Toward Building General Foundation Models for Language, Vision, and Vision-Language Understanding Tasks

Xinsong Zhang* ByteDance Research
Yan Zeng ByteDance Research
Jipeng Zhang HKUST
Hang Li ByteDance Research

Abstract

Foundation models or pre-trained models have substantially improved the performance of various language, vision, and vision-language understanding tasks. However, existing foundation models can only perform the best in one type of tasks, namely language, vision, or vision-language. It is still an open question whether it is possible to construct a general foundation model performing the best for all the understanding tasks. In this paper, we propose a new method for training the general foundation model, X-FM (the X-Foundation Model). X-FM has one language encoder, one vision encoder, and one fusion encoder, as well as a new training method. The training method includes two new techniques for learning X-FM from text, image, and image-text pair data. One is to stop gradients from the vision-language training when learning the language encoder. The other is to leverage the vision-language training to guide the learning of the vision encoder. Extensive experiments on benchmark datasets show that X-FM can significantly outperform existing general foundation models and perform better than or comparable to existing foundation models specifically for language, vision, or vision-language understanding. Code and pre-trained models are released at https://github.com/zhangxinsong-nlp/XFM.

1 Introduction

With the enormous power of foundation models, also known as pre-trained models, remarkable performance gains have recently been achieved in a variety of understanding tasks in natural language processing (NLP), computer vision (CV), and other fields (Devlin et al., 2019; Liu et al., 2019; Lewis et al., 2020; Raffel et al., 2020; Brown et al., 2020; Dosovitskiy et al., 2021; He et al., 2022; Bao et al., 2021; Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020, 2021a; Zeng et al., 2021, 2022) . Foundation models are usually equipped with Transformer (Vaswani et al., 2017) as the backbone, pre-trained with a tremendous amount of unlabeled data, and then fine-tuned with small amounts of labeled data in downstream tasks. The strong representation ability of the model, the massive amount of data, and the effective means of training make the foundation models powerful for successfully solving the tasks of vision, language, and vision-language (Li et al., 2021b,c; Singh et al., 2021; Wang et al., 2021b, 2022b; Diao et al., 2022; Wang et al., 2022a).

The state-of-the-art foundation models usually work the best for one type of tasks, namely language, vision, and vision-language. For example, RoBERTa (Liu et al., 2019), BEiTv2 (Peng et al., 2022), and X-VLM (Zeng et al., 2021, 2022) are language, vision, and vision-language foundation models respectively, and can achieve state-of-the-art performances for the specific type of tasks. It is still very challenging, however, to build a general foundation model that can perform the best in all types of tasks. Existing models, such as FLAVA (Singh et al., 2021), OFA (Wang et al., 2022b), DaVinci (Diao et al., 2022) and Uni-Perceiver-MoE (Zhu et al., 2022), are trying to achieve the goal. Their performances are still not satisfactory, however, when compared with the best performing foundation models for the individual types of tasks, as shown in Table 1. Previous work (Bingel and Søgaard, 2017; Wang et al., 2020) also shows that it is difficult to train a general foundation model in a multi-task learning setting that can effectively learn and utilize representations for all types of tasks. The reason is that language, vision, and vision-language are very different in nature, and a simple way of jointly training a model from language, vision, and vision-language data can easily create a suboptimal solution.

To address the challenge, we propose a new

*Correspondence to: <xszhang0320@gmail.com>.

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 551–568
December 6-10, 2023

Methods	Text Tasks GLUE	Vision Tasks ImageNet	Multi-modal Tasks (MSCOCO Retriveal & VQA & NLVR) Zero-Shot		Fine-Tune			
	MNLI	FT/LE	TR	IR	TR	IR	VQA	NLVR
Foundation models specifically for language, vision, or vision-language understanding								
RoBERTa (Liu et al., 2019)	87.6	–	–	–	–	–	–	–
BEiTv2 (Peng et al., 2022)	–	85.5/80.1	–	–	–	–	–	–
X-VLM (Zeng et al., 2021)	–	–	70.8/92.1/96.5	55.6/82.7/90.0	80.4/95.5/98.2	63.1/85.7/91.6	78.1	84.8
X^2-VLM (Zeng et al., 2022)	–	–	–	–	83.5/96.3/98.5	66.2/87.1/92.2	80.4	87.0
General foundation models								
UNIMO-2 (Li et al., 2021c)	87.5	80.8/-	–	–	–	–	76.3	–
SimVLM (Wang et al., 2021c)	83.4	-/80.6	–	–	–	–	77.9	81.8
FLAVA (Singh et al., 2021)	80.3	-/75.5	42.7/76.8/-	38.4/67.5/-	61.5/82.1/89.6	50.1/74.4/83.2	72.8	–
OFA (Wang et al., 2022b)	84.3	82.2/–	–	–	–	–	78.0	–
DaVinci (Diao et al., 2022)	83.1	83.9/78.8	–	–	–	–	76.3	77.9
OmniVL (Wang et al., 2022a)	–	–	–	–	76.8/93.6/97.3	58.5/82.6/89.5	78.3	–
Uni-Perceiver-MoE (Zhu et al., 2022)	81.5	84.5/–	64.6/–/–	51.6/–/–	70.5/–/–	54.1/–/–	–	–
mPLUG-2_{base} (Xu et al., 2023)	87.6	–/–	–/–/–	–/–/–	81.2/95.2/98.1	65.3/86.9/92.4	79.3	–
X-FM$_{\text{base}}$	**87.7**	**85.5/81.2**	**77.6/94.8/97.7**	**61.1/84.5/90.6**	**84.2/96.4/98.4**	**67.0/87.2/92.4**	**80.5**	**88.4**

Table 1: **Performance comparisons between foundation models.** All results are from *base*-size models. MSCOCO is a cross-modal retrieval task, and IR and TR are image-retrieval and text-retrieval, respectively. MNLI results are average accuracies of MNLI-m and MNLI-mm. For ImageNet1k classification, we report linear evaluation (LE) performance and fine-tuning (FT) performance, respectively. We report R@1/R@5/R@10 for all retrieval tasks at both zero-shot and fine-tune settings. We report the VQA test-dev result and the NLVR test-P result. **bold** denotes the best number across general foundation models. **underline** denotes the best across all models.

method for training general foundation model, and bring in X-FM (X-Foundation Model). X-FM consists of three modular encoders for language (text) encoding, vision (image) encoding, and fusion encoding, as shown in Fig 1. The language encoder, the vision encoder, and the entire model can be used in downstream tasks of language, vision, and vision-language understanding, respectively. The language encoder and the vision encoder follow the implementations of BERT (Devlin et al., 2019) and ViT (Dosovitskiy et al., 2021), respectively. Note that X-FM do not include any extra parameters for language and vision tasks. The fusion encoder has the same architecture as BERT except that there is a cross-attention sub-layer after the self-attention sub-layer in each Transformer layer.

In learning of X-FM, the language encoder, vision encoder, and fusion encoder are jointly trained with text data, image data, and image-text pair data as input. Given the text data, we train the language encoder by masked language modeling (MLM). Given the image data, we train the vision encoder by masked image modeling (MIM). Given the image-text pair data, we train the fusion encoder by image text matching (ITM), image-conditioned masked language modeling (IMLM), bounding box prediction (BBP), also train the vision encoder and the language encoder by image-text contrastive learning (ITC). (See Fig 1.)

The essential thinking of our learning method is that language is more abstract than vision, and there is an asymmetric relationship between language and vision. Therefore, we separate the learning of the three encoders. The language encoder is trained mainly from text data and is isolated from the training of the fusion encoder. The vision encoder is simultaneously trained from image data and image-text pair data, guided by the vision-language training. The fusion encoder is trained from image-text pair data.

Our learning method includes two new techniques. One technique is to stop gradients from the vision-language training when learning the language encoder. The gradient flow is stopped from the fusion encoder to the language encoder in training, while the activation flow from the language encoder to the fusion encoder is as usual. As a result, the language encoder is not affected by training of the fusion encoder with image-text pair data. Moreover, the training of the fusion encoder concentrates on learning the alignments between language and vision features.

