The Collapse of Patches

Wei Guo, Shunqi Mao, Zhuonan Liang, Heng Wang, Weidong Cai
School of Computer Science, The University of Sydney
{wei.guo, shunqi.mao, zhuonan.liang, heng.wang, tom.cai}@sydney.edu.au

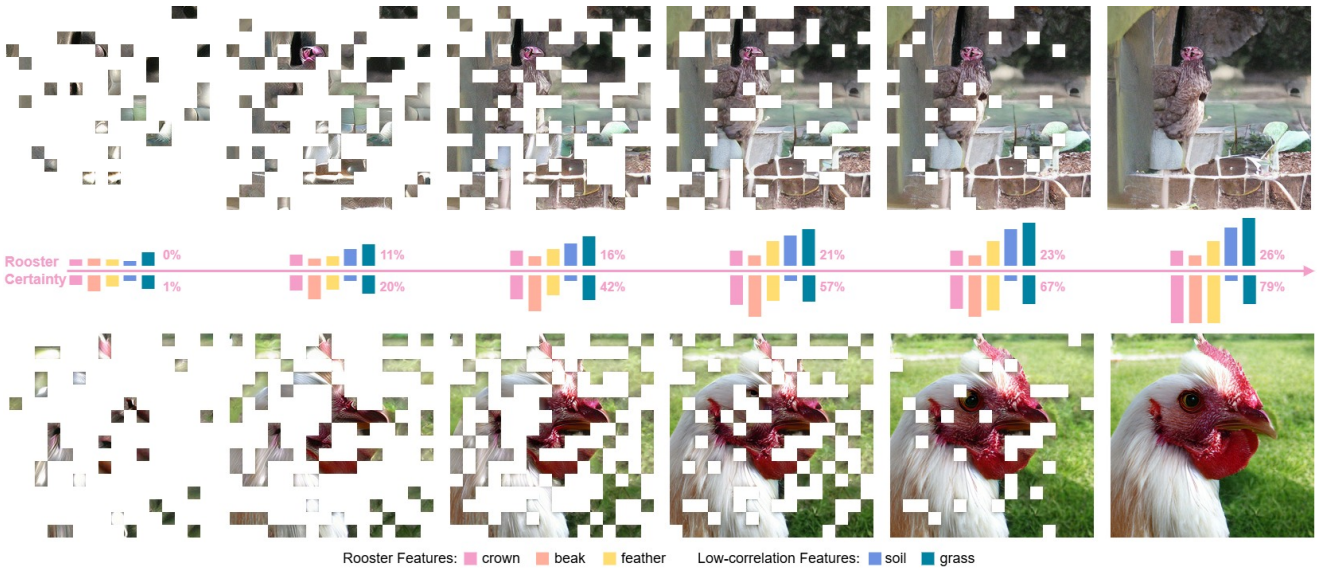

Figure 1. **Patch synthesis in random and collapse orders.** We autoregressively generate rooster image patches following random order (above) and collapse order (below). The latter synthesizes prominent rooster features and reduces image uncertainty more effectively.

Abstract

Observing certain patches in an image reduces the uncertainty of others. Their realization lowers the distribution entropy of each remaining patch feature, analogous to collapsing a particle's wave function in quantum mechanics. This phenomenon can intuitively be called patch collapse. To identify which patches are most relied on during a target region's collapse, we learn an autoencoder that softly selects a subset of patches to reconstruct each target patch. Graphing these learned dependencies for each patch's PageRank score reveals the optimal patch order to realize an image. We show that respecting this order benefits various masked image modeling methods. First, autoregressive image generation can be boosted by retraining the state-of-the-art model MAR. Next, we introduce a new setup for image classification by exposing Vision Transformers only to high-rank patches in the collapse order. Seeing 22% of such patches is sufficient to achieve high accuracy. With these experiments, we propose patch collapse as a novel image modeling perspective that promotes vision efficiency. Our project is available at https://github.com/wguo-ai/CoP.

1. Introduction

Images are more than collections of independent pixels or patches: they are structures of mutual dependence [42]. Observing parts of an image often reveals information about others. This observation (realization) process of an image, when modeled along a patch sequence, is then analogous to the **collapse** of a wave function in quantum mechanics [14]: once a particular patch is measured, the remaining unrealized patches' uncertainty reduces around the observed evidence. Intuitively, we refer to this image uncertainty reduction process as the **collapse of patches**.

Different patches collapse the uncertainty of the residual image with different effectiveness, which is a nontrivial phenomenon. Seeing the beak of a rooster first, for instance, constrains the leftover image information more than seeing a background field, as shown by the comparison of uncertainty reduction effects from an autoregressive (AR) image generator [27] following different patch synthesis orders in Fig. 1. In reality, a human painter also starts with a sketch of their subject's important parts in order to neatly constrain the underlying visual uncertainty [12]. Aside from image

synthesis, the human ability to correlate partial visual contents differently in a scene is also essential for completing vision tasks with high efficiency [32].

