# Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks

Wenhui Wang*, Hangbo Bao*, Li Dong*, Johan Bjorck, Zhiliang Peng, Qiang Liu
Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, Furu Wei†
Microsoft Corporation
https://aka.ms/beit-3

## Abstract

*A big convergence of language, vision, and multimodal pretraining is emerging. In this work, we introduce a general-purpose multimodal foundation model* **BEIT-3**, *which achieves excellent transfer performance on both vision and vision-language tasks. Specifically, we advance the big convergence from three aspects: backbone architecture, pretraining task, and model scaling up. We use Multiway Transformers for general-purpose modeling, where the modular architecture enables both deep fusion and modality-specific encoding. Based on the shared backbone, we perform masked "language" modeling on images (**Imglish**), texts (English), and image-text pairs ("parallel sentences") in a unified manner. Experimental results show that* BEIT-3 *obtains remarkable performance on object detection (COCO), semantic segmentation (ADE20K), image classification (ImageNet), visual reasoning (NLVR2), visual question answering (VQAv2), image captioning (COCO), and cross-modal retrieval (Flickr30K, COCO).*

## 1. Introduction: The Big Convergence

Recent years have featured a trend toward the big convergence of language [14, 15, 46], vision [3, 43], and multimodal [45, 62, 69] pretraining. By performing large-scale pretraining on massive data, we can easily transfer the models to various downstream tasks. It is appealing that we can pretrain a general-purpose foundation model that handles multiple modalities. In this work, we advance the convergence trend for vision-language pretraining from the following three aspects.

First, the success of Transformers [59] is translated from language to vision [16] and multimodal [26, 62] problems. The unification of network architectures enables us to seamlessly handle multiple modalities. For vision-language modeling, there are various ways to apply Transformers due to the different natures of downstream tasks. For example, the dual-encoder architecture is used for efficient retrieval [45], encoder-decoder networks for generation tasks [63], and the fusion-encoder architecture for image-text encoding [26]. However, most foundation models have to manually convert the end-task formats according to the specific architectures. Moreover, the parameters are usually not effectively shared across modalities. In this work, we adopt Multiway Transformers [62] for general-purpose modeling, i.e., one unified architecture shared for various downstream tasks. The modular network also comprehensively considers modality-specific encoding and cross-modality fusion.

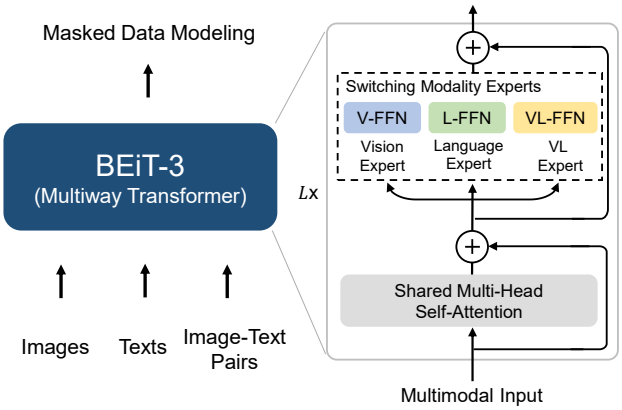


Figure 1. Overview of BEIT-3 pretraining. We perform masked data modeling on monomodal (i.e., images, and texts) and multimodal (i.e., image-text pairs) data with a shared Multiway Transformer as the backbone network.

Second, the pretraining task based on masked data modeling has been successfully applied to various modalities, such as texts [14] and images [3, 43]. Current vision-language foundation models usually multitask other pretraining objectives (such as image-text matching), rendering scaling-up unfriendly and inefficient. In contrast, we only use one pretraining task, i.e., mask-then-predict, to

* Equal contribution. † Corresponding author.

train a general-purpose multimodal foundation model. By regarding the image as a foreign language (i.e., *Imglish*), we handle texts and images in the same manner without fundamental modeling differences. Consequentially, image-text pairs are utilized as "parallel sentences" in order to learn the alignments between modalities. We also show that the simple yet effective method learns strong transferable representations, achieving remarkable performance on both vision and vision-language tasks. The prominent success demonstrates the superiority of generative pretraining [3, 14].

Third, scaling up the model size and data size universally improves the generalization quality of foundation models, so that we can transfer them to various downstream tasks. We follow the philosophy and scale up the model size to billions of parameters. Moreover, we scale up the pretraining data size while only using publicly accessible resources for academic reproducibility. Although without using any private data, our method outperforms state-of-the-art foundation models that rely on in-house data by a decent margin. In addition, the scaling up benefits from treating images as a foreign language, as we can directly reuse the pipeline developed for large-scale language model pretraining.

In this work, we take advantage of the above ideas to pretrain a general-purpose multimodal foundation model BEIT-3. We pretrain a Multiway Transformer by performing masked data modeling on images, texts, and image-text pairs. During the pretraining procedure, we randomly mask some proportion of text tokens or image patches. The self-supervised learning objective is to recover the original tokens (i.e., text tokens, or visual tokens) given corrupted inputs. The model is general-purpose in the sense that it can be repurposed for various tasks regardless of input modalities or output formats.

As shown in Table 1, BEIT-3 achieves remarkable transfer performance across a broad range of vision and vision-language tasks. We evaluate BEIT-3 on extensive downstream tasks and datasets, i.e., object detection (COCO), instance segmentation (COCO), semantic segmentation (ADE20K), image classification (ImageNet), visual reasoning (NLVR2), visual question answering (VQAv2), image captioning (COCO), and cross-modal retrieval (Flickr30K, COCO). Specifically, our model outperforms previous strong foundation models [1,69,70] despite that we only use public resources for pretraining and finetuning. The model also obtains better results than specialized models. Moreover, BEIT-3 not only performs well on vision-language tasks but also on vision tasks (such as object detection).

# 2. BEIT-3: A General-Purpose Multimodal Foundation Model

BEIT-3 is pretrained by masked data modeling on monomodal and multimodal data, using a shared Multiway Transformer network. The model can be transferred to various vision and vision-language downstream tasks.

