Supplementary Document
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

1. Ablation Studies

(We additionally report ImageNet-1K (IN1K) results for ablation studies compared with the submission.) 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 16×16 patch size and are trained at resolution 224×224. Most settings and hyperparameters are kept the same as the giant-size model. 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 [4]. The models are pretrained for 200K steps with 2e-3 peak learning rate and 6144 batch size. 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. Top1 accuracy is reported for ImageNet-1K. The models are finetuned as a dual encoder for Flickr30K. Gray indicates the default setting of BEiT-3.

Backbone Architecture We study the effects of different model architectures. Table 1 shows that Multiway Transformers perform better than standard Transformers on four benchmarks. Modality experts introduced in Multiway Transformers effectively capture modality-specific information and improve performance.

Transformer	VQA	NLVR2	F30K	IN1K
Standard	76.1	80.8	82.8	84.1
Multiway	**76.8**	**81.4**	**84.4**	**84.4**

Table 1. Multiway Transformer improves the performance over the conventional one.

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 2, separate masking outperforms joint masking on vision-language tasks and learns the alignment of images and texts more effectively. Two masking strategies perform similarly on ImageNet-1K.

Strategy	VQA	NLVR2	F30K	IN1K
Joint	75.7	79.0	83.1	**84.4**
Separate	**76.8**	**81.4**	**84.4**	**84.4**

Table 2. Separate masking in MVLM is helpful.

Monomodal and Multimodal Data We analyze the effects of monomodal and multimodal data in Table 3. Experimental results indicate that monomodal and multimodal data positively contribute to performance. Using both types of pretraining data achieves the best results.

Mono	Multi	VQA	NLVR2	F30K	IN1K
✓	✗	71.3	64.6	79.3	84.1
✗	✓	75.8	79.3	81.1	83.4
✓	✓	**76.8**	**81.4**	**84.4**	**84.4**

Table 3. Whether we conduct masked prediction for monomodal (mono) and multimodal (multi) data.

Image Reconstruction Target We compare different targets used for image reconstruction. As shown in Table 4, VQ-KD$_{\text{CLIP}}$ [14] performs better than the DALL-E [15] tokenizer used in BEiT [1] and per-patch-normalized pixels proposed by MAE [6]. In addition, we observe training instability and gradient imbalance between image reconstruc-

* Equal contribution. † Corresponding author.

tion loss of per-patch-normalized pixels and text reconstruction loss, which results in a performance drop on ImageNet-1K.

Target	VQA	NLVR2	F30K	IN1K
DALL-E [15]	73.2	77.7	76.6	82.7
Pixel (w/ norm) [6]	73.3	77.1	75.9	81.1
VQ-KD$_{\text{CLIP}}$ [14]	**76.8**	**81.4**	**84.4**	**84.4**

Table 4. Targets used for image reconstruction. VQ-KD$_{\text{CLIP}}$ [14] works the best.

Text Reconstruction We study the effects of text reconstruction on monomodal and multimodal data. As shown in Table 5, the text reconstruction tasks on monomodal and multimodal data bring improvements for vision-language tasks. 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 for vision-language tasks. We also observe that introducing text reconstruction results in a slight performance drop on ImageNet-1K. Using shared attention parameters between different modalities helps the model to align different modalities. While model capacity is constrained due to the shared parameters, especially for the base-size model. We perform architecture explorations on Table 16 and find that decoupling attention module of different modalities relieves the above issue.

Mono	Multi	VQA	NLVR2	F30K	IN1K
✗	✗	71.5	69.3	77.8	**84.7**
✓	✗	73.2	76.4	81.3	84.4
✗	✓	76.5	80.6	82.7	84.6
✓	✓	**76.8**	**81.4**	**84.4**	84.4

Table 5. Whether we enable text reconstruction for monomodal (mono) and multimodal (multi) data.

Image Reconstruction Table 6 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 on vision-language tasks.

Mono	Multi	VQA	NLVR2	F30K	IN1K
✗	✗	71.6	74.3	71.7	77.9
✓	✗	75.8	79.8	82.0	84.3
✗	✓	75.6	79.5	81.9	83.3
✓	✓	**76.8**	**81.4**	**84.4**	**84.4**

Table 6. Whether we enable image reconstruction for monomodal (mono) and multimodal (multi) data.

