Sample- *and* Parameter-Efficient Auto-Regressive Image Models

Elad Amrani
Apple, Technion

Leonid Karlinsky
MIT-IBM Watson AI Lab

Alex Bronstein
Technion

https://github.com/elad-amrani/xtra

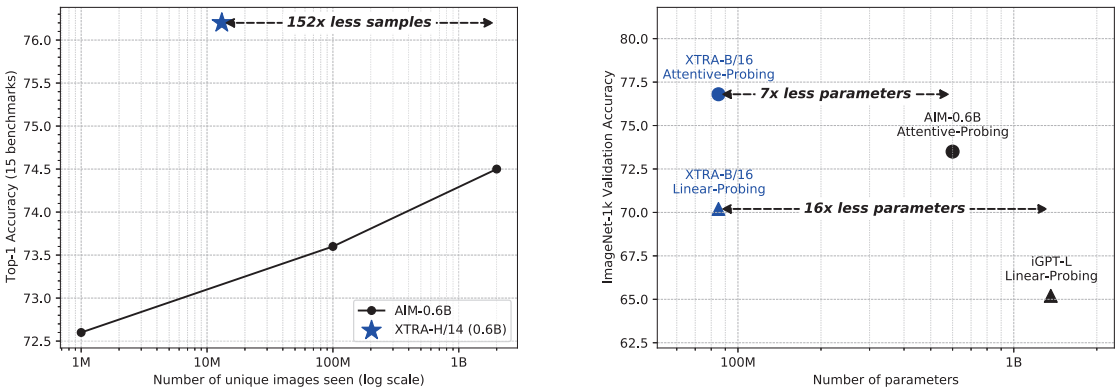

Figure 1. **Sample and Parameter Efficiency of XTRA.** (Left) XTRA-H/14 (0.6B parameters) outperforms prior state-of-the-art auto-regressive image model (AIM-0.6B [26]) in top-1 average accuracy across 15 diverse image recognition benchmarks, despite being trained on 152× fewer samples. (Right) XTRA-B/16 (85M parameters) outperforms prior auto-regressive image models trained on ImageNet-1k in linear and attentive probing tasks, while using 7–16× fewer parameters.

Abstract

*We introduce **XTRA**, a vision model pre-trained with a novel auto-regressive objective that significantly enhances both sample and parameter efficiency compared to previous auto-regressive image models. Unlike contrastive or masked image modeling methods, which have not been demonstrated as having consistent scaling behavior on unbalanced internet data, auto-regressive vision models exhibit scalable and promising performance as model and dataset size increase. In contrast to standard auto-regressive models, XTRA employs a Block Causal Mask, where each Block represents $k \times k$ tokens rather than relying on a standard causal mask. By reconstructing pixel values block by block, XTRA captures higher-level structural patterns over larger image regions. Predicting on blocks allows the model to learn relationships across broader areas of pixels, enabling more abstract and semantically meaningful representations than traditional next-token prediction. This simple modification yields two key results. First, **XTRA is sample-efficient**. Despite being trained on 152× fewer samples (13.1M vs. 2B), XTRA ViT-H/14 surpasses the top-1 average accuracy of the previous state-of-the-art auto-regressive model across 15 diverse image recognition benchmarks. Second, **XTRA is parameter-efficient**. Compared to auto-regressive models trained on ImageNet-1k, XTRA ViT-B/16 outperforms in linear and attentive probing tasks, using 7–16× fewer parameters (85M vs. 1.36B/0.63B).*

1. Introduction

Auto-regressive models have played a foundational role in recent advancements in Natural Language Processing (NLP). The simplicity of their objective — predicting the next word in a sequence based on its preceding context — allows these models to capture intricate dependencies and complex patterns over long sequences. This next-token prediction mechanism

Elad Amrani conducted this work while at IBM Research-AI.

has proven highly effective, leading to the development of models capable of sophisticated language understanding and generation. Most importantly, auto-regressive language models [10, 25, 38, 44, 45] demonstrate desirable scaling properties, where downstream performance improves consistently as both model capacity and data size grow [31].

In contrast, the progress in the field of Computer Vision (CV) has been driven mostly by Contrastive Learning (CL) methods [1, 11, 12, 14–17, 23, 28, 48], which aim to learn visual representations by maximizing the similarity of two different augmentations of the same image while simultaneously minimizing the similarity between different images; Masked Image Modeling (MIM) methods [2–4, 6, 19, 29, 49], which focus on predicting masked parts of an image based on its visible regions; and various hybrid approaches that combine elements of both [35, 50]. These methods, and specifically the hybrid approaches, set the current state-of-the-art performance for self-supervised visual representation learning. Despite their success, these methods often rely on intricate training recipes involving numerous tricks, such as multi-crop *handcrafted* augmentations, momentum networks, schedules for teacher momentum and weight decay, and complex regularization techniques like KoLeo [40] and LayerScale [43]. These modifications can introduce significant overhead, both in terms of computational resources and implementation complexity. While some methods, like DINOv2 [35], have shown promising scaling behavior, they lack the consistent scaling laws [31] that are a hallmark of auto-regressive models. For example, [41] demonstrated that, even when scaling the pre-training dataset from 1M to 3B samples using MAE ViT-H [29], the resulting improvement on downstream tasks is modest, with only a 0.5% gain on ImageNet1k [22] and a 1.2% gain on iNAT-18 [46]. This limited scaling effectiveness can impede these models' capacity to sustain performance gains as they increase in size or are trained on larger, uncurated internet datasets, in contrast to the more predictable scaling benefits observed in auto-regressive language models.