## 2.1. Backbone Network: Multiway Transformers

We use Multiway Transformers [62] as the backbone model to encode different modalities. As shown in Figure 1, each Multiway Transformer block consists of a shared self-attention module, and a pool of feed-forward networks (i.e., modality experts) used for different modalities. We route each input token to the experts depending on its modality. Each layer contains a vision expert and a language expert. Moreover, the top three layers have vision-language experts designed for fusion encoders. Refer to Figure 2 (a)(b)(c) for detailed modeling layouts. Using a pool of modality experts encourages the model to capture more modality-specific information. The shared self-attention module learns the alignment between different modalities and enables deep fusion for multimodal (such as vision-language) tasks.

As shown in Figure 2, the unified architecture enables BEIT-3 to support a wide range of downstream tasks. For example, BEIT-3 can be used as an image backbone for various vision tasks, including image classification, object detection, instance segmentation, and semantic segmentation. It can also be finetuned as a dual encoder for efficient image-text retrieval, and a fusion model for multimodal understanding and generation tasks.

## 2.2. Pretraining Task: Masked Data Modeling

We pretrain BEIT-3 via a unified masked data modeling objective on monomodal (i.e., images, and texts) and multimodal data (i.e., image-text pairs).

**Masked Language Modeling** BEIT-3 uses masked language modeling (MLM) to learn language representations from large-scale text-only data. Following BERT [14], we randomly mask 15% tokens of monomodal text data. Each masked token is replaced by a `[MASK]` token 80% of the time, a random token 10% of the time, and kept the original tokens 10% of the time. The pretraining objective is to recover the masked tokens from the corrupted input text.

**Masked Image Modeling** In addition to masked language modeling, we employ masked image modeling (MIM) to learn vision representations from large-scale image data. Following BEIT [3], given an input image, we apply a block-wise masking strategy to mask 40% of image patches. The pretraining objective of MIM is to reconstruct the discrete visual tokens of masked patches. We use the image tokenizer VQ-KD$_{\text{CLIP}}$ proposed in BEIT v2 [43], which is trained under the supervision of CLIP [45], to obtain the discrete tokens as the reconstruction targets.

**Masked Vision-Language Modeling** We introduce masked vision-language modeling (MVLM), which ex-

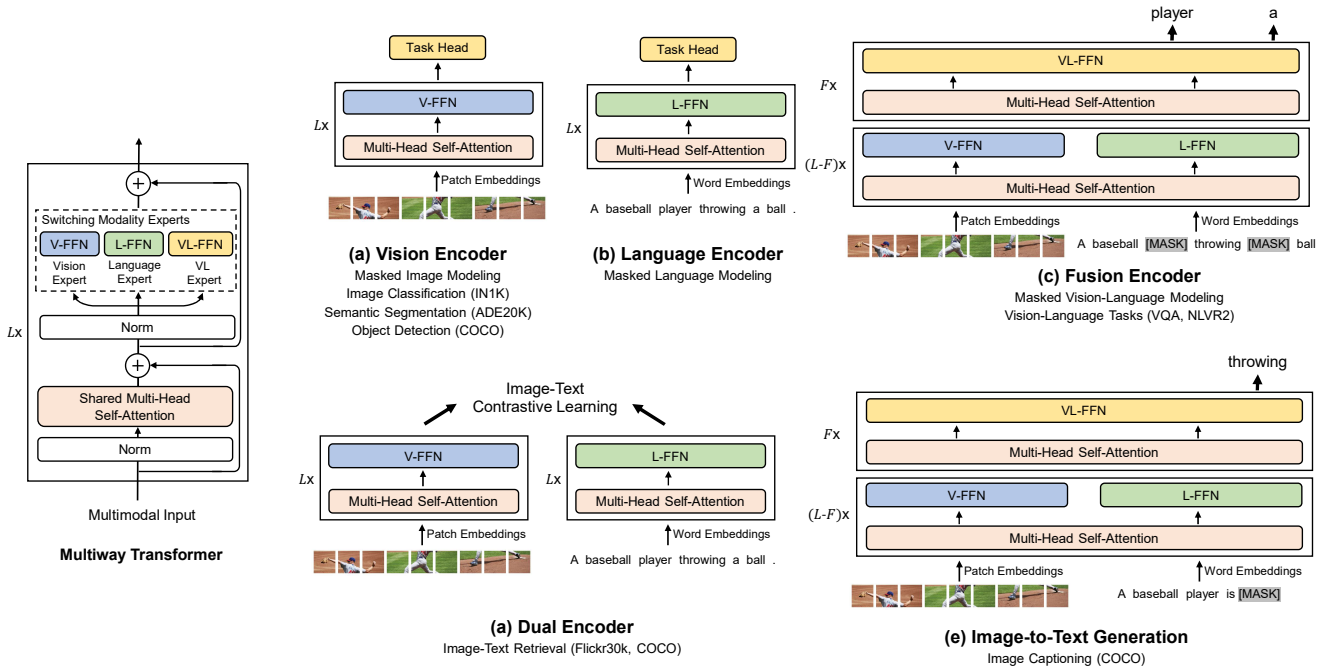


Figure 2. BEIT-3 can be transferred to various vision and vision-language downstream tasks. With a shared Multiway Transformer, we can reuse the model as (a)(b) vision or language encoders; (c) fusion encoders that jointly encode image-text pairs for deep interaction; (d) dual encoders that separately encode modalities for efficient retrieval; (e) sequence-to-sequence learning for image-to-text generation.

tends masked language modeling and masked image modeling to multimodal data. The task aims at recovering masked image patches and text tokens based on visual and linguistic clues. Specifically, we randomly mask text tokens (with 50% mask ratio) as in masked language modeling, and recover the masked text tokens based on the joint image-text representations. In addition, we mask image patches as in MIM and predict their corresponding visual tokens based on the image-text pair. The masking strategy is the same as in masked image modeling. The MVLM task encourages the model to learn alignments between the pairs of image and text.