2. Effects of Intermediate Finetuning for Retrieval

As shown in Table 7, we directly finetune BEiT-3 on COCO and Flickr30K. BEiT-3 still outperforms previous state-of-the-art models, even without using image-text contrastive objective during pretraining. The results demonstrate the effectiveness of masked data modeling for learning cross-modal representations. Next, we perform intermediate finetuning on the pretraining image-text pairs for 5 epochs with a 16k batch size. The peak learning is 3e-5, with linear warmup over the first epoch. The image input size is 224×224. The weight decay is set to 0.05. We disable dropout as in pretraining and use drop path with a rate of 0.3. The layer-wise learning rate decay is 0.95. We use the AdamW [11] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$.

3. Hyperparameters Used for Pretraining

Hyperparameters	BEiT-3
Layers	40
Hidden size	1408
FFN inner hidden size	6144
Attention heads	16
Patch size	14×14
Relative positional embeddings	✗
Training steps	1M
Batch size	6144
AdamW ϵ	1e-6
AdamW β	(0.9, 0.98)
Peak learning rate	1e-3
Learning rate schedule	Cosine
Warmup steps	10k
Gradient clipping	3.0
Dropout	✗
Drop path	0.1
Weight decay	0.05
Data Augment	RandomResizeAndCrop
Input resolution	224^2
Color jitter	0.4

Table 8. Hyperparameters for pretraining BEiT-3.

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
BEIT-3	82.7	96.0	98.2	65.1	86.6	92.3	97.5	99.9	100.0	89.1	98.6	99.3
+ Intermediate Finetuning	**84.8**	**96.5**	**98.3**	**67.2**	**87.7**	**92.8**	**98.0**	**100.0**	**100.0**	**90.3**	**98.7**	**99.5**

Table 7. Finetuning results of image-text retrieval on COCO and Flickr30K. BEIT-3 is directly finetuned on downstream benchmarks without intermediate finetuning on the pretraining data.

4. Hyperparameters Used for Finetuning on NLVR2 and VQAv2

Hyperparameters	NLVR2	VQAv2
Peak learning rate	1e-3	1e-5
Fine-tuning epochs	20	10
Warmup epochs	5	1
Layer-wise learning rate decay	0.8	1.0
Batch size	256	128
AdamW ϵ	1e-8	
AdamW β	(0.9, 0.999)	
Weight decay	0.05	0.01
Drop path	0.4	
Dropout	✗	
Input resolution	224^2	756^2

Table 9. Hyperparameters for fine-tuning BEIT-3 on NLVR2 and VQAv2.

5. Hyperparameters Used for Finetuning on COCO Captioning

Hyperparameters	COCO Captioning
Peak learning rate	8e-6
Fine-tuning steps	16k
Warmup steps	1600
Layer-wise learning rate decay	1.0
Batch size	256
AdamW ϵ	1e-8
AdamW β	(0.9, 0.999)
Weight decay	0.01
Drop path	0.3
Dropout	✗
Input resolution	392^2
Mask prob	0.6
Label smoothing ε	0.1
Beam size	3

Table 10. Hyperparameters for fine-tuning BEIT-3 on COCO captioning.

6. Hyperparameters Used for Finetuning on Image-Text Retrieval

Hyperparameters	COCO	Flickr30K
Peak learning rate	1e-5	
Fine-tuning epochs	15	20
Warmup epochs	3	5
Layer-wise learning rate decay	0.95	
Batch size	3k	
AdamW ϵ	1e-8	
AdamW β	(0.9, 0.999)	
Weight decay	0.05	
Drop path	0.3	
Dropout	✗	
Input resolution	420^2	

Table 11. Hyperparameters for fine-tuning BEIT-3 on image-text retrieval.

7. Hyperparameters Used for Finetuning on Semantic Segmentation

Hyperparameters	ADE20K
Peak learning rate	1e-5
Fine-tuning steps	80k
Warmup steps	1500
Layer-wise learning rate decay	0.95
Batch size	16
AdamW ϵ	1e-8
AdamW β	(0.9, 0.999)
Weight decay	0.05
Drop path	0.5
Dropout	✗
Input resolution	896^2

Table 12. Hyperparameters for fine-tuning BEIT-3 on semantic segmentation.

8. Hyperparameters Used for Finetuning on Object Detection

Hyperparameters	Object365	COCO
Learning rate	1e-4	5e-5
Fine-tuning epochs	15	20
Warmup steps	250	
Layer-wise learning rate decay	0.9	
Batch size	64	
AdamW ϵ	1e-8	
AdamW β	(0.9, 0.999)	
Weight decay	0.1	
Drop path	0.6	
Input resolution	1024^2	1280^2

Table 13. Hyperparameters for fine-tuning BEIT-3 on object detection.