Auto-regressive image models [13, 26], like their language counterparts, predict image pixels (or patches) sequentially based on the preceding context. These models aim to learn visual representations by capturing dependencies across pixel or patch sequences. First, iGPT [13] demonstrated the feasibility of self-supervised visual representation learning using an auto-regressive model that predicts the next pixel. Later, AIM [26], a model based on Vision Transformers (ViT), demonstrated that auto-regressive models for images can scale similarly to their NLP counterparts, offering a consistent relationship between the model's objective function and its downstream task performance. This characteristic is crucial for scalable visual representation learning, as it enables predictable improvements as model size or dataset size increases. However, despite their promising scaling properties, both iGPT and AIM suffer from significant sample- and parameter-efficiency drawbacks. iGPT, for example, required 7 billion parameters to achieve results on par with contrastive models that operate with 20 times fewer parameters. Similarly, AIM was trained on a massive dataset of 2 billion samples, whereas contrastive and MIM models can achieve competitive results with datasets that are 150 times smaller. This inefficiency poses a substantial barrier to their widespread adoption in resource-constrained environments. These limitations highlight the need for more efficient auto-regressive models for visual tasks.

In this work, we propose XTRA, an auto-regressive vision model that leverages a Block Causal Mask to enhance sample and parameter efficiency. By employing Block Causal Masking, XTRA more effectively utilizes its modeling capacity to capture low-frequency structures essential for object recognition, rather than focusing on high-frequency details. Empirical results demonstrate that this approach enables XTRA to learn abstract and semantically meaningful representations using less data and smaller model sizes.

The key contributions of this paper are:

- **High sample efficiency.** Although trained on 152× fewer samples (13.1M vs. 2B), XTRA ViT-H/14 outperforms the previous state-of-the-art auto-regressive model of the same size in top-1 average accuracy across 15 diverse image recognition benchmarks.
- **High parameter efficiency.** XTRA ViT-B/16 outperforms auto-regressive models trained on ImageNet-1k in linear and attentive probing tasks, while using 7–16× fewer parameters (85M vs. 1.36B/0.63B).

2. Related Work

2.1. Contrastive Self-Supervised Learning

Contrastive Learning methods aim to maximize the similarity of two *handcrafted* augmentations (also called views) of a given image, while preventing collapse. Collapse is defined as the trivial solution where all images in the dataset are assigned the same vector representation. The various methods differ by the way they prevent collapse. For instance, SimCLR [14, 15] tackles collapse by utilizing negative pairs, explicitly minimizing the similarity between different images. MoCo [17, 28], BYOL [27] and DINO [12], on the other hand, employ a momentum encoder (also utilized beyond contrastive methods) to mitigate collapse. SwAV [11] takes a different route by relying on an external clustering algorithm to prevent collapse effectively. Self-Classifier [1], meanwhile, counters collapse with the use of an explicit uniform prior. Lastly, SimSiam [16] employs a stop-gradient operation on one of the views as its means to avert collapse.

2.2. Generative Self-Supervised Learning

Generative self-supervised learning aims to learn representations that enable the prediction of masked regions within a sample based on the unmasked portions. While this paradigm has seen remarkable success in Natural Language Processing [10, 32, 38], it has recently made its way into the domain of