We only use one pretraining task, which makes the training process scaling-up friendly. In contrast, previous vision-language models [26, 30, 31, 34, 62, 69, 74] usually employ multiple pretraining tasks, such as image-text contrast, image-text matching, and word-patch/region alignment. We show that a much smaller batch size can be used with the mask-then-predict task. In comparison, contrastive-based models [24, 45, 69, 70] usually need a very large batch size for pretraining, which brings more engineering challenges, such as GPU memory cost. For example, CoCa [69] uses 65k batch size, CLIP [45] uses 32k batch size, and Florence [70] uses 24k batch size. In contrast, BEIT-3 enables a much smaller 6k batch size for pretraining. Moreover, unlike the global dependency between examples as in contrastive learning, it is straightforward to implement gradient accumulation for masked data modeling.

## 2.3. Scaling Up: BEIT-3 Pretraining

**Backbone Network** We scale up the model capacity of BEIT-3 to a giant-size Transformer model following the setup of ViT-giant [71]. The giant-size model consists of a 40-layer Multiway Transformer with 1408 hidden size, 6144 intermediate size, and 16 attention heads. All layers contain both vision experts and language experts. Vision-language experts are also employed in the top three Multiway Transformer layers. The self-attention module is shared across different modalities. BEIT-3 giant model consists of 1.9B parameters in total, including 692M parameters for vision experts, 692M for language experts, 52M for vision-language experts, 90M for word embeddings, and 317M for the shared self-attention module. Notice that only vision-related parameters (i.e., comparable size as ViT-giant; about 1B) are activated when the model is used as a vision encoder. Similarly, only text-related weights are used for language tasks.

**Pretraining Data** BEIT-3 is pretrained on both monomodal and multimodal data. For multimodal data, there are about 15M images and 21M image-text pairs collected from five public datasets: Conceptual 12M (CC12M) [8], Conceptual Captions (CC3M) [52], SBU Captions (SBU) [42], COCO [37] and Visual Genome (VG) [27]. Notice that the image tokenizer VQ-KD$_{\text{CLIP}}$ [43] is learned with the guidance from CLIP [45]. Given CLIP

| Task | Dataset | Metric | Previous Systems | BEIT-3 |
|---|---|---|---|---|
| Semantic Segmentation | ADE20K | mIoU | 61.4 [64] | **62.8** (+1.4) |
| Object Detection | COCO | AP | 63.3 [72] | **63.7** (+0.4) |
| Instance Segmentation | COCO | AP | 54.7 [29] | **54.8** (+0.1) |
| Image Classification | ImageNet† | Top-1 acc. | 89.0 [64] | **89.6** (+0.6) |
| Visual Reasoning | NLVR2 | Acc. | 87.0 [69] | **92.6** (+5.6) |
| Visual QA | VQAv2 | VQA acc. | 82.3 [69] | **84.0** (+1.7) |
| Image Captioning | COCO‡ | CIDEr | 145.3 [61] | **147.6** (+2.3) |
| Finetuned Retrieval | COCO | R@1 | 72.5 [70] | **76.0** (+3.5) |
| Finetuned Retrieval | Flickr30K | R@1 | 92.6 [70] | **94.2** (+1.6) |
| Zero-shot Retrieval | Flickr30K | R@1 | 86.5 [69] | **88.2** (+1.7) |

Table 1. Overview of BEIT-3 results on various vision and vision-language benchmarks. We compare with previous strong models, including FD-SwinV2 [64], DINO [72], Mask DINO [29], FD-CLIP [64], CoCa [69], OFA [61], Florence [70]. The comparison models are state-of-the-art when we collect the results (timestamp: 08/22/2022). We report the average of top-1 image-to-text and text-to-image results for retrieval tasks. "†" indicates ImageNet results only using publicly accessible resources. "‡" indicates image captioning results without CIDEr optimization.

is trained with 400M image-text pairs, we note that our model also indirectly touches these data. For monomodal data, we use 14M images from ImageNet-21K and 160GB text corpora [5] from English Wikipedia, BookCorpus [76], OpenWebText[1], CC-News [38], and Stories [57].

**Pretraining Settings** We pretrain the model for 1M steps. Each batch contains 6144 samples in total, including 2048 images, 2048 texts, and 2048 image-text pairs. The batch size is much smaller than contrastive models [24, 45, 69]. BEIT-3 giant model uses $14 \times 14$ patch size and is pretrained at resolution $224 \times 224$. We use the same image augmentation as in BEIT [3], including random resized cropping, horizontal flipping, and color jittering [66]. A SentencePiece tokenizer [28] with 64k vocab size is employed to tokenize the text data. We use the AdamW [40] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon =$1e-6 for optimization. We use a cosine learning rate decay scheduler with a peak learning rate of 1e-3 and a linear warmup of 10k steps. The weight decay is 0.05. Stochastic depth [20] with a rate of 0.1 is used. The pretraining process takes about two weeks using 256 A100 40GB GPUs.

## 3. Experiments

We extensively evaluate BEIT-3 on major public benchmarks for both vision-language and vision tasks. Table 1 presents the overview of the results. BEIT-3 obtains remarkable performance on a wide range of vision and vision-language tasks.

[1] http://skylion007.github.io/OpenWebTextCorpus

| Model | VQAv2 | | NLVR2 | |
|---|---|---|---|---|
| | test-dev | test-std | dev | test-P |
| Oscar [34] | 73.61 | 73.82 | 79.12 | 80.37 |
| VinVL [74] | 76.52 | 76.60 | 82.67 | 83.98 |
| ALBEF [31] | 75.84 | 76.04 | 82.55 | 83.14 |
| BLIP [30] | 78.25 | 78.32 | 82.15 | 82.24 |
| SimVLM [63] | 80.03 | 80.34 | 84.53 | 85.15 |
| Florence [70] | 80.16 | 80.36 | - | - |
| OFA [61] | 82.00 | 82.00 | - | - |
| Flamingo [1] | 82.00 | 82.10 | - | - |
| CoCa [69] | 82.30 | 82.30 | 86.10 | 87.00 |
| **BEIT-3** | **84.19** | **84.03** | **91.51** | **92.58** |

Table 2. Results of visual question answering and visual reasoning tasks. We report *vqa-score* on VQAv2 test-dev and test-standard (test-std) splits, accuracy for NLVR2 development set (dev) and public test set (test-P).