Most modern vision models, however, treat image patches as uniformly correlated samples in masked image modeling (MIM), *e.g.*, for stochastic AR generation [5, 11, 24, 27, 35, 45, 47, 48] or masked classification [6, 16, 19, 28, 51–53]. This assumption ignores the contributive differences among patches during collapse.

In this work, we formalize the problem of patch collapse respecting its patch-wise priorities. We assume that when certain patches of an image are observed, the feature distributions of the unobserved patches shift from broad, high-entropy shapes to concentrated, low-entropy states. The further assumption is that an image's patches can be ranked based on their elicited shift strengths during this process. To study how this collapse unfolds, we train a **Collapse Masked Autoencoder (CoMAE)** whose encoder selectively masks image patches with noise injection, conditioning the decoder in reconstructing a target patch. This encoder-predicted mask is a continuous vector weighting each patch's collapse contribution between 0 and 1. Our experiments validate the assumptions above: only a subset of patches are most responsible for each given patch's collapse, as polarized selection weights emerge to optimize reconstructions. Furthermore, we observe that this selection set varies across patches, suggesting that each patch has a distinct collapse dependency and contributes to the global image certainty with different effectiveness.

To analyze these patch dependencies at the image level, we map out a directed acyclic graph of patches with the CoMAE selection masks as edge weights. Applying PageRank [3] to this graph yields an ordering of patches by their collapse independence, defining an optimal uncertainty reduction sequence in which an image realizes itself. We term these sequences **collapse orders**. Our visualizations show that high-rank patches in the collapse order outlines important shapes in an image. Additionally, we observe that the inter-class collapse orders across ImageNet [7] samples exhibit moderate similarity, which suggests their depiction of a consistent underlying structure behind different images.

We demonstrate that collapse order offers beneficial supervision to MIM methods. When integrated into an AR image generator MAR [27], respecting the collapse order improves sample quality both quantitatively and qualitatively over the original model. Alongside generation, respecting the collapse order also leads to efficient image classification: by training on patches with high collapse priorities, a Vision Transformer (ViT) [10] can maintain high accuracy while processing just 22% of the image content. In contrast, conventional full-image classifiers sacrifice computation to redundant patches that contribute little discriminative value.

In summary, our work has the following contributions:

1. We introduce and formalize the problem of patch collapse, which offers a novel perspective to describe image structures that are fundamental to vision modeling.
2. We present CoMAE, an effective method to pinpoint the image-level order of patch collapse.
3. By supervising with the collapse order, we improve MIM methods' performance in AR image generation and masked image classification. These experiments show the generalizable applicability of patch collapse modeling to different vision tasks.

2. Related Works

Stochastic Masked Image Modeling. Inspired by masked language modeling [8], masked image modeling (MIM) predicts corrupted local units from an image to learn generalizable visual features. Stochastic MIM (SMIM) methods reconstruct randomly masked image portions to obtain self-supervised visual features, assuming uniform patch correlations. The pioneering denoising autoencoder [49] explores robust feature learning through partial latent pattern corruption. Later, Context Encoder [33] employs convolutional neural networks to inpaint stochastic image regions for representation learning. Masked Autoencoders (MAE) [16] apply high random masking ratios (e.g., 75%) in an asymmetric transformer for scalable learning. SimMIM [51] simplifies this process with larger patches and RGB regression. Painter [50] expands MIM to various image-to-image mapping tasks. MixMAE [28] incorporates image mixing alongside MIM for data augmentation. VideoMAE [46] extends MIM to consider spatio-temporal masking for videos, while OmniMAE [15] explores masked modeling with multimodal data. Recently, CAPI [6] enhances SMIM features by predicting missing patches w.r.t. an unmasked teacher. MIMIR [52] improves SMIM's adversarial robustness via mutual information-based reconstruction. These SMIM methods learn generalizable visual features but ignore the variance of inter-patch dependencies, leading to inefficient representations. In contrast, our CoMAE adaptively masks trivial patches to model patch dependencies accurately, deriving stronger guidance for MIM methods.