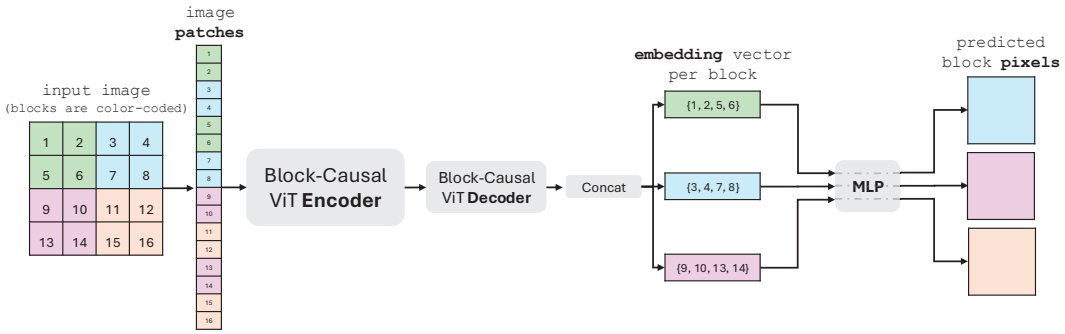

Figure 2. **XTRA Architecture.** Following ViT [24] an image is partitioned into a sequence of patches (numbered grid) and processed by a standard ViT encoder-decoder architecture with our proposed Block Causal Masking. I.e., causality is enforced at the block level with a rasterized pattern (see Figure 3 for detailed explanation). In the example above a block represents 2×2 patches/tokens. The token representations within each block at the output of the decoder are concatenated in a predetermined order such that each block of pixels is represented by a single embedding vector. Finally, each block embedding vector is passed through an MLP (same MLP for all blocks) to predict **all** pixel values of the next block in the sequence.

computer vision. Generative learning approaches in Computer Vision can be categorized into two categories, each with its own techniques and objectives. These two categories are further detailed in the following subsections.

2.2.1. Masked Image Modeling

Following the introduction of the Vision Transformer (ViT) by [24], BEiT [6] extended the concept of masked language modeling from NLP to visual tasks via a Masked Image Modeling (MIM) approach, inspired by BERT [32]. BEiT partitions an image into a grid of patches and tokenizes each patch into discrete visual tokens using latent codes obtained from a pre-trained discrete variational autoencoder [39] (dVAE). The objective is to predict the masked visual tokens based on the unmasked patches, mirroring BERT's approach of recovering masked words using the surrounding text context. iBOT [50] built on BEiT by replacing the fixed dVAE tokenizer with an online tokenizer, learned through a momentum encoder that is updated during training. Furthermore, iBOT integrates contrastive learning, minimizing the similarity between two augmented views of the same image to improve performance. Other BERT-inspired methods, such as MAE [29] and SimMIM [49], deviate from BEiT by abandoning discrete tokenization. Instead, they focus on directly predicting the pixel values of masked patches. MAE, in particular, adopts an asymmetric encoder-decoder architecture, where the encoder processes only visible patches, and a lightweight decoder reconstructs the missing patches from the latent representation and mask tokens. This approach has proven highly effective, demonstrating that pixel-level reconstruction can achieve impressive results in self-supervised visual learning. Further MIM approaches, such as data2vec [3, 4] and I-JEPA [2], move away from pixel reconstruction loss and instead predict the *latent representations* of masked regions based on unmasked parts. These methods rely on a momentum encoder to avoid collapse—preventing all patches from being assigned the same representation. Lastly, Context Autoencoder [19] (CAE) blends two pre-training tasks: (1) masked patch reconstruction (as seen in MAE and SimMIM) and (2) masked representation prediction (as seen in I-JEPA and data2vec).

2.2.2. Auto-Regressive Image Modeling

One of the early auto-regressive models for computer vision is iGPT [13] (Image-GPT). It is a sequence Transformer designed for auto-regressively predicting pixel values in images by minimizing the negative log-likelihood over a fixed set of pixel values. It represents a straightforward yet effective extension of GPT [38] to image pixels. Although it did not surpass the performance of state-of-the-art Contrastive Learning methods, the main contribution of iGPT was demonstrating that the auto-regressive paradigm can be smoothly extended from language to vision. Primary limitations of iGPT, however, lie in its memory and computation requirements. First, due the quadratic nature of self-attention, processing a sequence of *pixels* in an image is both memory and computationally expensive. Second, the sample- and parameter-efficiency of iGPT is considerably low in comparison to other self-supervised methods, as it requires more than 15× more parameters and more than 78× more samples for results to be competitive with self-supervised benchmarks on ImageNet. A more recent work is AIM [26] (Auto-regressive Image Models). AIM applies the standard auto-regressive objective using a causal Vision Transformer to a sequence of non-overlapping image patches (where a patch is embedded linearly following ViT), and predicts the pixel values of the next patch by minimizing the Mean Squared Error loss. The main contribution of AIM was showing that auto-regressive *image* models exhibit similar scaling properties to auto-regressive *language* models. I.e., the performance of the learned visual features scale with both the model capacity and the quantity of

Figure 3. **Block Causal Masking.** In the image above the fine-grained grid represents a grid of patches (following ViT [24]) that are processed by the model. The coarse-grained (numbered) grid represents a grid of blocks, where each block represents 4×4 patches/tokens. Block Causal Masking enforces causality at the block level with a rasterized pattern (numbered sequence), ensuring that tokens can attend to others within the same block and also to preceding blocks.

data, and the value of the objective function correlates with the performance of the model on downstream tasks. Yet, similarly to iGPT, the sample- and parameter-efficiency, though improved, is still low. Specifically, when trained on ImageNet-1k only, AIM requires more than 7× more parameters in comparison to other Contrastive and Masked Image Modeling methods for competitive results (see Tab. 4). Additionally, for same model size (e.g., ViT-H/14), AIM requires more than 150× more samples for achieving stronger results on a diverse set of 15 image recognition benchmarks (see Tab. 3).