### 3.1. Vision-Language Downstream Tasks

We evaluate the capabilities of BEIT-3 on the widely used vision-language understanding and generation benchmarks, including visual question answering [18], visual reasoning [55], image-text retrieval [37, 44], and image captioning [37].

**Visual Question Answering (VQA)** The task requires the model to answer natural language questions about input images. Following previous work [2, 26, 74], we conduct finetuning experiments on the VQA v2.0 dataset [18] and formulate the task as a classification problem. The model is trained to predict answers from the 3129 most frequent answer candidates in the training set. BEIT-3 is finetuned as a fusion encoder to model deep interactions of images and questions for the VQA task. We concatenate the embeddings of a given question and an image, and then feed the input embeddings into Multiway Transformers to jointly encode the image-question pair. The final pooled output is fed into a classifier layer to predict the answer. The results are reported in Table 2, BEIT-3 outperforms previous models by a large margin (more than 1.7 points), achieving 84.03 with a single model.

**Visual Reasoning** The task needs models to perform joint reasoning about images and natural language descriptions. We evaluate the model on the popular NLVR2 [55] benchmark, which is to determine whether a textual description is true about a pair of images. Following previous work [26, 74], we construct two image-text pairs based on the triplet input. We finetune BEIT-3 as a fusion encoder to jointly encode the image-text pairs. The final pooled outputs of the two pairs are concatenated and then fed into a classifier layer to predict the label. As shown in Table 2, BEIT-3 achieves prominent results for visual reasoning,

| Model | MSCOCO (5K test set) | | | | | | Flickr30K (1K test set) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Image → Text | | | Text → Image | | | Image → Text | | | Text → Image | | |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| UNITER | 65.7 | 88.6 | 93.8 | 52.9 | 79.9 | 88.0 | 87.3 | 98.0 | 99.2 | 75.6 | 94.1 | 96.8 |
| VinVL | 75.4 | 92.9 | 96.2 | 58.8 | 83.5 | 90.3 | - | - | - | - | - | - |
| ALBEF | 77.6 | 94.3 | 97.2 | 60.7 | 84.3 | 90.5 | 95.9 | 99.8 | **100.0** | 85.6 | 97.5 | 98.9 |
| BLIP | 82.4 | 95.4 | 97.9 | 65.1 | 86.3 | 91.8 | 97.4 | 99.8 | 99.9 | 87.6 | 97.7 | 99.0 |
| ALIGN | 77.0 | 93.5 | 96.9 | 59.9 | 83.3 | 89.8 | 95.3 | 99.8 | **100.0** | 84.9 | 97.4 | 98.6 |
| FILIP | 78.9 | 94.4 | 97.4 | 61.2 | 84.3 | 90.6 | 96.6 | **100.0** | **100.0** | 87.1 | 97.7 | 99.1 |
| Florence | 81.8 | 95.2 | - | 63.2 | 85.7 | - | 97.2 | 99.9 | - | 87.9 | 98.1 | - |
| **BEIT-3** | **84.8** | **96.5** | **98.3** | **67.2** | **87.7** | **92.8** | **98.0** | **100.0** | **100.0** | **90.3** | **98.7** | **99.5** |

(a) Finetuned results on COCO and Flickr30K.

| Model | Flickr30K (1K test set) | | | | | |
|---|---|---|---|---|---|---|
| | Image → Text | | | Text → Image | | |
| | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| FLAVA | 67.7 | 94.0 | - | 65.2 | 89.4 | - |
| CLIP | 88.0 | 98.7 | 99.4 | 68.7 | 90.6 | 95.2 |
| ALIGN | 88.6 | 98.7 | 99.7 | 75.7 | 93.8 | 96.8 |
| FILIP | 89.8 | 99.2 | 99.8 | 75.0 | 93.4 | 96.3 |
| Florence | 90.9 | 99.1 | - | 76.7 | 93.6 | - |
| Flamingo | 89.3 | 98.8 | 99.7 | 79.5 | 95.3 | **97.9** |
| CoCa | 92.5 | 99.5 | 99.9 | 80.4 | **95.7** | 97.7 |
| **BEIT-3** | **94.9** | **99.9** | **100.0** | **81.5** | 95.6 | 97.8 |

(b) Zero-shot results on Flickr30K.

Table 3. Finetuning and zero-shot results of image-to-text and text-to-image retrieval. UNITER [9] and VinVL [74] are fusion-encoder models. ALBEF [31] and BLIP [30] first obtain top-$k$ candidates using their dual encoders and then use fusion encoders to rerank the candidates. BEIT-3 and the other models [1, 24, 45, 53, 68–70] are dual-encoder models. Notice that dual-encoder models are more efficient than fusion-encoder-based models for retrieval tasks because of representation caching.

outperforming CoCa by about 5.6 points. The performance on NLVR2 reaches above 90% for the first time.

**Image-Text Retrieval** The task is to measure the similarity between images and texts. There are two directions depending on the modality of the retrieved target: image-to-text retrieval, and text-to-image retrieval. Two popular retrieval benchmarks, i.e., COCO [37], and Flickr30K [44], are used to evaluate the model. Following previous work [26, 74], we use the Karpathy split [25] for the two benchmarks. BEIT-3 is finetuned as a dual encoder for efficient image-text retrieval. Dual-encoder models separately encode images and texts to obtain their representations. Then we calculate the cosine similarity scores of these representations. Dual-encoder models are more efficient than fusion-encoder models. Because they do not have to jointly encode all possible image-text pairs.