The other technique is to leverage the vision-language training to guide the learning of the vision encoder with masked image modeling (MIM). In MIM, the masked image is compared with the original image by the differences between the predicted representations and target representations at the masked and `[CLS]` positions. The vision encoder creates both the predicated and target rep-

resentations, while there is gradient flow from the predicted representations but no gradient flow from the target representations. The vision encoder can create the target representations because it is also trained in the vision-language training.

We conduct experiments on a variety of twenty-three tasks of language, vision, and vision-language understanding. X-FM can outperform other general foundation models by a large margin and can even achieve better or comparable performance than SOTA foundation models specifically designed for language, vision, or vision-language understanding tasks, as shown in Table 1.

Our contribution is as follows.

(1) We address the problem of how to build a general foundation model that can perform the best for all the understanding tasks of language, vision, and vision-language.

(2) We propose a general foundation model, X-FM, which can achieve better or competitive performance on both unimodal understanding tasks and multi-modal understanding tasks through two training techniques.

(3) The stop gradient technique is useful in maintaining text understanding capability and enhancing multi-modal understanding capability at the same time. We also propose a convenient method for mask image modeling with multi-modal learning. The technique can enhance both vision and multi-modal understanding.

2 Related Work

Following the success of language model pre-training (Devlin et al., 2019; Liu et al., 2019; Sun et al., 2019; Joshi et al., 2020; Clark et al., 2020; Lan et al., 2020; Zhang et al., 2020; He et al., 2021), vision pre-training (Dosovitskiy et al., 2021; He et al., 2022; Bao et al., 2021; Peng et al., 2022; Wei et al., 2022a) and vision-language pre-training (Radford et al., 2021; Jia et al., 2021; Li et al., 2021a, 2022; Yuan et al., 2021; Wang et al., 2021a; Bao et al., 2022; Zeng et al., 2021, 2022) with Transformer as the backbone have also made significant progress recently, pushing the state-of-the-art of various understanding tasks of language, vision, and vision-language.

Recently, the fact that Transformer can model multi-modal data within a single architecture has inspired research to develop general foundation models that can solve language, vision, and vision-language tasks at the same time. UNIMO (Li et al., 2021b,c) jointly learns vision representations, language representations, and vision-language alignments in a shared space. FLAVA (Singh et al., 2021) performs pre-training with masked uni-modal and multi-modal modeling objectives. OFA (Wang et al., 2022c) formulates vision-language tasks as sequence-to-sequence (seq2seq) problems and pre-trains a seq2seq model in multi-task learning. SimVLM (Wang et al., 2021c) pre-trains a seq2seq model with a single objective of language generation (prefix language modeling). DaVinci (Diao et al., 2022) combines prefix language modeling and prefix image modeling to learn a general foundation model for a wide range of tasks. Uni-Perceiver (Zhu et al., 2021, 2022) builds a unified perception architecture that processes various modalities and tasks with a single Transformer and shared parameters.

Previous studies on general foundation models have shown that different capabilities can be established with only one model. Still, few studies demonstrate that the best performance can be achieved in all tasks with one model. In this paper, we propose a new method for training general foundation model and show that it can perform the best for all the understanding tasks of language, vision, and vision-language. We compare our model extensively with recent general foundation models on multiple dimensions, as shown in Appendix A.

Several super-large foundation models (over 1B parameters) are proposed recently, most of which are trained on super-large *in-house* datasets (over 900M image-text pairs). The authors do not report results at the base (about 300M parameters) scale on *public* datasets, which we consider in this paper. CoCa (Yu et al., 2022) pre-trains an image-text sequence-to-sequence model with contrastive loss and captioning loss. BEiT-3 (Wang et al., 2022d) uses a multi-way Transformer and a unified objective of masked "language" modeling for learning from image, text, and image-text pair data. Flamingo (Alayrac et al., 2022) makes use of a large language model in vision-language pre-training to solve the "in-context learning" problem for vision-language tasks. PaLI (Chen et al., 2022) jointly scales up the vision encoder and language encoder to cover a variety of language, vision, vision-language, and multilingual tasks.

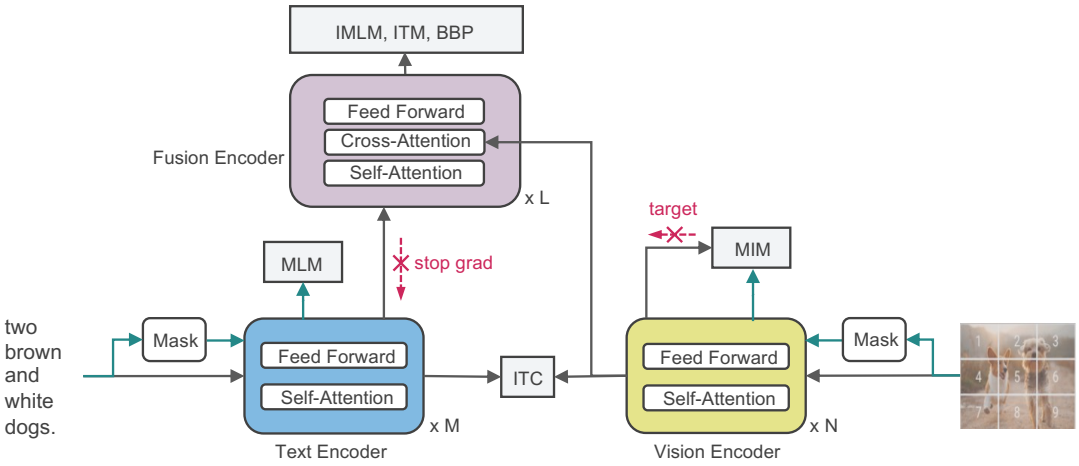

Figure 1: **The architecture and pre-training process of X-FM, a Transformer-based general foundation model.** Given a text, we learn the language encoder by MLM. Given an image, we learn the vision encoder by MIM. Given an image-text pair, we learn the fusion encoder by BBP, ITM, IMLM and ITC, and further learn the vision encoder by MIM. The gradients of BBP, ITM, and IMLM are stopped from the fusion encoder to the language encoder. The vision encoder is trained by MIM with both the image-text pair data and the image data. M, N and L denote numbers of encoder layers.

3 Method

3.1 Model Architecture and Training Process

We propose a new method for training general foundation model and bring in X-FM, having a language encoder, a vision encoder, and a fusion encoder, shown as Fig 1. The architectures of language encoder, vision encoder and fusion encoder are following precious works (Devlin et al., 2019; Dosovitskiy et al., 2021; Li et al., 2021a). We propose a new method for training general foundation model. Text, image, and image-text pair data are used as input to train X-FM. The language encoder is trained by masked language modeling (MLM) and image text contrastive learning (ITC). The vision encoder is trained by masked image modeling (MIM) and ITC. The fusion encoder is trained by image text matching (ITM), image-conditioned masked language modeling (IMLM), and bounding box prediction (BBP). There are two new techniques developed for the training.

Stop Gradient. We stop gradients from the vision-language training when learning the language encoder. Specifically, when the fusion encoder is trained with image-text pair data by ITM, IMLM, and BBP, there are forward flows (activations) from the language encoder to the fusion encoder, but there are no backward flows (gradients) from the fusion encoder to the language encoder. In this way, the language encoder is only trained with text data by MLM and with image-text pair data by ITC. The former helps the language encoder to learn text representations, and the latter helps to make alignments between text representations and image representations. Meanwhile, the training of the fusion encoder is performed separately with the focus of learning cross-modal alignments.

Vision-Language Guided Masked Image Modeling. The training of vision encoder by MIM is carried out as follows. The image data is first masked and then predicted by the vision encoder. The differences between predicted representations and 'target' representations at masked positions and `[CLS]` position are then measured with MSE (mean squared error) loss. The target representations are obtained from the same image data (without masking) by the vision encoder. There are no gradients from the target representations in the learning of the vision encoder. The vision encoder can create target representations because it is also trained with image-text pair data. In this way, the vision encoder is trained by both the cross-modal objectives (ITC, ITM, BBP, IMLM) with image-text pair data and the uni-modal objective (MIM) with image data. The representations obtained from the vision-language training are highly semantic, which is necessary for MIM as demonstrated in previous work (Bao et al., 2021; Peng et al., 2022; Wei et al., 2022a,b).