Building upon the foundation set by iGPT and AIM, our approach introduces an auto-regressive image model that leverages a Vision Transformer architecture with a novel Block Causal Mask (see Sec. 3). By processing blocks of image patches rather than individual patches or pixels, XTRA makes more efficient use of its modeling capacity to capture low-frequency structures that enhance object recognizability, rather than focusing on redundant high-frequency details. Empirically, this design allows us to substantially enhance sample- and parameter-efficiency (Tab. 3 and Tab. 4) while maintaining the simplicity and scalability that characterize auto-regressive models.

3. Method

This section provides details on the XTRA architecture (Figure 2), including its use of *Block Causal Masking* (Figure 3), the training objective and loss function.

Architecture. XTRA is an encoder-decoder network, where a Vision Transformer [24] (ViT) with *Block Causal Masking* is used for both components.

Block Causal Masking. Block Causal Masking, visualized in Figure 3, structures attention such that the image is divided into spatial blocks of $k \times k$ tokens, with causality enforced at the block level with a rasterized pattern. This structure allows for: (1) auto-regressive modeling with image regions bigger than a single patch size ; and (2) efficient modeling of both local (within-block) and global (cross-block) dependencies, ensuring that tokens can attend to others within the same block and also to preceding blocks.

Next Block Reconstruction. At the output of the decoder, the token representations within each block are concatenated in a predetermined order (e.g., raster order) and passed through a fully-connected MLP to reconstruct the pixel values of the next block. This block-wise reconstruction strategy leverages the structured attention, enabling accurate pixel prediction for subsequent blocks based on previously processed regions.

Training Objective & Loss. The training objective employs a standard auto-regressive approach, generating predictions sequentially, where each block's prediction is based solely on previously observed blocks. The loss function used is the Mean Squared Error (MSE) applied to pixel values normalized per block, as per [26, 29]:

$$\ell(\theta) = \frac{1}{N(K-1)} \sum_{n=1}^{N} \sum_{k=2}^{K} \| \hat{x}_k^n(\theta; x_{<k}^n) - x_k^n \|_2^2, \qquad (1)$$

where, θ represents the network's parameters, N is the batch size, K denotes the number of blocks in an image, x_k^n is the ground truth pixel values for the k-th block in the n-th image, and $\hat{x}_k^n(\theta; x_{<k}^n)$ represents the reconstructed values based on the network parameters (θ) and the preceding blocks in the sequence for the same image ($x_{<k}^n$).

4. Implementation Details

Default setting for the pre-training stage is in Table 1. Default setting for the attentive probing stage is in Table 2. Our training hyper-parameters are taken from AIM [26].

5. Quantitative Results

In this section, we evaluate the performance of XTRA in comparison to state-of-the-art methods on two probing tasks: linear (LIN) and attentive (ATT). The linear probing task assesses the quality of the frozen pre-trained features by training a simple linear classifier, while the attentive probing task evaluates the ability to learn a lightweight attention mechanism over these frozen representations. These tasks

config	ViT-B/16	ViT-H/14
Optimizer	AdamW	
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$	
Peak learning rate	$1e^{-3}$	
Minimum Learning rate	$1e^{-6}$	
Weight decay	0.05	
Batch size	2048	
Patch size	16×16 px	14×14 px
Block size	64×64 px	56×56 px
Decoder width	768	640
Decoder depth	8	
Gradient clipping	1.0	
Drop path rate	0.2	
Dataset	ImageNet-1K	ImageNet-21K
Warmup epoch	15	3
Total epochs	800	100
Learning rate schedule	cosine decay	
Augmentations:		
`RandomResizedCrop`		
`size`	256px	224px
`scale`	[0.3, 1.0]	
`ratio`	[0.75, 1.33]	
`RandomHorizontalFlip`	$p = 0.5$	

Table 1. **Pre-training hyperparameters.**

config	IN-1k	Others
Optimizer	AdamW	
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	
Peak learning rate grid	[1, 3, 5, 10, 15, 20, 40, 80] $\times 1e^{-4}$	
Minimum Learning rate	$1e^{-5}$	
Weight decay	0.1	
Batch size	4k	[128, 256, 512]*
Gradient clipping	3.0	
Drop path rate	0.0	
Warmup epochs	10	0
Epochs	100	
Learning rate schedule	cosine decay	
Augmentations:		
`RandomResizedCrop`		
`size`	224px	
`scale`	[0.08, 1.0]	
`ratio`	[0.75, 1.33]	
`RandomHorizontalFlip`	$p = 0.5$	
`AutoAugment`	`rand-m9-mstd0.5-inc1`	

Table 2. **Attentive probe hyperparameters.** $*$: Small, medium and large datasets (excluding imagenet) used batch sizes of 128, 256 and 512, respectively.

provide a comprehensive measure of how effectively the pre-trained representations can be leveraged for downstream image recognition tasks with minimal fine-tuning.