We directly finetune BEIT-3 on COCO and Flickr30K, although the model is not pretrained with image-text contrastive loss. Surprisingly, BEIT-3 outperforms previous strong models only using a small amount of contrastive training. The results demonstrate that BEIT-3 effectively learns alignments between images and texts via masked data modeling. In order to improve the performance, we perform intermediate finetuning with an image-text contrastive objective on the pretraining image-text pairs. We finetune the model with much fewer steps than pretraining. Then we use the model to evaluate zero-shot and finetuned image-text retrieval. The finetuned results are reported in Table 3a, dual-encoder BEIT-3 outperforms prior models by a large margin, achieving 3.0/4.0 absolute improvement on COCO top-1 image-to-text/text-to-image retrieval, and 0.8/2.4 absolute improvement on Flickr30K top-1 image-to-text/text-to-image retrieval. BEIT-3 also significantly outperforms fusion-encoder-based models, which require more computation cost than dual-encoder models for inference. As shown in Table 3b, BEIT-3 also achieves better performance on Flickr30K zero-shot retrieval.

| Model | COCO Captioning | | | |
|---|---|---|---|---|
| | BLEU@4 | METEOR | CIDEr | SPICE |
| Oscar [34] | 37.4 | 30.7 | 127.8 | 23.5 |
| VinVL [74] | 38.5 | 30.4 | 130.8 | 23.4 |
| BLIP [30] | 40.4 | - | 136.7 | - |
| SimVLM [63] | 40.6 | 33.7 | 143.3 | **25.4** |
| OFA [61] | 43.9 | 31.8 | 145.3 | 24.8 |
| Flamingo [1] | - | - | 138.1 | - |
| CoCa [69] | 40.9 | **33.9** | 143.6 | 24.7 |
| **BEIT-3** | **44.1** | 32.4 | **147.6** | **25.4** |

Table 4. Results of COCO image captioning. We report BLEU@4, METEOR, CIDEr, and SPICE on the Karpathy test split. For simplicity, we report results without using CIDEr optimization.

**Image Captioning** The task aims to generate a natural language caption for the given image. We use the COCO [37] benchmark, finetune and evaluate the model on Karpathy split [25]. Following UNILM [15] and s2s-ft [4], BEIT-3 is used as a conditional generation model via masked finetuning. To be more specific, a special self-attention mask is employed for the image captioning task. Image tokens (i.e., image patches) can only attend to each other bidirectionally within the image sequence. Tokens of the caption can attention to image tokens, their leftward caption tokens, and themselves. During finetuning, we randomly mask some percentage of caption tokens. The model is trained to recover these tokens based on the clues of the image and its leftward caption context. We also mask the special boundary token `[SEP]` to help the model learn to terminate the generation. For simplicity, BEIT-3 is trained with simple cross-entropy loss, without using CIDEr optimization. During inference, we generate the caption tokens

| Model | Maximum Image Size | COCO test-dev $AP^{box}$ | COCO test-dev $AP^{mask}$ |
|---|---|---|---|
| DyHead [12] | 2000 | 60.6 | - |
| Soft Teacher [67] | - | 61.3 | 53.0 |
| GLIP [33] | - | 61.5 | - |
| GLIPv2 [73] | - | 62.4 | - |
| Florence [70] | 2500 | 62.4 | - |
| SwinV2-G [39] | 1536 | 63.1 | 54.4 |
| Mask DINO [29] | 1280 | - | 54.7 |
| DINO [72] | 2000 | 63.3 | - |
| **BEIT-3** | 1280 | **63.7** | **54.8** |

Table 5. Results of object detection and instance segmentation on COCO benchmark. BEIT-3 uses Cascade Mask R-CNN [7] as the detection head. Our results are reported with multi-scale evaluation. We report the maximum image size used for training. DyHead [12] generates pseudo labels on ImageNet and use the data as extra OD data. Florence [70] uses FLOD-9M and GLIP [33, 73] is trained with FourODs. The other models use Object365 as the extra OD data. FLOD-9M and FourODs also contain Object365. The results of the comparison systems are from the paperswithcode.com leaderboard (timestamp: 08/22/2022).

one by one in an autoregressive manner. Table 4 presents the results on COCO captioning. BEIT-3 achieves better results compared with previous image captioning models. The results demonstrate the superiority of BEIT-3 for vision-language generation.

### 3.2. Vision Downstream Tasks

In addition to vision-language downstream tasks, BEIT-3 can be transferred to a wide range of vision downstream tasks, including object detection, instance segmentation, semantic segmentation, and image classification. Notice that only vision-related parameters are activated when BEIT-3 is used as a vision encoder. So the number of effective parameters is comparable to ViT-giant [71], i.e., the effective model size is about 1B.

**Object Detection and Instance Segmentation** We conduct finetuning experiments on the COCO 2017 benchmark [37], which consists of 118k training, 5k validation, and 20k test-dev images. We use BEIT-3 as the backbone and follow ViTDet [35], including a simple feature pyramid and window attention, for the object detection and instance segmentation tasks. Following common practices [39, 72], we first conduct intermediate finetuning on the Objects365 [51] dataset. Then we finetune the model on the COCO dataset. Soft-NMS [6] is used during inference. Table 5 compares BEIT-3 with previous strong models on COCO object detection and instance segmentation. BEIT-3 achieves the best results on the COCO test-dev set with a smaller image size used for finetuning, reaching up to 63.7 box AP and 54.8 mask AP.

| Model | Crop Size | ADE20K mIoU | ADE20K +MS |
|---|---|---|---|
| HorNet [48] | $640^2$ | 57.5 | 57.9 |
| SeMask [23] | $640^2$ | 57.0 | 58.3 |
| SwinV2-G [39] | $896^2$ | 59.3 | 59.9 |
| ViT-Adapter [10] | $896^2$ | 59.4 | 60.5 |
| Mask DINO [29] | - | 59.5 | 60.8 |
| FD-SwinV2-G [64] | $896^2$ | - | 61.4 |
| **BEIT-3** | $896^2$ | **62.0** | **62.8** |