There are mainly two advantages by exploiting the new MIM technique. First, it is convenient to conduct MIM with the signals from the vision-language training. Note that most previous work for MIM uses an external image tokenizer such as VQ-VAE (Bao et al., 2021; Singh et al., 2021), CLIP (Wei et al., 2022b), and VQ-KL (Peng et al., 2022). Second, the learning of the vision encoder and that of the fusion encoder are mutually enhanced. Once the vision encoder is trained, it is also utilized to train the fusion encoder. Fortunately, image data for training the vision encoder is relatively easy to obtain.

3.2 Pre-training Objectives

We explain six objectives in learning of X-FM. Here, $\mathcal{T}$ represents the distribution of text data, $\mathcal{I}$ represents the distribution of image data, and $\mathcal{D}$ represents the distribution of image-text pair data.

Masked Language Modeling (MLM) We perform MLM on text data to learn the language encoder of X-FM. Specifically we recover the masked tokens in a text by minimizing the cross entropy loss below.

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{T\sim\mathcal{T}}\text{H}(\vec{y}(\bar{T}), \hat{\vec{p}}(\bar{T})) \quad (1)$$

where T denotes a text, $\bar{T}$ denotes the masked text of T, $\hat{\vec{p}}$ denotes the predicted probability vectors of masked tokens of $\bar{T}$, $\vec{y}$ denotes the one-hot vectors representing the original tokens of $\bar{T}$, and H denotes cross-entropy.

Image-Text Contrastive Learning (ITC). We use a contrastive loss as in CLIP (Radford et al., 2021) to learn the alignments between images and texts in ITC. Given a batch of images and texts, we calculate the cosine similarities between all image-text pairs. For each image, there is one text matched and the rest is unmatched. For each text, there is one image matched and the rest is unmatched. The contrastive loss is defined as follows.

$$\mathcal{L}_{\text{itc}} = \frac{1}{2}\mathbb{E}_{(I,T)\sim\mathcal{D}}\big[\text{H}(\vec{y}^{\text{i2t}}(I), \vec{p}^{\text{i2t}}(I)) + \text{H}(\vec{y}^{\text{t2i}}(T), \vec{p}^{\text{t2i}}(T))\big] \quad (2)$$

where (I,T) denotes an image-text pair, $\vec{p}^{\text{i2t}}(I)$ denotes the in-batch image-to-text similarities, $\vec{p}^{\text{t2i}}(T)$ denotes the in-batch text-to-image similarities, $\vec{y}^{\text{i2t}}(I)$ denotes the one-hot vectors representing the image-to-text matching relations, $\vec{y}^{\text{t2i}}(T)$ denotes the one-hot vectors representing the text-to-image matching relations, and H is cross-entropy.

Image-Text Matching (ITM). We also learn the alignments between images and texts in ITM, using a loss indicating whether an image-text pair is matched. For each image in a batch there is a matched (positive) text, and we sample an unmatched (negative) text in the batch. For each text there is a matched (positive) image, and we sample an unmatched image in the batch. The loss is defined as follows.

$$\mathcal{L}_{\text{itm}} = \mathbb{E}_{(I,T)\sim\mathcal{D}}\big[\text{H}(p^{\text{match}}(I,T)) + \text{H}(p^{\text{match}}(\tilde{I},T)) + \text{H}(p^{\text{match}}(I,\tilde{T}))\big] \quad (3)$$

where (I,T) denotes a positive image-text pair, $(\tilde{I},T)$ and $(I,\tilde{T})$ denote negative image-text pairs, $p^{\text{match}}(I,T)$ denotes a predicted matching probability of (I,T), and H denotes logistic loss.

Image-conditioned Masked Language Modeling (IMLM) We conduct IMLM on image-text pair data to learn the fusion encoder. We recover the masked text tokens given for an image-text pair by minimizing the cross entropy loss below.

$$\mathcal{L}_{\text{imlm}} = \mathbb{E}_{(I,T)\sim D}\text{H}(\vec{y}(\bar{T}), \hat{\vec{p}}(I,\bar{T})) \quad (4)$$

where (I,T) denotes an image-text pair, $\bar{T}$ denotes the masked text of T, $\hat{\vec{p}}(I,\bar{T})$ denotes the predicted probability vectors of the masked tokens of $\bar{T}$ based on I, $\vec{y}$ denotes the one-hot vectors representing the original tokens of $\bar{T}$, and H denotes cross-entropy.

Bounding Box Prediction (BBP) We adopt the BBP in X-VLM (Zeng et al., 2021, 2022), which locates the visual concept in the image by a bounding box given the text. With BBP we learn the alignments between the images and texts in multi-granularity. In BBP, two losses are simultaneously minimized to measure the differences between the predicted bounding box and the ground-truth bounding box. One is generalized intersection over union $GIoU$ (Rezatofighi et al., 2019) and the other is ℓ_1 distance.

$$\mathcal{L}_{\text{bbp}} = \mathbb{E}_{(I,T)\sim D}\{GIoU(\vec{b}, \hat{\vec{b}}) + \|\vec{b} - \hat{\vec{b}}\|_1\} \quad (5)$$

where $\vec{b} = (cx, cy, w, h)$ denotes the ground truth bounding box, $\hat{\vec{b}} = (\hat{cx}, \hat{cy}, \hat{w}, \hat{h})$ denotes the predicted bounding box. A bounding box is represented by two coordinates, width, and height.

Masked Image Modeling (MIM) We perform MIM on image data and image-text pair data to learn the vision encoder. Specifically, we recover

Task	Eval.	RoBERTa 1 –	BEiTv2 2 –	X^2-VLM 3 4M	X^2-VLM 4 1.3B	UNIMO-2 5 4M	FLAVA 6 70M	SimVLM 7 1.8B	OFA 8 21M	DaVinci 9 46M	DaVinci 10 648M	Uni-Per. 11 30M	OmniVL 12 14M	mPLUG-2_{base} 13 17M	X-FM_{base} 14 4M	X-FM_{base} 15 1.3B
MNLI	FT	87.6	–	–	–	87.5	80.3	83.4	84.3	82.3	83.1	81.5	–	87.6	**87.7**	**87.7**
CoLA	FT	63.6	–	–	–	62.1	50.7	46.7	52.3	52.1	54.8	52.2	–	–	65.3	**65.7**
MRPC	FT	90.2	–	–	–	–	84.2	79.8	88.7	83.1	84.5	–	–	87.3	**91.7**	91.2
QQP	FT	**91.9**	–	–	–	–	88.7	90.4	91.3	88.2	88.9	–	–	91.3	91.8	91.7
SST-2	FT	94.8	–	–	–	94.7	90.9	90.9	92.7	90.5	91.4	90.9	–	93.5	**95.0**	94.6
QNLI	FT	92.8	–	–	–	–	87.3	88.6	91.1	87.2	87.9	88.2	–	**93.2**	92.9	92.8
RTE	FT	78.70	–	–	–	–	57.8	63.9	70.8	60.7	64.2	75.8	–	**85.2**	83.8	82.7
STS-B	FT	**91.2**	–	–	–	**91.2**	85.7	87.2	–	86.3	87.1	–	–	–	90.8	90.7
Language Avg.		86.4	–	–	–	–	78.2	78.9	–	78.8	80.2	–	–	–	**87.4**	87.1
ImageNet	FT	–	**85.5**	–	–	80.8	–	–	82.2	–	83.9	84.5	–	–	85.3	**85.5**
ImageNet	LE	–	80.1	–	–	–	75.5	80.6	71.4†	75.9	77.7	–	–	–	81.0	**81.2**
Food101	LE	–	88.2†	–	–	–	88.5	–	75.2†	89.3	90.1	–	87.4	–	88.7	**90.5**
CIFAR10	LE	–	95.3†	–	–	–	92.9	–	86.1†	93.0	94.0	–	96.2	–	97.2	**97.4**
CIFAR100	LE	–	81.5†	–	–	–	77.7	–	66.7†	79.0	80.1	–	83.2	–	**86.7**	86.2
Pets	LE	–	**93.1**†	–	–	–	84.8	–	81.0†	85.5	88.2	–	87.1	–	90.8	90.2
DTD	LE	–	**78.4**†	–	–	–	77.3	–	70.3†	77.1	78.3	–	76.2	–	78.4	**80.0**
Flowers102	LE	–	95.7†	–	–	–	96.4	–	86.3†	96.1	96.9	–	89.8	–	**97.1**	96.4
Vision Avg.		–	88.7	–	–	–	86.3	–	79.2	86.7	87.9	–	86.7	–	89.8	**90.1**
VQAv2	FT	–	–	79.2	80.4	76.3	72.5	77.9	78.0	73.9	76.4	–	78.3	79.3	79.1	**80.5**
NLVR2	FT	–	–	86.1	87.0	–	–	81.8	–	77.9	–	–	–	–	86.7	**88.4**
Flickr30K TR R@1	ZS	–	–	85.1†	85.1†	84.6†	88.5	67.7	–	–	–	82.1	–	–	90.1	**93.4**
Flickr30K IR R@1	ZS	–	–	77.3†	79.2†	72.7	65.2	–	–	–	–	72.4	–	–	79.1	**84.1**
Flickr30K TR R@1	FT	–	–	97.4	**98.5**	92.0	–	–	–	–	–	93.6	94.9	96.9	97.4	98.1
Flickr30K IR R@1	FT	–	–	90.0	**90.4**	80.1	–	–	–	–	–	79.8	83.4	88.2	88.6	89.9
COCO TR R@1	ZS	–	–	68.4†	71.7†	–	42.7	–	–	–	–	64.6	–	–	73.8	**77.6**
COCO IR R@1	ZS	–	–	55.2†	58.3†	–	38.4	–	–	–	–	51.6	–	–	59.4	**61.1**
COCO TR R@1	FT	–	–	80.5	83.5	–	–	–	–	–	–	70.5	76.8	81.2	81.8	**84.2**
COCO IR R@1	FT	–	–	62.7	66.2	–	–	–	–	–	–	52.6	58.5	65.3	64.7	**67.0**
Vision-Language Avg.		–	–	78.2	80.0	–	–	–	–	–	–	–	–	–	80.1	**82.4**