5.1. Transfer Learning

Table 3 summarizes the results of attentive probing across 15 diverse image recognition benchmarks. These benchmarks cover a wide range of domains, including fine-grained recognition, medical imaging, satellite imagery, natural environments, and infographic images. This diversity highlights XTRA's robustness and adaptability across varied visual tasks. For detailed hyperparameters see Table 2. Specific details of each benchmark dataset are in Table 7 in the Appendix.

XTRA-H (ViT-H/14, 632M parameters), pre-trained on ImageNet-21K (filtered to 13.1M samples due to broken URL links), achieves superior or competitive performance on 9 out of 15 datasets, setting a new benchmark for generative models with the highest average accuracy across tasks. Notably, XTRA outperforms the previous state-of-the-art auto-regressive model, AIM-0.6B [26], by 1.7% when compared to AIM-0.6B trained on DFN-2B and by 0.6% against AIM-0.6B trained on DFN-2B+ (80% DFN-2B and 20% ImageNet-1K). Despite training on 152× fewer samples (13.1M vs. 2B), XTRA's sample efficiency enables it to learn rich, semantically meaningful representations, which lead to these improvements.

5.2. ImageNet-1K

Here, we focus on models trained and evaluated exclusively on ImageNet-1K.

In Table 4, we compare XTRA-B (ViT-B/16, 85M parameters) to prior auto-regressive image models. Despite having 16× fewer parameters, XTRA-B achieves a 5.0% higher accuracy in the linear probing task compared to iGPT-L (1.36B parameters). In the attentive probing task, XTRA-B outperforms AIM-0.6B (0.63B parameters) by 3.3%, using 7.5× fewer parameters. These results demonstrate XTRA's high parameter efficiency, delivering superior performance with a fraction of the model size, which translates to faster inference times and lower computational costs.

Next, we compare XTRA-B with other state-of-the-art models that utilize the same ViT-B/16 backbone (Table 5). To ensure fair comparisons, we limit the analysis to models trained with single-crop views and without image-specific augmentations or multi-crop training. Under these conditions, XTRA-B establishes a new state-of-the-art in both linear and attentive probing tasks on ImageNet-1K. Consistent improvements are observed across training durations of 300 and 800 epochs. Notably, XTRA-B trained for 800 epochs surpasses both data2vec and MAE, which require 1600 epochs to train, further highlighting XTRA's ability to achieve competitive results with significantly fewer training resources.

These findings underscore the advantages of the auto-regressive pre-training objective used by XTRA, particularly in terms of sample and parameter efficiency. XTRA achieves competitive results with reduced training time and computational resources, making it a highly efficient approach compared to previous methods.

6. Ablation Study

In this section, we evaluate the impact of the main components of XTRA. For each experiment, we pre-train a ViT-B/16 model for 100 epochs, followed by a supervised attentive probing training stage using the ImageNet-1K training set. We report the attentive probing accuracy on the ImageNet-1K validation set in Table 6.

Model	Arch.	Data	Cost†	IN-1k	iNAT-18	Cifar10	Cifar100	Food101	DTD	Pets	Cars	iWildCam	Camelyon17	PCAM	RxRX1	EuroSAT	fMoW	Infographic	Avg
methods using extra view data augmentations																			
DINO [12]	ViT-B/8	IN-1k	19.2	80.1	66.0	97.8	87.3	89.5	78.4	92.3	89.2	58.5	93.7	90.2	6.1	98.2	57.0	41.1	75.0
iBOT [50]	ViT-L/16	IN-21k	1.4	83.5	70.5	99.2	93.3	93.5	81.6	92.8	90.8	61.8	94.5	90.0	5.9	98.0	60.3	47.7	77.6
methods without view data augmentations																			
— masked image modeling methods																			
BEiT [6]	ViT-L/14	IN-21k	**4.2**	62.2	44.4	94.4	78.7	79.0	64.0	80.9	69.5	52.0	92.8	88.2	4.2	97.5	47.7	25.9	65.4
MAE-H [29]	ViT-H/14	IN-1k	8.0	**80.9**	64.6	97.1	85.8	90.2	78.1	**95.0**	**93.7**	58.1	94.2	89.8	5.4	98.1	56.9	42.2	75.3
— auto-regressive image modeling methods																			
AIM-0.6B [26]	ViT-H/14	DFN-2B	20.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	74.5
AIM-0.6B [26]	ViT-H/14	DFN-2B+	20.7	78.5	64.0	97.2	86.8	90.1	**80.1**	93.0	93.0	57.9	94.3	**90.0**	**7.8**	98.4	58.3	**45.2**	75.6
XTRA-H (ours)	ViT-H/14	IN-21k	5.8	**80.9**	**67.0**	**98.2**	**90.0**	**90.8**	79.7	93.7	93.1	**59.5**	93.3	**90.0**	5.7	**98.5**	**58.6**	43.9	**76.2**

Table 3. **Downstream evaluation with a frozen trunk.** Similarly to AIM, We assess the quality of XTRA by evaluating against a diverse set of 15 image recognition benchmarks (specific details of each dataset are in Table 7 in the Appendix). XTRA and the baseline methods are evaluated using attentive probing with a frozen trunk. The attentive probing results for all other methods are from AIM [26]. †: Computational cost is estimated using $Parameters \times Samples \times Epochs \times Views^2 \times Tokens^2$ (see Section 11 in Appendix for detailed explanation of this formula).

