, CIFAR10, MNLI and SST-2. “–” 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 ∗.

| Method          | COCO | VQA  | SNLI-VE | NLVR2 | ImageNet | Food101 | CIFAR10 | MNLI | SST-2 |
|-----------------|------|------|---------|-------|----------|---------|---------|------|-------|
| No Pre-training | 32.1 | 52.73| 54.23   | 51.08 | –        | –       | –       | 66.32| 79.84 |
| Ours            | 35.8 | 117.30| 69.25   | 72.55 | 48.88    | 75.32   | 73.82   | 81.76| 90.25 |
| – PLM           | 33.6 | 111.17| 65.15   | 73.91 | 48.05    | 74.17   | 72.98   | 81.42| 89.97 |
| – PIM           | 34.3 | 116.58| 68.89   | 75.79 | 45.54    | 71.18   | 70.11   | 81.94| 90.53 |
| – Text2Text     | 34.1 | 115.21| 68.14   | 70.34 | 48.67    | 74.26   | 73.23   | 76.48| 88.14 |

Figure 3: Generation samples by DAVINCI.