Table 2: **Experimental results on vision, language and vision-language tasks.** The multi-modal data size used for pre-training are reported under the model name. MNLI results are average of MNLI-m and MNLI-mm. MRPC results are average accuracies and F1 scores. Matthews correlation coefficient (MCC) is reported for CoLA, and Pearson correlation coefficient (PCC) is reported for STS-B. We report accuracies for all the vision and multi-modal tasks. FT is short for fine-tuning, LE for linear evaluation, ZS for zero-shot, TR for text retrieval, and IR for image retrieval. Results for RoBERTa are from its corresponding paper (Liu et al., 2019), and they use the mid-training (Phang et al., 2018) on MNLI for RTE, MRPC, and STS-B while other models (e.g., DaVinci, X-FM) do not use this trick. Note that mPLUG-2 used more layers and parameters than RoBERTa and X-FM for the language understanding tasks. Language Avg. is the average score of all the language tasks, while Vision Avg. is the average score of six line evaluation tasks except ImageNet. Vision-Language Avg. is the average score of all vision-language tasks. † are our reproduced results with the officially released models.

the masked image patches in an image by minimizing the loss below.

$$\mathcal{L}_{\text{mim}} = \mathbb{E}_{(I,T)\sim\mathcal{D}}||\vec{v}(\bar{I}) - \hat{\vec{v}}(\bar{I})||_2 + \mathbb{E}_{I\sim\mathcal{I}}||\vec{v}(\bar{I}) - \hat{\vec{v}}(\bar{I})||_2 \quad (6)$$

where (I, T) and I denote an image-text pair and a single image respectively, $\bar{I}$ denotes the masked image I, $\hat{\vec{v}}(\bar{I})$ denotes the predicted representations at the masked positions and `[CLS]` of $\bar{I}$, and $\vec{v}(\bar{I})$ denotes the target representations at the masked positions and `[CLS]` of $\bar{I}$. $||||_2$ is the MSE loss. We employ block masking following previous work (Bao et al., 2021; Peng et al., 2022). Note that (I, T) and I are independently sampled from $\mathcal{D}$ and $\mathcal{I}$, and the sample sizes are not necessarily equal.

Finally, the pre-training objective of X-FM is defined as the sum of the losses described above.

$$\mathcal{L} = \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{itc}} + \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{imlm}} + \mathcal{L}_{\text{bbp}} + \mathcal{L}_{\text{mim}}$$

4 Experiments

4.1 Implementation Details

Pre-training Datasets. We mainly conduct our experiments on several widely used public datasets, consisting of two in-domain datasets, COCO (Lin et al., 2014) and Visual Genome (VG) (Krishna et al., 2017), and two out-of-domain datasets, SBU Captions (Ordonez et al., 2011) and Conceptual Captions (CC) (Sharma et al., 2018). Following X-VLM (Zeng et al., 2021, 2022), we also include annotations of objects and regions from RefCOCO (Yu et al., 2016), Objects365 (Shao et al., 2019) and OpenImages (Kuznetsova et al., 2018). Since we assume also using uni-modal data, we include RoBERTa corpus (Liu et al., 2019), $C4$ datasets (Raffel et al., 2020) and Imagenet21K (Ridnik et al., 2021). In addition, we also scale up the pre-training dataset with Conceptual 12M dataset (CC-12M) (Changpinyo et al., 2021) and LAION (Schuhmann et al., 2022) as the "more data" setting, which contains around 1.3B

image-text pairs. Please refer to Appendix B for statistics of the pre-training datasets.

Pre-training Settings. Our model is of base size, and the detailed parameters are explained in Appendix D. The vision encoder is initialized with BEiTv2. The language encoder is initialized with RoBERTa. The fusion encoder is trained from scratch. X-FM is pre-trained at image resolution of 224×224 with patch size of 16×16. We pre-train X-FM for 200K steps with a batch size of 3072 image-text pairs, 3072 images, and 8192 sentences on 32 A100, which takes about six days. The learning rate for both models is warmed-up to $1e^{-4}$ in the first 2500 steps and decayed following a linear schedule. We set the maximum number of text tokens to 30 for image-text pairs, while that of pure text corpus is set to 128. For the "more data" setting, we pre-train X-FM for 400k steps with 18k batch size on 64 A100. Due to the consideration of computational cost, we did not pre-train the large or giant models. We apply mixed precision for pre-training. We choose widely used downstream tasks whose details are shown in Appendix C.

4.2 Comparison with Foundation Models

We extensively compare the performance of X-FM with state-of-the-art foundation models on vision, language, and multi-modal tasks. We first compare our model with general foundation models, including UNIMO-v2 (Li et al., 2021c), FLAVA (Singh et al., 2021), SimVLM (Wang et al., 2021c), OFA (Wang et al., 2022b), DaVinci (Diao et al., 2022), Uni-Perceiver-MoE (Zhu et al., 2022), OmniVL (Wang et al., 2022a), and mPLUG-2 (Xu et al., 2023). We also include comparisons with SOTA foundation models specifically designed for language, vision, or vision-language tasks, RoBERTa (Liu et al., 2019), BEiTv2 (Peng et al., 2022), and X^2-VLM (Zeng et al., 2022). There are several observations in Table 2. First, X-FM_{base} (column 15) outperforms all the previous general foundation models (column 5-13) across almost all tasks by a large margin, becoming a new and stronger general foundation model. When we use less pre-training data, X-FM can also achieve competitive performance compared with previous general foundation models (column 5-13 vs 14). Second, we compare X-FM with state-of-the-art foundation models specifically designed for language, vision, and vision-language tasks, RoBERTa, BEiTv2 and X^2-VLM. We observe that X-FM is also better than or comparable with the foundation models (column 1,2,3,4 vs 15). We further compare our model, X-FM_{base}, with three previous foundation models on 18 image classification tasks on the linear evaluation setting to evaluate generalization performance on vision understanding tasks. The results are shown in Table 4. X-FM_{base} wins 11 of 18 tasks, 7 for CLIP, 2 for FLAVA, and 2 for DaVinci.