Method	# Params (M)	LIN	ATT
iGPT-L† [13]	1362	65.2	-
AIM-0.6B [26]	600	-	73.5
XTRA (ours)	85	**70.2**	**76.8**

Table 4. **Comparison to previous Auto-Regressive Image Modeling methods**. LIN: linear probing accuracy. ATT: attentive probing accuracy. All models were pre-trained and evaluated with ImageNet-1k. †: iGPT use linear probing by concatenating the output of 5 layers, which indirectly inflates the capacity of the evaluation head.

Block size & multi-block prediction. Block size is a critical component of XTRA. Table 6a shows results for various block sizes: 16×16, 32×32, and 64×64 pixels, corresponding to a single token/patch, 2×2 tokens, and 4×4 tokens, respectively. We also experimented with 'multi-block' prediction, where multiple blocks are predicted based on prior context (e.g., predicting the next two blocks). As shown, this does not significantly impact the results, except in the case of a single token versus two tokens (as seen in the first row of the table). When the block size is set to a single patch/token (16×16 pixels), it replicates the standard auto-regressive image model (AIM [26]), with *all* other training details kept identical. Notably, increasing the block size has a substantial effect on performance, with a 3.0% accuracy improvement when comparing a block size of 16×16 pixels (AIM re-implementation) to 64×64 pixels (XTRA). This demonstrates the clear benefits of XTRA and our proposed Block Causal Masking in auto-regressive image models.

Impact of block size relative to resolution. To further explore the impact of block size, we evaluate various block sizes in relation to the image resolution, as shown in Table 6b. Specifically, we compare two different resolutions (224 and 256 pixels) with corresponding patch sizes of 14 and 16 pixels, respectively. We observe that when the block size to resolution ratio is kept constant across experiments (as reflected in each row of the table), the results remain consistent, particularly when the block size to resolution is sufficiently large (e.g., 4/256 or 16/256). This consistency holds even when the patch size, image size, and absolute block size (in pixels) differ, suggesting that for auto-regressive image models, the block size to resolution ratio is the key factor determining performance, rather than the absolute block size.

Loss function. In Table 6c, we compare L1 loss (Mean Absolute Error) with L2 loss (Mean Squared Error). The results show that L2 loss outperforms L1 loss (+1.0%).

Auto-regressive pattern. In Table 6d, we compare the simple raster pattern with a fixed random pattern. The results indicate that the raster pattern significantly outperforms the random pattern, with an improvement of 11.9%. We hypothesize that this large margin arises from the substantial difficulty introduced by randomizing the sequence of blocks in the auto-regressive prediction task. For example, certain random permutations may force the model to predict the top-left block of pixels based *solely* on the bottom-right block, an arrangement which,

Method	Epochs	LIN	ATT
methods using extra augmented views$^{\diamond}$			
MoCo-v3 [18]	300	76.2	77.0
DINO [12]	400	77.3	77.8
iBOT [50]	400	79.5	79.8
methods using extra masked views*			
I-JEPA [2]	600	70.9	-
StoP [7]	600	72.6	-
methods with a single view			
MAE [29]	300	61.5	71.1
CAE† [19]	300	64.1	73.8
XTRA (ours)	300	**66.1**	**74.3**
SimMIM [49]	800	56.7	-
MAE [29]	1600	67.8	74.2
Data2Vec [3]	1600	68.0	-
CAE† [19]	800	68.6	75.9
XTRA (ours)	800	**70.2**	**76.8**

Table 5. **ViT-B ImageNet-1k evaluation with a frozen trunk.** All models were trained and evaluated exclusively on ImageNet-1K. LIN: linear probing accuracy. ATT: attentive probing accuracy. $\star$: I-JEPA and StoP use multi-mask views. For each 1 context block mask, 4 target blocks masks are sampled; both results are taken from StoP [7] for linear probing using the single last layer. $\diamond$: MoCo-v3, DINO and iBOT use multi-crop *hand-crafted* view augmentations. MoCo-v3: 2 crops. DINO and iBOT: 12 crops. Thus, the number of effective epochs for both $\star$ and $\diamond$ is larger (equivalent to taking a larger number of epochs compared to one-crop augmentation). In contrast, Data2Vec, MAE, CAE and XTRA use a single view without any image-specific augmentations (i.e., only random crops). The attentive probing results for all other methods are from CAE [19]. †: denotes using the DALL-E tokenizer (trained with d-VAE on 400M images).

for non-object-centric images, may yield an unsolvable task and limit the model's ability to learn semantically meaningful representations. By contrast, predicting the next nearest block in raster order is generally feasible, even with only a single prior block as context, allowing for more coherent learning.