9. Hyperparameters Used for Finetuning on Image Classification

Hyperparameters	ImageNet-21K	ImageNet-1K
Peak learning rate	5e-5	3e-5
Fine-tuning epochs	50	15
Warmup epochs	5	3
Layer-wise learning rate decay	0.85	0.95
Batch size	16k	2k
AdamW ϵ	1e-6	1e-8
AdamW β	(0.9, 0.98)	(0.9, 0.999)
Weight decay	0.05	
Drop path	0.4	
Dropout	✗	
Input resolution	224^2	336^2
Label smoothing ε	0.1	

Table 14. Hyperparameters for fine-tuning BEIT-3 on image classification.

10. Video Downstream Tasks

We evaluate a base-size BEIT-3 model on video retrieval (MSR-VTT [19]) and action recognition (Kinetics-400 [7]) tasks. The results are present in Table 15. We directly adopt the framework of X-CLIP [12] for Kinetics-400 and keep all the hyperparameters, except the learning rate, the same for a fair comparison. For MSR-VTT, we evaluate the zero-shot text-to-video retrieval result of a BEIT-3 checkpoint after intermediate image-text contrastive finetuning. We follow VIOLET [5] and use the same protocol. Table 15 shows that BEIT-3 achieves better performance than CLIP on both two tasks.

Model	K400 (Top1 Acc)	MSR-VTT (R@1)
CLIP Base	83.8	30.0
BEIT-3 Base	**84.2**	**30.7**

Table 15. Finetuning results on Kinetics-400 (K400) and zero-shot text-to-video retrieval results on MSR-VTT 1K-A test set.

11. Additional Architecture Exploration

We perform architecture exploration on decoupling attention parameters of different modalities and introducing MAGNETO [18]. Multiway Transformers use a shared self-attention module between different modalities to enable the model to be used for vision-language tasks requiring deep fusion. While the shared attention parameters limit the model capacity for different modalities. We explore encoding different modalities using different attention parameters, and fuse image-text pairs via concatenating queries, keys, and values of images and texts in the self-attention module to model their interactions. As present in Table 16, decoupling the self-attention module improves model capacity and brings improvements to the vision task (ImageNet-1K) and language task (SST-2). It also achieves similar performance on vision-language tasks. Moreover, introducing MAGNETO brings further improvements across different downstream tasks.

Architecture	VQA	IN1K	SST-2
Multiway Transformer	76.8	84.4	92.6
Decoupled Transformer	76.8	84.7	92.8
+ MAGNETO [18]	**77.5**	**84.9**	**93.5**

Table 16. Architecture exploration of decoupling self-attention module and introducing MAGNETO [18].

12. Model Configuration

We scale up the model capacity of BEIT-3 to a giant-size Transformer model following the setup of ViT-giant [20]. As shown in Table 17, the 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.

Model	#Layers	Hidden Size	MLP Size	#Parameters				
				V-FFN	L-FFN	VL-FFN	Shared Attention	Total
BEiT-3	40	1408	6144	692M	692M	52M	317M	1.9B

Table 17. Model configuration of BEiT-3. The architecture layout follows ViT-giant [20].

13. Data Statistics

BEiT-3 is pretrained on both monomodal and multimodal data shown in Table 18. For multimodal data, there are about 15M images and 21M image-text pairs collected from five public datasets: Conceptual 12M (CC12M) [3], Conceptual Captions (CC3M) [16], SBU Captions (SBU) [13], COCO [9] and Visual Genome (VG) [8]. For monomodal data, we use 14M images from ImageNet-21K and 160GB text corpora [2] from English Wikipedia, BookCorpus [21], OpenWebText[1], CC-News [10], and Stories [17].

Data	Source	Size
Image-Text Pair	CC12M, CC3M, SBU, COCO, VG	21M pairs
Image	ImageNet-21K	14M images
Text	English Wikipedia, BookCorpus, OpenWebText, CC-News, Stories	160GB documents

Table 18. Pretraining data of BEiT-3. All the data are academically accessible.

References

[1] 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] 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. 5

[3] 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. 5

[4] 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

[5] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. VIOLET : End-to-end video-language transformers with masked visual-token modeling. *CoRR*, abs/2111.12681, 2021. 4

[6] 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. 1, 2

[7] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 4

[8] 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. 5

[9] 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. 5

[10] 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. 5

[11] 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. 2

[12] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, 2022. 4

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

[13] 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. 5

[14] 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

[15] 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. 1, 2

[16] 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. 5

[17] Trieu H. Trinh and Quoc V. Le. A simple method for commonsense reasoning. *ArXiv*, abs/1806.02847, 2018. 5

[18] Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, and Furu Wei. Foundation transformers. *CoRR*, abs/2210.06423, 2022. 4

[19] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society, 2016. 4

[20] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021. 4, 5

[21] 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. 5