4.3 Comparison with multi-modal Models

In addition to general foundation models, we also compare X-FM with state-of-the-art vision-language models. The results are shown in Table 3 and Table 6. X-FM demonstrates its superiority on five downstream vision-language tasks including MSCOCO Retrieval, Flick Retrieval, VQA, NLVR and RefCOCO+. Note that X-FM_{base} outperforms CLIP, ALIGN and Florence on image-text retrieval tasks with fewer parameters and much less training data. Compared to the recently released SOTA vision-language model, X^2-VLM, X-FM is much better on zero-shot image-text retrieval tasks. When we scale up pre-training datasets, X-FM_{base} is consistently better than previous vision-language models for most cases.

4.4 Ablation Study

To verify the contributions of different modules in our framework, we ablate them and evaluate the performance of X-FM on all downstream tasks. The results are shown in Table 5. We first explain several abbreviations in the table. S-MLM means that we only stop the gradient of language representations in IMLM task, while S-ITM means stopping the gradient of language representations for computing ITM and BBP. wostop indicates without stopping the gradients of all language representations. woMIM means that we do not learn by MIM, while wBEiTv2 tokenizer means that we learn by MIM with the image tokenizer used in BEiTv2. Multi-task is a variation that uses straightforward multi-task learning to optimize the three encoders in X-FM. To make a fair comparison, we also train RoBERTa, BEiTv2 and X^2-VLM with the same data noted as RoBERTa†, BEiTv2† and X^2-VLM†. Note that we also increase the fusion layers in X^2-VLM† to make the parameter sizes comparable to our models. RoBERTa†, BEiTv2† and X^2-VLM† all have slightly better results on average than the

Model	# Params	MSCOCO (5K test set) TR-Fine-Tune R@1/R@5/R@10	MSCOCO (5K test set) IR-Fine-Tune R@1/R@5/R@10	Flickr30K (1K test set) TR-Fine-Tune R@1/R@5/R@10	Flickr30K (1K test set) IR-Fine-Tune R@1/R@5/R@10	MSCOCO (5K test set) TR-Zero-Shot R@1/R@5/R@10	MSCOCO (5K test set) IR-Zero-Shot R@1/R@5/R@10	Flickr30K (1K test set) TR-Zero-Shot R@1/R@5/R@10	Flickr30K (1K test set) IR-Zero-Shot R@1/R@5/R@10
ALBEF	210M	73.1/91.4/96.0	56.8/81.5/89.2	94.3/99.4/99.8	82.8/96.7/98.4	–	–	**90.5**/98.8/99.7	76.8/93.7/96.7
$\text{VLMo}_{\text{base}}$	175M	74.8/93.1/96.9	57.2/82.6/89.8	92.3/99.4/99.9	79.3/95.7/97.8	–	–	–	–
VL-BEiT	175M	79.5/–/–	61.5/–/–	95.8/–/–	83.9/–/–	–	–	–	–
OmniVL	288M	76.8/93.6/97.3	58.5/82.6/89.5	94.9/9.6/99.9	83.4/97.0/98.6	–	–	–	–
X-VLM	216M	80.4/95.5/98.2	63.1/85.7/91.6	96.8/99.8/**100**	86.1/97.4/98.7	70.8/92.1/96.5	55.6/82.7/90.0	85.3/97.8/99.6	71.9/93.3/96.4
$\text{X}^2\text{-VLM}_{\text{base}}$	255M	80.5/95.5/97.8	62.7/84.7/90.7	**97.4**/99.9/**100**	**90.0/98.6/99.3**	$68.4^{\dagger}$/92.5†/96.8†	$55.2^{\dagger}$/82.2†/89.3†	$85.1^{\dagger}$/**99.2**†/**100.0**†	$77.3^{\dagger}$/**95.3**†/**97.6**†
X-FM$_{\text{base}}$	327M	**81.8/96.0/98.3**	**64.7/86.1/91.6**	**97.4/100/100**	88.6/97.9/98.9	**73.8/93.9/97.2**	**59.4/83.6/90.0**	90.1/**99.2**/99.9	**79.1**/95.2/97.3
More Data									
CLIP	490M	–	–	88.7/98.0/99.2	76.7/93.6/96.4	58.4/81.5/88.1	37.8/62.4/72.2	88.0/98.7/99.4	68.7/90.6/95.2
ALIGN	490M	77.0/93.5/96.9	59.9/83.3/89.8	95.3/99.8/100	84.9/97.4/98.6	58.6/83.0/89.7	45.6/69.8/78.6	88.6/98.7/99.7	75.7/93.8/96.8
Florence	893M	81.8/95.2/–	63.2/85.7/–	97.2/99.9/–	87.9/98.1/–	64.7/85.9/–	47.2/71.4/–	90.9/99.1/–	76.7/93.6/–
$\text{X}^2\text{-VLM}_{\text{base}}$	255M	83.5/96.3/**98.5**	66.2/87.1/92.2	**98.5/100/100**	**90.4**/98.2/99.3	$71.7^{\dagger}$/93.4†/97.5†	$58.3^{\dagger}$/**84.7**†/**91.0**†	$84.6^{\dagger}$/99.1†/**99.9**†	$79.2^{\dagger}$/96.4†/98.0†
X-FM$_{\text{base}}$	327M	**84.2/96.4**/98.4	**67.0/87.2/92.4**	98.1/**100/100**	89.9/**98.6/99.4**	**77.6/94.8/97.7**	**61.1**/84.5/90.6	**93.4/99.8/99.9**	**84.1/96.5/98.1**
Super-Large Models									
CoCa	2.1B	–	–	–	–	66.3/86.2/91.8	51.2/74.2/82.0	92.5/99.5/99.9	80.4/95.7/97.7
BEiT-3	1.9B	84.8/96.5/98.3	67.2/87.7/92.8	98.0/100/100	90.3/98.7/99.5	–	–	94.9/99.9/100.0	81.5/95.6/97.8

Table 3: Results of text-retrieval (TR) and image-retrieval (IR) on COCO and Flickr30K. † denotes our reproduced results with the officially released models. In more data setting, we use Conceptual 12M dataset (CC-12M) (Changpinyo et al., 2021) and LAION (Schuhmann et al., 2022) as additional datasets. More details are explained in Appendix B. Giant models with over 1B parameters (e.g., BEiT-3) are in grey since they are not directly comparable with other models.

		ImageNet	Food101	CIFAR10	CIFAR100	StanfordCars	Aircraft	DTD	OxfordiiitPets	Flowers102	MNIST	STL10	Country211	Sun397	SST	Caltech101	GTSRB	PCAM	FER2013
CLIP	B/16-224px	80.2	**92.8**	96.2	83.1	86.7	**59.5**	79.2	**93.1**	**97.1**	**99.0**	99.0	**30.1**	78.4	**75.5**	94.7	86.6	83.5	69.5
FLAVA	B/16-224px	75.5	88.5	92.9	77.7	70.9	47.3	77.3	84.8	98.1	**99.0**	98.9	28.9	82.1	57.1	95.7	79.5	**85.3**	61.1
DaVinci	B/16-224px	77.7	90.1	94.0	80.1	74.6	49.6	78.3	88.2	96.9	**99.0**	**99.2**	29.9	–	–	–	–	–	–
X-FM$_{\text{base}}$	B/16-224px	**81.2**	90.5	**97.4**	**86.2**	**88.3**	47.4	**80.0**	90.2	96.4	**99.0**	**99.2**	24.9	**93.9**	60.6	**97.1**	**90.9**	82.4	**72.6**

Table 4: **Linear evaluation performance of four foundation models over 18 datasets.** B/16-224px means base size model, 16*16 patches, and 224*224 resolution, respectively. The best performance is identified with bold.

official ones. From the results, we have the following observations.