Decoder depth & width. Unlike previous auto-regressive image models (AIM and iGPT), which use only an encoder, XTRA is an encoder-decoder architecture. Although the decoder is not crucial to XTRA's success, it serves two purposes: (1) as observed in previous works (e.g., AIM and MAE), the final encoder layers often specialize in pixel reconstruction rather than semantic recognition, so adding a decoder can localize this reconstruction specialization, allowing the encoder output to retain more abstract representations; and (2) it enables down-sampling of the embedding size before pixel predictions, improving reconstruction efficiency. In Table 6e and Table 6f, we assess the impact of the decoder's depth and width. The results show that when the width is fixed at 384 (half the encoder's width), the decoder depth has little effect on performance. However, with a fixed depth, increasing the width to 768 (matching the encoder's width) yields a substantial 2.0% performance improvement. Notably: (1) as with previous methods, the decoder is used only during pre-training for block reconstruction, ensuring that linear/attentive probing comparisons with other methods remain fair; and (2) in Table 6a, we also re-implement AIM with an encoder-decoder structure, validating that XTRA's Block Causal Masking provides performance gains irrespective of architecture.

7. Discussion

Our model employs *Block Causal Masking* to improve both **sample efficiency** and **parameter efficiency** by structuring autoregressive attention to better align with the 2D structure of images.

Sample efficiency. Standard autoregressive models process tokens sequentially, making it inefficient to capture local spatial structure. In contrast, our block causal mask groups tokens into local regions, allowing **unrestricted intra-block attention** while enforcing causality between blocks. This design improves sample efficiency by explicitly modeling local pixel dependencies, which reduces the number of training samples required to learn meaningful representations. Additionally, by allowing free attention within each block, local structures are captured in fewer autoregressive steps, making representation learning faster and requiring less data to generalize effectively.

Parameter efficiency. The block-wise autoregressive factorization also enhances parameter efficiency by reducing the depth needed to model long-range dependencies. Since tokens within a block attend freely to each other, information propagates efficiently without requiring excessively deep hierarchies. This leads to a more compact and computationally efficient model, as local structures are captured without redundant layers or parameters.

Empirical Validation. These theoretical advantages manifest in practical improvements. As shown in Section 5, models with block causal masking achieve **higher accuracy with fewer training samples** and perform competitively even with fewer parameters compared to fully autoregressive baselines.

8. Qualitative Results

XTRA captures semantically meaningful representations by learning to predict unseen pixel values based on prior contextual information. In Figure 4, we present visualizations of the generative capabilities of XTRA to better understand the model's learned knowledge visually. Although XTRA's primary focus is on representation learning rather than generating photorealistic images, these visualizations offer valuable insights into the model's predictive reasoning. Notably, the

block size	# of blocks to predict 1	2
16×16	64.6	65.1
32×32	67.4	67.3
64×64	**67.6**	67.4

(a) **Block size (pixels) & multi-block prediction.**

block / res. ratio	res.|patch size 256|16	224|14
1/256	64.6	65.2
4/256	67.4	67.3
16/256	67.6	**67.7**

(b) **Block size relative to resolution.**

Function	Accuracy
L1 (MAE)	66.6
L2 (MSE)	**67.6**

(c) **Loss function.**

AR Pattern	Accuracy
Fixed Random	55.7
Raster	**67.6**

(d) **Auto-regressive pattern.**

blocks	Accuracy
1	67.4
2	**67.9**
4	67.6
8	67.6
16	67.8

(e) **Decoder depth.**

width	Accuracy
192	67.1
384	67.6
576	67.8
768	**69.6**

(f) **Decoder width.**

Table 6. **Ablation study.** Default settings for ablation baseline are marked in gray. Best in **bold**.