Table 6. Results of semantic segmentation on ADE20K. "MS" is short for multi-scale. The results of the comparison systems are from the paperswithcode.com leaderboard (timestamp: 08/22/2022).

| Model | Extra Data | Image Size | ImageNet |
|---|---|---|---|
| *With extra **private** image-tag data* | | | |
| SwinV2-G [39] | IN-22K-ext-70M | $640^2$ | 90.2 |
| ViT-G [71] | JFT-3B | $518^2$ | 90.5 |
| CoAtNet-7 [13] | JFT-3B | $512^2$ | 90.9 |
| Model Soups [65] | JFT-3B | $500^2$ | 91.0 |
| CoCa [69] | JFT-3B | $576^2$ | 91.0 |
| *With only **public** image-tag data* | | | |
| BEIT [3] | IN-21K | $512^2$ | 88.6 |
| CoAtNet-4 [13] | IN-21K | $512^2$ | 88.6 |
| MaxViT [58] | IN-21K | $512^2$ | 88.7 |
| MViTv2 [36] | IN-21K | $512^2$ | 88.8 |
| FD-CLIP [64] | IN-21K | $336^2$ | 89.0 |
| **BEIT-3** | IN-21K | $336^2$ | **89.6** |

Table 7. Top-1 accuracy on ImageNet-1K.

**Semantic Segmentation** Semantic segmentation aims to predict the label for each pixel of the given image. We evaluate BEIT-3 on the challenging ADE20K dataset [75], which includes 150 semantic categories. ADE20K contains 20k images for training and 2k images for validation. We directly follow the task transfer settings of ViT-Adapter [10]. We use a dense prediction task adapter and employ Mask2Former [11] as the segmentation framework. As shown in Table 6, BEIT-3 achieves 62.8 mIoU, outperforming FD-SwinV2 [64] giant model with 3B parameters by 1.4 points. It shows that BEIT-3 achieves superior performance on the dense prediction task.

**Image Classification** We evaluate the model on ImageNet-1K [50], which contains 1.28M training images and 50k validation images in 1k classes. Rather than appending a task layer to the vision encoder [3, 16], we formulate the task as an image-to-text retrieval task. We use the category names as texts to construct image-text pairs. BEIT-3 is trained as a dual encoder to find the most relevant label for an image. During inference, we first compute the feature embeddings of possible class names and the feature embedding of the image. Their

cosine similarity scores are then calculated to predict the most probable label for each image. Table 7 reports the results on ImageNet-1K. We first perform intermediate finetuning on ImageNet-21K, then we train the model on ImageNet-1K. For a fair comparison, we compare with the previous models only using public image-tag data. BEIT-3 outperforms prior models when only using public image-tag data.

### 3.3. Ablation Studies

We conduct ablation studies on base-size models, having 12-layer Multiway Transformer blocks with 768 hidden size and 3072 intermediate size. The base-size models use $16 \times 16$ patch size and are trained at resolution $224 \times 224$. Most settings and hyperparameters are kept the same as in Section 2.3. We use multimodal data including CC3M, SBU, COCO, and VG to pretrain the model. The monomodal data include ImageNet-21K and 16GB text corpora from English Wikipedia and BookCorpus. Notice that we use the same text corpora as BERT [14] so that we can directly compare the language-only performance in Table 9. The models are pretrained for 200K steps with 2e-3 peak learning rate and 6144 batch size.

**Backbone Architecture** We study the effects of different model architectures. Table 8a shows that Multiway Transformers perform better than standard Transformers on three benchmarks. Modality experts introduced in Multiway Transformers effectively capture modality-specific information and improve performance.

**Masking Strategy in MVLM** We compare two masking strategies for MVLM, i.e., joint masking, and separate masking. Specifically, for joint masking, we simultaneously mask image patches and text tokens for the same input image-text pair. In contrast, for separate masking, given an input pair, we randomly mask tokens of one modality (image or text) while keeping tokens of another modality unmasked. As shown in Table 8b, separate masking outperforms joint masking and learns the alignment of images and texts more effectively.

**Monomodal and Multimodal Data** We analyze the effects of monomodal and multimodal data in Table 8c. Experimental results indicate that monomodal and multimodal data positively contribute to performance. Using both types of pretraining data achieves the best results.

**Image Reconstruction Target** We compare different targets used for image reconstruction. As shown in Table 8d, VQ-KD$_{\text{CLIP}}$ [43] performs better than the DALL-E [47] tokenizer used in BEIT [3] and per-patch-normalized pixels proposed by MAE [19].

**Text Reconstruction** We study the effects of text reconstruction on monomodal and multimodal data. As shown in Table 8e, the text reconstruction tasks on monomodal and multimodal data bring improvements. Text reconstruction on text corpora learns language representations. Moreover, text reconstruction on multimodal data encourages the model to learn cross-modal alignments. In addition, we find that masked language modeling on multimodal data plays a more important role than on text-only data.

**Image Reconstruction** Table 8f presents the ablation study of masked image modeling on monomodal and multimodal data. The results indicate that the image reconstruction tasks on both types of pretraining data improve the results. In contrast to text reconstruction, we find that monomodal data and multimodal data contribute similarly to image reconstruction.

**Language Downstream Tasks** Table 9 shows that our method also achieves competitive performance on language-only tasks. Following previous work [53, 63], we conduct experiments on the widely used GLUE [60] benchmark with a base-size model. Compared with previous vision-language pretrained models [9, 32, 41, 45, 53, 54, 56, 63], BEIT-3 achieves better performance. BEIT-3 even outperforms SimVLM [63] trained on a much larger text corpora.

## 4. Related Work

Vision-language pretraining aims to learn multimodal representations from large-scale image-text pairs. Model architecture and pretraining objectives are critical to the effectiveness of vision-language models.