Adaptive Masked Image Modeling. Adaptive MIM (AMIM) methods adjust image-wise masking during training to effectively target informative regions. CMAE [19] unifies contrastive learning with MIM for stronger representation disambiguation. SiamMAE [53] leverages asymmetric masking for video correspondences. AttMask [21] generates attention-based masks from a teacher model for AMIM. AdaMAE [2] masks visible video tokens adaptively with an additional sampling network. SemMAE [25] uses semantic masks from a ViT to mask an image's dominant shapes. CL-MAE [29] employs curriculum learning to

adapt MIM to harder masks that hinder reconstruction. The recent RAM++ [56] uses adaptive masks for blind image restoration, while Self-Guided MAE [41] introduces self-guided informed masking based on early-stage patch clustering in MIM. Although AMIM methods efficiently adapt to data salience, they do not explicitly model image uncertainty reduction across patches. Our CoMAE formalizes patch collapse, identifying the global patch orders for image realization. CoMAE provides a unique perspective of image structures absent in prior methods, implemented with AMIM but not limited to its coverage in vision modeling.

Autoregressive Image Generation. Autoregressive (AR) models generate images sequentially with localized representations, often employing masking for unified learning. Classic AR methods, *i.e.*, PixelRNN [48] and PixelCNN [47], follow fixed raster orders during generation. VQ-GAN [11] generates quantized image tokens in the same order. MaskGIT [5] predicts quantized tokens with a scheduled stochastic order. MAGE [26] learns discrete image tokens and their generation in a unified manner. MAR [27] follows the same order but replace MaskGIT's discrete tokens with continuous latents. VAR [45] employs next-scale prediction for AR. MAGVIT [55] extends AR generation to videos. Recently, HMAR [24] improves AR efficiency with hierarchical multi-step prediction. xAR [35] generalizes next-token generation to flexible units such as cell or scales (next-x predictio). While different random or fixed orders are explored by these methods, our collapse order offers a data-salient strategy in modeling the generation priorities of AR units, yielding significant gains without drastic remodeling and retraining. Our strategy can also be integrated with various next-x methods straightforwardly.

Efficient Vision Transformers via Token Pruning. Token pruning methods enhance the computational efficiency of Vision Transformers (ViTs) by selectively removing redundant tokens. DynamicViT [34] estimates token importance scores to hierarchically prune trivial tokens. Similarly, ATS [13] and A-ViT [54] prune tokens with attention scores. AdaViT [30] learns a decision policy model to drop various inference units. EViT [1] fuse attention-pruned tokens into a single representation to retain more information. SPViT [23] employs an attention-based selector for this fusion. For semantic segmentation, DToP [44] dynamically prunes easy tokens via early exiting based on confidence thresholds. While these methods investigates feature pruning across the model architecture, our collapse order directly operates on explicit image patches. We show that vision efficiency can be significantly boosted by pruning on the model-agnostic image space alone. This decoupling of data and model also suggests that our method can readily combine with token pruning techniques.

3. Method

For modeling efficiency, we first pass an image through a variational autoencoder (VAE) [22, 36, 38] to represent N patches with a set of embeddings as $\{\mathbf{e}_n\}^N$. Each patch's feature distribution is conditioned on other patches, which can be expressed by a probability distribution $P\left(\mathbf{e}_n | \{\mathbf{e}_i\}_{i\neq n}^N\right)$. Our problem can then be formulated as finding a ranking function $R(\mathbf{e}_n)$ such that:

$$H_c = \sum_{i=1}^{N} H\left[P\left(\mathbf{e}_n | \{\mathbf{e}_i; R(\mathbf{e}_i) > R(\mathbf{e}_n)\}_{i\neq n}^N\right)\right], \quad (1)$$

where H measures the distribution entropy, is minimized.

To model R, we first learn each patch's dependencies on other patches with an autoencoder's encoded masking weights in Sec. 3.1 through reconstruction. A directed acyclic graph of patches can then be drawn in Sec. 3.2, where the edge weights are patch dependencies. Computing the PageRank [3] scores of this graph yields the output of R, which constitutes our collapse order as the optimal patch realization sequence to reduce image uncertainty.

With the collapse order identified, we propose to improve the existing stochastic AR image generator MAR [27] by respecting this image structure in Sec. 3.3. To accomplish this, we train an MAR model with collapse order guidance. Finally, we investigate our collapse modeling's effectiveness on image classification by training a Vision Transformer (ViT) [10] with collapse-order masks in Sec. 3.4.