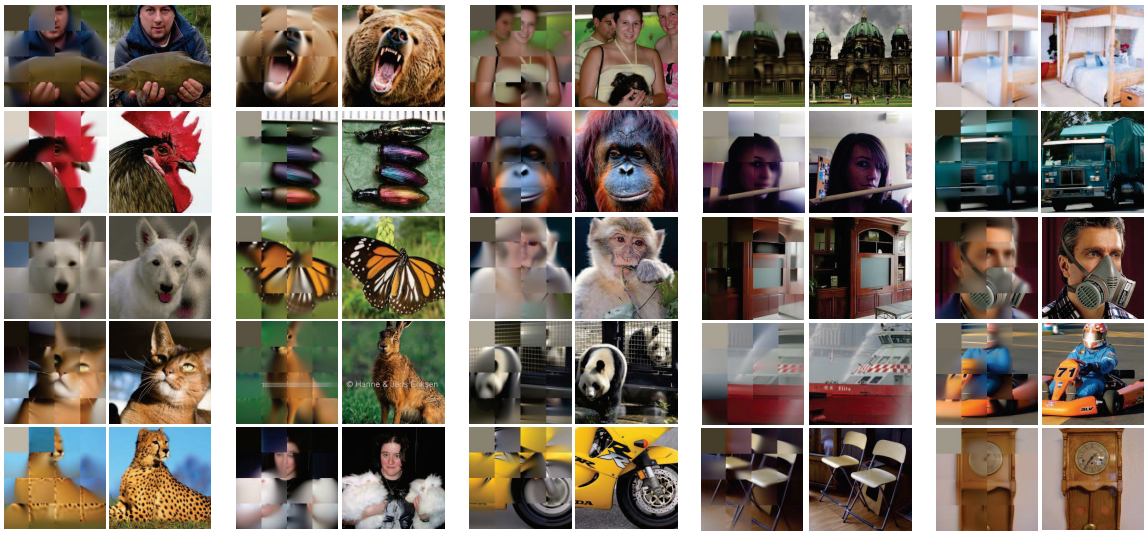

Figure 4. **Visualization of XTRA's Predictions on the ImageNet-1k Validation Set.** XTRA generates predictions auto-regressively, producing one block of pixels at a time, with each new block conditioned on the preceding sequence of ground-truth blocks. *Note that no loss is applied to the left upper block (first in sequence), yet it is of different color from image to image due to post-generation per block de-normalization, since loss is applied to normalized block pixels.*

effects of the Block Causal Mask are immediately observable: some blocks lack details of entire objects due to limited information from previous image blocks. Yet, even with these gaps, the predictions remain plausible given the current context.

9. Summary

Recent advancements in Deep Learning have focused on simple, efficient, and scalable algorithms. Notably, the field of Natural Language Processing (NLP) has made remarkable progress by scaling architecture, data, and compute resources. In this work, we aimed to create a similarly simple, efficient, and scalable approach for the field of Computer Vision, inspired by the successful scaling strategies applied to auto-regressive language models. We proposed XTRA, an auto-regressive image model that employs a novel Block Causal Masking technique. This model predicts the next block of pixels based on prior context, demonstrating significant improvements in sample and parameter efficiency compared to existing auto-regressive image models. Specifically, our empirical results indicate that XTRA achieves 152× greater sample efficiency and 7–16× improved parameter efficiency. We hope our work inspires further exploration in scalable auto-regressive models for Computer Vision.

References

[1] Elad Amrani, Leonid Karlinsky, and Alex Bronstein. Self-supervised classification network. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022. 2

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 2, 3, 7

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 3, 7

[4] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR, 2023. 2, 3

[5] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 1

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

[7] Amir Bar, Florian Bordes, Assaf Shocher, Mido Assran, Pascal Vincent, Nicolas Ballas, Trevor Darrell, Amir Globerson, and Yann LeCun. Stochastic positional embeddings improve masked image modeling. In *International conference on machine learning*. PMLR, 2024. 7

[8] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 1

[9] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 1

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 2

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020. 2

[12] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 6, 7

[13] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 2, 3, 6

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning, ICML*, 2020. 2

[15] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems*, 2020. 2

[16] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[17] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. In *arXiv preprint arXiv:2003.04297*, 2020. 2

[18] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 7

[19] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024. 2, 3, 7

[20] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 1

[21] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 1

[23] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014. 2

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3, 4

[25] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan

Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[26] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *International conference on machine learning*. PMLR, 2024. 1, 2, 3, 4, 5, 6

[27] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. 2

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 4, 6, 7

[30] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12 (7):2217–2226, 2019. 1

[31] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2

[32] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2. Minneapolis, Minnesota, 2019. 2, 3

[33] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. 1

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[36] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 1

[37] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 1

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 3

[39] Jason Tyler Rolfe. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016. 3

[40] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Spreading vectors for similarity search. *arXiv preprint arXiv:1806.03198*, 2018. 2

[41] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, et al. The effectiveness of mae pre-pretraining for billion-scale pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5484–5494, 2023. 2

[42] James Taylor, Berton Earnshaw, Ben Mabey, Mason Victors, and Jason Yosinski. Rxrx1: An image set for cellular morphological variation across many experimental batches. In *International Conference on Learning Representations (ICLR)*, page 23, 2019. 1

[43] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021. 2

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[45] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[46] Grant Van Horn, Oisin Mac Aodha, Yin Cui, Yang Song, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. iNaturalist 2018 competition dataset. `https://github.com/visipedia/inat_comp/tree/master/2018`, 2018. 2, 1

[47] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018. 1

[48] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2

[49] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022. 2, 3, 7

[50] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2, 3, 6, 7