First, both designs (stop gradient and vision-language guided MIM) bring improvements, and the combination can make further improvements on all three downstream tasks (column 10 vs. others). Second, without separated language representations, models always perform worse on language understanding tasks (column 10 vs. 2,3,4). Besides, the separate language representations in the IMLM task on image-text data are helpful for multi-modal tasks (column 2 vs. 4). As we point out in section 1, the fusion encoder can concentrate on learning the alignments between language and vision features instead of predicting masked tokens with clues from other visible text tokens. Although S-ITM shows slight side effects (column 4 vs. 3), stopping the gradients of language representation in the fusion encoder is necessary to simultaneously achieve strong language understanding and vision-language understanding capability. Third, the vision-language guided MIM task is useful for both vision-language and vision learning (column 10 vs. 6). Meanwhile, the targets in our MIM task are better than the BEiTv2 tokenizer (column 10 vs. 7). Four, X-FM is much better than a naive multi-task learning strategy for a foundation model, compared with which, $\text{X-FM}_{\text{base}}$ improves an average of 0.9%, 1.7% and 1.6% on language, vision, and vision-language tasks, respectively (column 10 vs. 9). Five, X-FM is also better than foundation models specifically designed for language, vision, and vision-language tasks with the same training corpus (column 10 vs. 1,5,8).

5 Limitations and Potential Risks

Limitations. Like most existing work on foundation models, the entire project consumed over 5 A100 GPU years on a computing cluster with high electricity costs, although we only tested base models. There is still potential for efficiency improvement through sparse attention (Zaheer et al., 2020) or the lottery ticket hypothesis (Frankle and Carbin, 2018). We will explore the techniques to improve the training efficiency and reduce the carbon footprint so that we can adhere to the proposals on "green" deep learning (Schwartz et al., 2020; Xu et al., 2021).

Task	Eval.	RoBERTa†	X-FM$_{base}$ S-MLM	S-ITM	wostop	BEiTv2†	woMIM	wBEiTv2 Tokenizer	X^2-VLM†	Multi-task	**ALL**
		1	2	3	4	5	6	7	8	9	10
MNLI	FT	**87.7**	87.4	87.3	**87.7**	–	–	–	–	87.4	87.6
CoLA	FT	63.2	61.6	63.6	64.2	–	–	–	–	62.2	**65.2**
MRPC	FT	90.7	92.2	91.1	90.7	–	–	–	–	92.0	**92.5**
QQP	FT	91.5	**91.6**	**91.6**	**91.6**	–	–	–	–	**91.6**	**91.6**
SST-2	FT	95.0	95.1	94.2	94.6	–	–	–	–	94.4	**95.3**
QNLI	FT	93.1	93.0	**93.2**	92.5	–	–	–	–	92.8	92.9
RTE	FT	80.9	79.1	81.6	81.2	–	–	–	–	79.8	**81.9**
STS-B	FT	**90.9**	90.7	90.7	90.4	–	–	–	–	90.1	90.8
Language Avg.		86.6	86.4	86.7	86.6	–	–	–	–	86.3	**87.2**
ImageNet	FT	–	–	–	–	**85.5**	84.8	85.0	–	85.0	85.3
ImageNet	LE	–	–	–	–	80.5	79.1	79.4	–	79.3	**81.1**
Food101	LE	–	–	–	–	88.2	86.9	87.2	–	86.9	**88.7**
CIFAR10	LE	–	–	–	–	95.3	96.6	96.5	–	96.6	**97.5**
CIFAR100	LE	–	–	–	–	81.5	83.3	83.9	–	84.1	**86.9**
Pets	LE	–	–	–	–	**93.1**	88.1	88.5	–	88.2	90.7
DTD	LE	–	–	–	–	78.4	77.7	76.9	–	78.0	**78.7**
Flowers102	LE	–	–	–	–	95.7	94.1	94.5	–	94.2	**97.1**
Vision Avg.		–	–	–	–	87.3	86.3	86.5	–	86.5	**88.2**
VQAv2	FT	–	**78.8**	78.5	78.7	–	78.3	78.2	78.0	78.2	78.6
NLVR2	FT	–	86.3	86.0	86.4	–	85.9	85.5	86.2	86.1	**86.7**
Flickr30K TR R@1	ZS	–	88.3	87.2	87.1	–	87.1	87.2	87.7	85.0	**89.3**
Flickr30K IR R@1	ZS	–	76.6	74.9	75.8	–	76.1	75.3	75.1	75.6	**77.4**
Flickr30K TR R@1	FT	–	97.5	97.0	97.2	–	96.4	96.7	97.0	97.0	**97.7**
Flickr30K IR R@1	FT	–	**87.4**	86.9	87.3	–	86.2	86.6	86.2	86.4	**87.4**
COCO TR R@1	ZS	–	72.0	72.1	70.5	–	73.0	72.1	**73.2**	69.9	72.8
COCO IR R@1	ZS	–	58.4	57.1	57.7	–	58.2	57.7	57.7	56.5	**59.0**
COCO TR R@1	FT	–	**81.2**	80.2	80.9	–	80.6	80.1	80.3	80.0	**81.2**
COCO IR R@1	FT	–	**64.2**	63.4	63.6	–	63.7	63.0	63.1	63.0	64.0
Vision-Language Avg.		–	79.1	78.3	78.5	–	78.6	78.2	78.5	77.8	**79.4**

Table 5: **Ablation studies on vision, language, and vision-language tasks.** We use the same settings as Table 2. "ALL" means we use both of our proposed techniques. To compare fairly, we pre-train all variants with the same data at the same settings for both pre-training and fine-tuning. Avg. means the average score.

Method	# Params	VQA test-dev	VQA test-std	NLVR2 dev	NLVR2 test-P	RefCOCO+ val	RefCOCO+ testA^d	RefCOCO+ testB^d
ALBEF	210M	74.5	74.7	80.2	80.5	–	–	–
VLMo$_{base}$	175M	76.6	76.9	82.8	83.3	–	–	–
METER	341M	77.7	77.6	82.3	83.1	–	–	–
VL-BEiT	175M	77.5	77.8	81.9	82.7	–	–	–
BLIP$_{base}$	240M	78.2	78.2	82.5	83.1	–	–	–
X-VLM	216M	78.1	78.1	84.2	84.2	80.2	86.4	71.0
OFA$_{base}$	182M	78.0	78.1	–	–	81.4	87.2	74.3
OmniVL	288M	78.3	78.4	–	–	–	–	–
X^2-VLM$_{base}$	255M	**79.2**	**79.3**	85.9	86.1	**85.4**	89.2	77.3
X-FM$_{base}$	327M	79.1	79.2	**86.3**	**86.5**	84.8	**89.7**	**79.1**
More Data								
SimVLM$_{base}$	273M	77.9	78.1	81.7	81.8	–	–	–
X^2-VLM$_{base}$	255M	80.4	80.2	86.2	87.0	85.2	90.3	78.4
X-FM$_{base}$	327M	**80.5**	**80.4**	**87.6**	**88.4**	**86.1**	**90.4**	**79.8**
Super-Large Models								
CoCa	2.1B	82.3	82.3	86.1	87.0	–	–	–
BEiT-3	1.9B	84.2	84.0	91.5	92.6	–	–	–

Table 6: Results on VQA, visual reasoning and visual grounding. Giant models with over 1B parameters (e.g., CoCa and BEiT-3) are in grey because they are not directly comparable with other models.

Due to considerations of fair comparisons and computational resources, we did not try super-large models which use at least 1.9B or more parameters like BEITv3 (Wang et al., 2022d), CoCa (Yu et al., 2022) and PaLI (Chen et al., 2022). We also did not pre-train large size model on large-scale datasets. However, scalability is also an important factor for foundation models. We leave the investigations to future work.

Potential Risks. The image-text pairs use for training our model are mostly derived from lexical databases and image queries in English, resulting in source material with a North American or Western European bias.

6 Conclusion

In this work, we address the problem of how to build a general foundation model that can perform the best for all the understanding tasks of language, vision, and vision-language. We propose a new method for training general foundation model with two new and effective techniques, bringing in X-FM, to learn rich language, vision, and vision-language representations at the same time. Experimental results demonstrate that X-FM outperforms other general foundation models by a large margin. Moreover, X-FM can even be better than or comparable to the SOTA foundation models specifically designed for language, vision, or vision-language understanding tasks.