**Vision-Language Architectures** There are two mainstream architectures widely used in previous vision-language pretrained models: *dual-encoder* and *fusion-encoder* models. Dual-encoder model [24, 45] consists of an image encoder and a text encoder. It encodes images and text separately, and then employs cosine similarity to model the interaction of image and text vectors. Dual-encoder models achieve promising results for image-text retrieval tasks with linear time complexity. However, the simple fusion module is not enough to handle complex vision-language understanding tasks such as visual reasoning. Fusion-encoder models employ a complex fusion module with cross-modal attention, to jointly encode images and text. Previous models [34, 41, 54, 74] use an off-the-shelf object detector like Faster R-CNN [49] to obtain image region features. Text features are usually word embeddings or contextual vectors encoded by a text encoder. These image and text features are then jointly encoded

| Transformer | VQA | NLVR2 | F30K |
|---|---|---|---|
| Standard | 76.1 | 80.8 | 82.8 |
| Multiway | **76.8** | **81.4** | **84.4** |

(a) Multiway Transformer improves the performance over the conventional one.

| Strategy | VQA | NLVR2 | F30K |
|---|---|---|---|
| Joint | 75.7 | 79.0 | 83.1 |
| Separate | **76.8** | **81.4** | **84.4** |

(b) Separate masking in MVLM is helpful.

| Mono | Multi | VQA | NLVR2 | F30K |
|---|---|---|---|---|
| ✓ | ✗ | 71.3 | 64.6 | 79.3 |
| ✗ | ✓ | 75.8 | 79.3 | 81.1 |
| ✓ | ✓ | **76.8** | **81.4** | **84.4** |

(c) Whether we conduct masked prediction for monomodal (mono) and multimodal (multi) data.

| Target | VQA | NLVR2 | F30K |
|---|---|---|---|
| DALL-E [47] | 73.2 | 77.7 | 76.6 |
| Pixel (w/ norm) [19] | 73.3 | 77.1 | 75.9 |
| VQ-KD$_{\text{CLIP}}$ [43] | **76.8** | **81.4** | **84.4** |

(d) Targets used for image reconstruction. VQ-KD$_{\text{CLIP}}$ [43] works the best.

| Mono | Multi | VQA | NLVR2 | F30K |
|---|---|---|---|---|
| ✗ | ✗ | 71.5 | 69.3 | 77.8 |
| ✓ | ✗ | 73.2 | 76.4 | 81.3 |
| ✗ | ✓ | 76.5 | 80.6 | 82.7 |
| ✓ | ✓ | **76.8** | **81.4** | **84.4** |

(e) Whether we enable text reconstruction for monomodal (mono) and multimodal (multi) data.

| Mono | Multi | VQA | NLVR2 | F30K |
|---|---|---|---|---|
| ✗ | ✗ | 71.6 | 74.3 | 71.7 |
| ✓ | ✗ | 75.8 | 79.8 | 82.0 |
| ✗ | ✓ | 75.6 | 79.5 | 81.9 |
| ✓ | ✓ | **76.8** | **81.4** | **84.4** |

(f) Whether we enable image reconstruction for monomodal (mono) and multimodal (multi) data.

Table 8. Ablation studies of BEIT-3. We conduct experiments on a base-size model. We report vqa-score on VQA test-dev set, accuracy on NLVR2 dev set, and average of top1 recall of image-to-text and text-to-image retrieval on Flickr30K dev set. The models are finetuned as a dual encoder for Flickr30K. Gray indicates the default setting of BEIT-3.

| **Model** | **SST-2** | **RTE** | **QQP** | **MNLI** | **QNLI** | **Avg** |
|---|---|---|---|---|---|---|
| BERT [14] | 92.5 | 62.5 | 90.6 | 84.4 | 91.0 | 84.2 |
| VisualBERT [32] | 89.4 | 56.6 | 89.4 | 81.6 | 87.0 | 80.8 |
| UNITER [9] | 89.7 | 55.6 | 89.2 | 80.9 | 86.0 | 80.3 |
| VL-BERT [54] | 89.8 | 55.7 | 89.0 | 81.2 | 86.3 | 80.4 |
| ViLBERT [41] | 90.4 | 53.7 | 88.6 | 79.9 | 83.8 | 79.3 |
| LXMERT [56] | 90.2 | 57.2 | 75.3 | 80.4 | 84.2 | 77.5 |
| CLIP [45] | 88.2 | 55.2 | 76.8 | 33.5 | 50.5 | 60.8 |
| SimVLM [63] | 90.9 | 63.9 | 90.4 | 83.4 | 88.6 | 83.4 |
| FLAVA [53] | 90.9 | 57.8 | 90.4 | 80.3 | 87.3 | 81.3 |
| BEIT-3$_{\text{base}}$ | **92.6** | **66.5** | **91.0** | **83.8** | **90.8** | **84.9** |

Table 9. Finetuning results of base-size models on the dev set of the GLUE [60] benchmark. Comparison results are taken from [22]. Our numbers are averaged over three runs with different seeds. We also report the average results (Avg) across datasets.

by the fusion module, which usually adopts a multi-layer Transformer network. Recently, Pixel-BERT [21] and ALBEF [31] use CNN/vision Transformer to encode images and remove object detectors. ViLT [26] uses a shared Transformer network to jointly encode image patches and word embeddings. Fusion-encoder models achieve superior performance on vision-language understanding tasks such as vision reasoning. But it requires quadratic time complexity for retrieval tasks, which leads to a much slower inference speed than the dual-encoder models. VLMO [62] unifies dual-encoder and fusion-encoder models and introduces Multiway Transformers to encode various modalities within a shared Transformer block. In this work, we adopt the Multiway Transformers as the backbone network given its simplicity and flexibility. BEIT-3 can also be finetuned as a dual-encoder model or fusion-encoder model.