References

Eneko Agirre, Lluís Màrquez, and Richard Wicentowski, editors. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

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

Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. 2022. Vl-beit: Generative vision-language pretraining. *arXiv preprint arXiv:2206.01127*.

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

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. *arXiv preprint arXiv:1702.08303*.

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

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. 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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Shizhe Diao, Wangchunshu Zhou, Xinsong Zhang, and Jiawei Wang. 2022. Prefix language models are unified modal learners. *arXiv preprint arXiv:2206.07699*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

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. 2021. 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.

Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

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

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. 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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *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.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Andrej Karpathy and Fei-Fei Li. 2015. 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.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*.

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

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. 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. Association for Computational Linguistics.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021c. 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. Association for Computational Linguistics.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. 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.

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

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *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.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. 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.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*.

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

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. 2021. Learning transferable visual models from natural language supervision. In *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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 658–666. Computer Vision Foundation / IEEE.

Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. 2021. Imagenet-21k pretraining for the masses.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

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

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. 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.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2021. Flava: A foundational language and vision alignment model. *ArXiv preprint*, abs/2112.04482.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. 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. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019a. 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. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019b. 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. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Hao Tan and Mohit Bansal. 2019. 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. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *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.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. 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.

Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022a. Omnivl: One foundation model for image-language and video-language tasks. *arXiv preprint arXiv:2209.07526*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

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

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705.

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

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

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *CoRR*, abs/2108.10904.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2022a. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678.

Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. 2022b. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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. Association for Computational Linguistics.

Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. *arXiv preprint arXiv:2302.00402*.

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

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer.

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. 2021. Florence: A new foundation model for computer vision. *arXiv preprint*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

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

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2022. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*.

Xinsong Zhang, Pengshuai Li, and Hang Li. 2020. Ambert: A pre-trained language model with multi-grained tokenization. *arXiv preprint arXiv:2008.11869*.

Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. 2022. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *arXiv preprint arXiv:2206.04674*.

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Xiaogang Wang, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2021. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *arXiv preprint arXiv:2112.01522*.

A Comparison of Foundation Models

Table 7 shows an extensive comparison of recent foundation models and X-FM on multiple axes. Previous work either (i) perform best on uni-modal tasks (Liu et al., 2019; Peng et al., 2022) or vision-language tasks (Zeng et al., 2021, 2022); (2) target a specific uni-modal domain along with part of vision-and-language tasks (Wang et al., 2021a; Radford et al., 2021; Jia et al., 2021; Wang et al., 2021c; Yu et al., 2022; Wang et al., 2022b; Diao et al., 2022); or (3) target all domains but cannot perform best on all the tasks (Li et al., 2021c; Singh et al., 2021; Zhu et al., 2022). Our model, X-FM, is a general foundation model that can perform the best for all the understanding tasks of language, vision, and vision language.

B Details of Pre-training Datasets

We conduct our experiments on several widely used public datasets, consisting of two in-domain datasets, COCO (Lin et al., 2014) and Visual Genome (VG) (Krishna et al., 2017), and two out-of-domain datasets, SBU Captions (Ordonez et al., 2011) and Conceptual Captions (CC) (Sharma et al., 2018). Following X-VLM (Zeng et al., 2021, 2022), we use annotations of objects and regions from RefCOCO (Yu et al., 2016), Objects365 (Shao et al., 2019) and OpenImages (Kuznetsova et al., 2018). We also include uni-modal data, RoBERTa corpus (Liu et al., 2019), $C4$ datasets (Raffel et al., 2020) and Imagenet21K (Ridnik et al., 2021).

For our "more data" setting, we scale up the pre-training dataset by including image-text pairs from Conceptual 12M dataset (CC-12M) (Changpinyo et al., 2021) and LAION (Schuhmann et al., 2022). Thanks to LAION, we can use a large-scale public corpus of image-text pairs. However, we note that there are amounts of "low-quality" image text pairs, as it is only filtered by the CLIP score. The clip score is deceptive when an image contains word tokens in its caption. Therefore, we apply three filters, OCR filter, text filter, and image filter, to capture "high-quality" image-text pairs from LAION. Note that we only use English data in LAION. The OCR filter will remove an image (image-text pair) when its OCR text contains more than four words or any token in the caption. The text filter will remove a text image (image-text pair) if it is an address or contains only digits or symbols. The image filter will remove an image (image-text pair) if the shorter edge is smaller than 224 pixels, and also remove an image (image-text pair) if the height/width or width/height ratio is greater than 3. Finally, we have 1.3B paired data after all three filters. Statistics of the pre-training datasets are shown in Table 8.

C Details of Downstream Tasks

We report overall performance on eight language tasks from GLUE (Wang et al., 2019), eight vision tasks following OmniVL (Wang et al., 2022a) (More image classification tasks can be found in Appendix **??**.), four multi-modal tasks, which are text-image retrieval on MSCOCO and Flickr, visual question answering (VQA (Goyal et al., 2017)), visual reasoning (NLVR2 (Suhr et al., 2019a)) and visual grounding (RefCOCO+ (Yu et al., 2016)). For image-text retrieval task, we report both zero-shot results and fine-tuned results. For the ImageNet classification task, we report both linear evaluation results and fine-tuning results. The other vision tasks are evaluated in the linear evaluation setting. All the other tasks are evaluated in the fine-tuning setting. Because the image resolution differs between pre-training and fine-tuning, the position parameters are adapted using linear interpolation. For all downstream tasks, we apply random resize crops and horizontal flips augmentation for the images during training. More details of network architectures and hyper-parameters setups are given in Appendix D.

Language Understanding.
We conduct experiments on GLUE benchmark including MNLI (Williams et al., 2018), CoLA (Warstadt et al., 2019), MRPC (Dolan and Brockett, 2005), QQP (Iyer et al., 2017), SST-2 (Socher et al., 2013), QNLI (Rajpurkar et al., 2016), RTE (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), and STS-B (Agirre et al., 2007). We follow the practice of BERT (Devlin et al., 2019; Liu et al., 2019) and feed the input into the language encoder, and the hidden state of the `[CLS]` is fed into a new multi-class linear classifier or regression head.

Vision Understanding.
We conduct vision experiments on both fine-tuning and linear evaluation (linear eval). The linear evaluation follows a common practice (Caron et al., 2021; He et al., 2020; Singh et al., 2021) in self-

Methods	Multimodal data			Pretraining Objectives					Fusion Arch.			Target Modalities			
	public	dataset(s)	size	Contr.	ITM	BBP	(M/P)LM	Unimodal	ST	CT	MT	V	CV&L	MV&L	L
RoBERTa (Liu et al., 2019)	–	–	–	–	–	–	–	MLM	–	–	–	–	–	–	✓
BEiTv2 (Peng et al., 2022)	–	–	–	–	–	–	–	MIM	–	–	–	✓	–	–	–
X-VLM (Zeng et al., 2021, 2022)	✓	Combination	5M	✓	✓	✓	MLM	–	–	✓	–	–	✓	✓	–
VLMo (Wang et al., 2021a)	✓	Combination	5M	✓	✓	–	MLM	MLM+MIM	–	–	✓	–	✓	✓	–
CLIP (Radford et al., 2021)	✗	WebImageText	400M	✓	–	–	–	–	–	–	–	✓	✓	–	–
ALIGN (Jia et al., 2021)	✗	JFT	1.8B	✓	–	–	–	–	–	–	–	✓	✓	–	–
SimVLM (Wang et al., 2021c)	✗	JFT	1.8B	–	–	–	PrefixLM	PrefixLM	✓	–	–	*	–	✓	✓
CoCa (Yu et al., 2022)	✗	JFT	4.8B	✓	–	–	LM	–	✓	–	–	✓	✓	✓	–
UNIMO-2 (Li et al., 2021c)	✓	Combination	5M	–	✓	–	MLM	VCL	✓	–	–	✓	✓	✓	✓
OFA (Wang et al., 2022b)	✓	Combination	15M	–	–	–	LM	LM	✓	–	–	*	–	✓	✓
DaVinci (Diao et al., 2022)	✓	Combination	46M	–	–	–	PrefixLM + PrefixIM	PrefixLM	✓	–	–	✓	–	✓	✓
FLAVA (Singh et al., 2021)	✓	Combination	70M	✓	✓	–	MLM	MLM+MIM	✓	–	–	✓	✓	✓	✓
Uni-Perceiver-MoE (Zhu et al., 2022)	✓	Combination	116M	–	✓	–	LM+MLM	LM+MLM+Classify.	✓	–	–	✓	✓	✓	✓
X-FM	✓	Combination	5M	✓	✓	✓	MLM+MIM	MLM+MIM	–	✓	–	✓	✓	✓	✓
Super-Large Models															
Flamingo (Alayrac et al., 2022)	✗	Combination	2.2B	–	–	–	LM	–	✓	–	–	–	✓	✓	–
BEiT-v3 (Wang et al., 2022d)	✓	Combination	21M	–	–	–	MLM	MLM+MIM	–	–	✓	*	✓	✓	–
PaLI (Chen et al., 2022)	✗	WebImageText	41B	–	–	–	LM	–	✓	–	–	✓	✓	✓	✓

Table 7: **Comparison of recent foundation models in different modalities.** Contr. indicates contrastive learning. ITM is short for image-text matching. BBP represents boundary box prediction. (M/P)LM means image-conditioned (masked/prefix) language modeling. V, CV&L, MV&L and L stand for vision tasks, cross-modal retrieval tasks, multi-modal fusion tasks and language tasks respectively. ST, CT and MT are abbreviations for single Transformer, cross-attention Transformer and multiway Transformer. VCL stands for visual contrastive learning. $*$ means the modality is partially targeted (SimVLM and OFA include ImageNet.). Giant models with over 1B parameters (e.g. BEiT-3) are in grey since they are not directly comparable with other models.

Dataset	# Images	# Texts	# Objects	# Regions
COCO	0.11M	0.55M	0.45M	-
VG	0.10M	-	2.0M	3.7M
SBU	0.86M	0.86M	-	-
CC-3M	2.9M	2.9M	-	-
Objects365	0.58M	-	2.0M	-
OpenImages	1.7M	-	4.2M	-
C4	-	800GB	-	-
RoBERTa Corpus	-	160GB	-	-
ImageNet-21k	14M	-	-	-
More Data				
CC-12M	11.1M	11.1M	-	-
LAION	1.3B	1.3B	-	-

Table 8: Statistics of the pre-training datasets.

supervised learning to evaluate the representation quality, where the pre-trained backbone model is frozen, and an MLP head is appended on top of it. We choose 7 popular datasets following OmnVL (Wang et al., 2022a): ImageNet (Russakovsky et al., 2015), Food101 (Bossard et al., 2014), CIFAR10 (Krizhevsky et al., 2009), CIFAR100 (Krizhevsky et al., 2009), DTD (Cimpoi et al., 2014), Pets (Parkhi et al., 2012) and Flowers102 (Nilsback and Zisserman, 2008).

Vision-Language Understanding.

Image-Text Retrieval We evaluate X-FM on both MSCOCO and Flickr30K datasets. We adopt the widely used Karpathy split (Karpathy and Li, 2015) for both datasets. Following the previous work (Li et al., 2021a; Zeng et al., 2021, 2022), we first encode images and texts separately and calculate $s(I, T)$ to obtain the top-k candidates, and then use the fusion encoder to re-rank the candidates.

Visual Question Answering The task requires the model to predict an answer given an image and a question. We evaluate X-FM on the VQA v2.0 dataset (Goyal et al., 2017). Following the previous work (Zeng et al., 2021), we use a Transformer decoder to generate answers based on the outputs of the fusion module. The decoder network shares the same network architecture with the fusion encoder. Note that we use an image resolution of 768*768 for the final result of X-FM$_{\text{base}}$, and use an image resolution of 480*480 for X-FM$_{\text{base}}$ in ablation studies for efficient fine-tuning.

Visual Reasoning We evaluate X-FM on a widely used benchmark NLVR2 (Suhr et al., 2019b). The task allows the model to determine whether a text describes the relations between two images. Following previous work (Wang et al., 2021a; Bao et al., 2022), we formulate the triplet input into two image-text pairs, each containing the text description and an image. We then concatenate the final output [CLS] features of the fusion module of the two pairs to predict the label.

Visual Grounding We evaluate X-FM on RefCOCO+ (Yu et al., 2016). Given an image and a text description as input, the final output $[CLS]$ features of the fusion module is utilized to predict the bounding box (cx, cy, w, h), i.e. the normalized center coordinates, width, and height.

Model	Param		Hidden	Layers		
	Total	Trans.		Vision	Text	Fusion
X-FM$_{\text{base}}$	327M	284M	768	12	12	12

Table 9: Size variants of X-FM. All modules consist of transformer layers. Param indicates the parameter. Total means the total parameter number, and Trans. indicates the parameter number for Transformer layers.

D Details of hyper parameters

Pre-training X-FM$_{\text{base}}$ is implemented with a 12-layer language encoder, a 12-layer vision encoder, and a 12-layer fusion encoder, 768 dimensions for hidden states, 3072 for intermediate size, and 128 for maximum input length. We initialize the language encoder with RoBERTa and the vision encoder with BEiTv2. The weight decay is set to 0.01 with $\beta_1 = 0.9, \beta_2 = 0.98$. The learning rate is 1e-4 with a warm-up period for the first 2500 steps and then linearly decayed to 0. In each batch, there are 3072 image-text pairs, 3072 images, and 8192 text-only sentences. We use center-crop to resize each image to the size of 224×224. The model sizes and default hyper-parameter settings are shown in Table 9 and Table 10, respectively.

config	value
optimizer	AdamW
learning rate	1e-4
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.999
language batch size	8192
vision batch size	3072
vision-language batch size	3072
learning rate schedule	linear decay
warmup steps	2500
training steps	200k
augmentation	RandomResizedCrop
image res	224*224
patch size	16
text length for MLM	128
text length for IMLM	30

Table 10: Pre-training setting.

Fine-tuning The learning rate is $\in$ {1e-5, 2e-5, 5e-5} and our model is optimized by AdamW. Because the image resolution differs between pre-training and fine-tuning, the position parameters are adapted using linear interpolation. For all downstream tasks, we apply random resize crops and horizontal flips augmentation during training. The default settings for text classification, image classification and vision-language understanding are shown in Tables 11, 12, 13 and 14, respectively. Note that the resolution for VQA is different as described in Section C.

config	value
optimizer	AdamW
learning rate	{1e-5, 2e-5, 5e-5}
weight decay	0.0
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	{16, 32, 64}
learning rate schedule	linear decay
warmup ratio	0.0
training epochs	{5, 10, 20}

Table 11: Text classification: GLUE setting.

config	value
optimizer	AdamW
learning rate	[2e-5, 4e-5]
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	[256, 2048]
learning rate schedule	linear decay
warmup rate	0.1
training epochs	100
augmentation	RandomResizedCrop
image res	224*224
patch size	16

Table 12: Image classification: Linear probing setting.

config	value
optimizer	AdamW
learning rate	4e-5
minimal learning rate	1e-7
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	1024
learning rate schedule	linear decay
warmup rate	0.1
training epochs	100
augmentation	RandomResizedCrop
image res	224*224
patch size	16
label smoothing	0.1
mixup prob.	1.0
cutmix prob.	1.0

Table 13: ImageNet classification: Fine-tuning setting.

config	value
optimizer	AdamW
learning rate	{1e-5, 2e-5, 5e-5}
weight decay	0.01
optimizer momentum	β_1, β_2=0.9, 0.999
batch size	{64, 192, 512}
learning rate schedule	linear decay
warmup rate	0.1
training epochs	{10, 15, 20}
augmentation	RandomResizedCrop
image res	384*384
patch size	16

Table 14: Vision-Language understanding: fine-tuning setting.