**Pretraining Tasks** Many multimodal pretraining objectives have been proposed, including image-text contrastive learning [24, 45, 68], image-text matching [26, 31, 56, 62], masked language modeling [17, 26, 34, 54, 56] or prefix language modeling [63], masked region classification [56], word-patch/region alignment [9, 26]. SimVLM [63] trains the vision-language model using prefix language modeling on image-text pairs and text-only data. FLAVA [53] combines masked image modeling with masked language modeling, image-text contrast and matching based on a fusion-encoder model. Masked image modeling and masked language modeling are applied to the monomodal encoders. Masked multimodal modeling, image-text contrast and matching losses are used for the multimodal encoder. Compared with SimVLM, BEIT-3 introduces richer visual supervision via masked image modeling and masked vision-language modeling. Different from FLAVA, we use a shared Multiway Transformer for different modalities and adopt one-stage training from scratch.

## 5. Conclusion

In this paper, we present BEIT-3, a general-purpose multimodal foundation model, which achieves remarkable performance across a wide range of vision and vision-language benchmarks. The key idea of BEIT-3 is that image can be modeled as a foreign language, so that we can conduct masked "language" modeling over images, texts, and image-text pairs in a unified way. We also demonstrate that Multiway Transformers can effectively model different vision and vision-language tasks, making it an intriguing option for general-purpose modeling. BEIT-3 is simple and effective, and is a promising direction for scaling up multimodal foundation models. For future work, we are working on pretraining multilingual BEIT-3 and including more modalities (e.g., audio) in BEIT-3 to facilitate the cross-lingual and cross-modality transfer, and advance the big convergence of large-scale pretraining across tasks, languages, and modalities.

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198, 2022. 2, 4, 5

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society, 2018. 4

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 1, 2, 4, 6, 7

[4] Hangbo Bao, Li Dong, Wenhui Wang, Nan Yang, and Furu Wei. s2s-ft: Fine-tuning pretrained transformer encoders for sequence-to-sequence learning. *CoRR*, abs/2110.13640, 2021. 5

[5] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR, 2020. 4

[6] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms - improving object detection with one line of code. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5562–5570. IEEE Computer Society, 2017. 6

[7] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: high quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5):1483–1498, 2021. 6

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE, 2021. 3

[9] 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 Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 5, 7, 8

[10] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *CoRR*, abs/2205.08534, 2022. 6

[11] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *CoRR*, abs/2112.01527, 2021. 6

[12] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7373–7382. Computer Vision Foundation / IEEE, 2021. 6

[13] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 3965–3977, 2021. 6

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1, 2, 7, 8

[15] 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 *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054, 2019. 1, 5

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *preprint arXiv:2010.11929*, 2020. 1, 6

[17] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In 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. 8

[18] 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. 4

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021. 7, 8

[20] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pages 646–661. Springer, 2016. 4

[21] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020. 8

[22] 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*, pages 2189–2196, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. 8

[23] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masking transformer backbones for effective semantic segmentation. *arXiv*, 2021. 6

[24] 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. 3, 4, 5, 7, 8

[25] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society, 2015. 5

[26] 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. 1, 3, 4, 5, 8

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 3

[28] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. 4

[29] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask DINO: towards A unified transformer-based framework for object detection and segmentation. *CoRR*, abs/2206.02777, 2022. 4, 6

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022. 3, 4, 5

[31] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651, 2021. 3, 4, 5, 8

[32] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 7, 8

[33] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *CoRR*, abs/2112.03857, 2021. 6

[34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. 3, 4, 5, 7, 8

[35] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *CoRR*, abs/2203.16527, 2022. 6

[36] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 6

[37] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture*

*Notes in Computer Science*, pages 740–755. Springer, 2014. 3, 4, 5, 6

[38] 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. *CoRR*, abs/1907.11692, 2019. 4

[39] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: scaling up capacity and resolution. *CoRR*, abs/2111.09883, 2021. 6

[40] 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. 4

[41] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. 7, 8

[42] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151, 2011. 3

[43] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *CoRR*, abs/2208.06366, 2022. 1, 2, 3, 7, 8

[44] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society, 2015. 4, 5

[45] 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. 1, 2, 3, 4, 5, 7, 8

[46] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 1

[47] A. Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 7, 8

[48] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu. HorNet: Efficient high-order spatial interactions with recursive gated convolutions. *ArXiv*, abs/2207.14284, 2022. 6

[49] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. 7

[50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[51] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8429–8438. IEEE, 2019. 6

[52] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018. 3

[53] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. *CoRR*, abs/2112.04482, 2021. 5, 7, 8

[54] 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. 7, 8

[55] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics, 2019. 4

[56] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019. 7, 8

[57] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847, 2018. 4

[58] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit:

Multi-axis vision transformer. *CoRR*, abs/2204.01697, 2022. 6

[59] 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, 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. 1

[60] 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. 7, 8

[61] 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. *CoRR*, abs/2202.03052, 2022. 4, 5

[62] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. *CoRR*, abs/2111.02358, 2021. 1, 2, 3, 8

[63] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904, 2021. 1, 4, 5, 7, 8

[64] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *CoRR*, abs/2205.14141, 2022. 4, 6

[65] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR, 2022. 6

[66] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. Computer Vision Foundation / IEEE Computer Society, 2018. 4

[67] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3040–3049. IEEE, 2021. 6

[68] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. FILIP: fine-grained interactive language-image pre-training. *CoRR*, abs/2111.07783, 2021. 5, 8

[69] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *CoRR*, abs/2205.01917, 2022. 1, 2, 3, 4, 5, 6

[70] 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. *CoRR*, abs/2111.11432, 2021. 2, 3, 4, 5, 6

[71] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021. 3, 6

[72] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *CoRR*, abs/2203.03605, 2022. 4, 6

[73] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *CoRR*, abs/2206.05836, 2022. 6

[74] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. 3, 4, 5, 7

[75] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *Int. J. Comput. Vis.*, 127(3):302–321, 2019. 6

[76] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 4