3.1. Collapse Masked Autoencoder

Following the motivation that some patches are more responsible for a specific patch's collapse than others, we assume masking K patches from $\{\mathbf{e}_i\}_{i\neq n}^N$ leaves P's shape approximately unchanged. To find these influential patches, we train a Collapse Masked Autoencoder (CoMAE) model as shown in Fig. 2 (a). During reconstruction, the encoder f follows ViT [10] to pool visual information from $\{\mathbf{e}_i\}_{i\neq n}^N$ with self-attention blocks. It predicts a soft weight vector $\mathbf{w} \in [0,1]^n$ for patch selection as:

$$\mathbf{w}_n = f\left(\{\mathbf{e}_i\}_{i\neq n}^N; \mathbf{q}_n\right), \quad (2)$$

where $\mathbf{q}_n$ is a learned positional embedding to inform the encoder of the under-reconstruction patch location. We then mask each patch embedding by noise injection w.r.t. $\mathbf{w}_n$ as:

$$\mathbf{e}_i^m = \alpha_i \mathbf{e}_i + (1-\alpha_i)\mathcal{N}(0, \mathbf{I}), \; i \neq n, \quad (3)$$

where α is an exponential decay interpolant with a hyperparameter σ for controlled steepness:

$$\alpha_i = \exp\left(-\frac{\left(1-\mathbf{w}_n^i\right)^2}{2\sigma^2}\right). \quad (4)$$

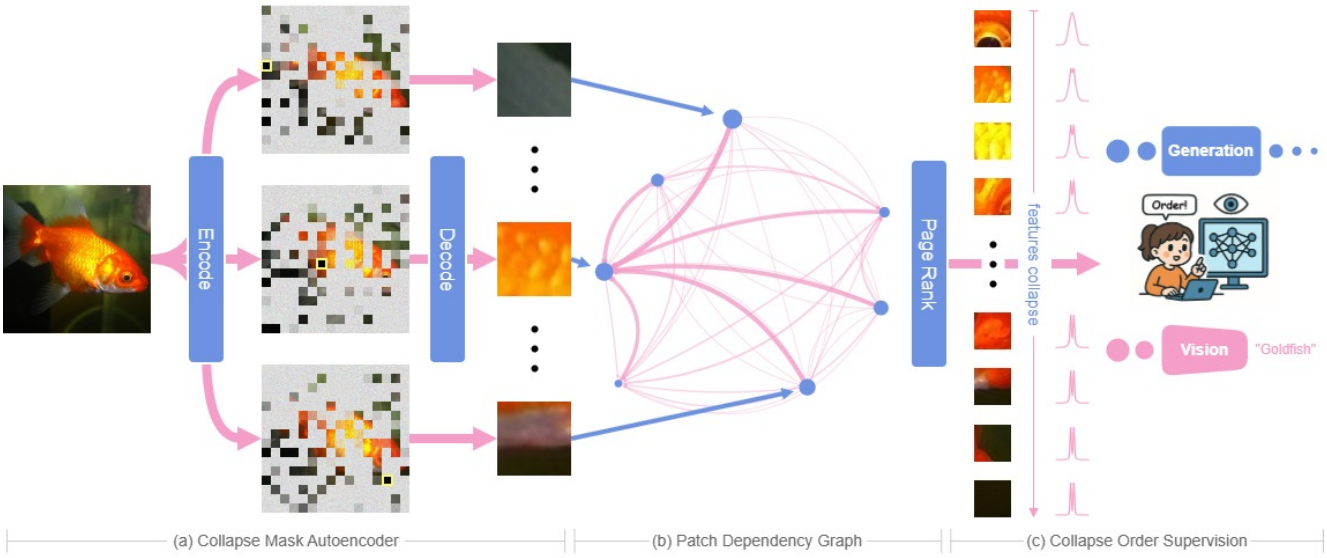

Figure 2. **Pipeline overview.** Given an image, the CoMAE encoder selects the most influential patches needed to reconstruct each patch, while trivial patches are masked with heavier noise injection. These selection weights form a patch dependency graph on which we compute the PageRank scores to determine the collapse order of patches, where higher-rank patches are less dependent on the rest of the image. Finally, we use this ranking to supervise image generation and classification tasks to follow the correct patch processing order.

The decoder g is another stack of ViT-like self-attention blocks that pools information from the selected patches to reconstruct the target patch $\mathbf{e}_n^*$:

$$\mathbf{e}_n^* = g\left(\{\mathbf{e}_i^m\}_{i\neq n}^N; \mathbf{q}_n\right). \tag{5}$$

The patch reconstruction loss is simply $\mathcal{L}_r = \|\mathbf{e}_n - \mathbf{e}_n^*\|_1$. We alternatively optimize f and g with $\mathcal{L}_r$ in a batch-wise manner during training. Please refer to Sec. 8 for additional architecture details of CoMAE.

Polarization. To test our assumptions in Sec. 1 that there exists a priority ranking of patches in patch collapse, we observe the training metrics of CoMAE in Sec. 4.1. It can be seen that $\mathbf{w}$ polarizes to 0 and 1 as the reconstruction loss $\mathcal{L}_r$ converges to minima instead of staying uniform. This effect confirms that different patches contribute to an image's uncertainty reduction with different effectiveness, since only a subset of patches significantly contribute to each target patch's feature collapse.

Contrastive Regularization. Do different patches rely on different subset selections for their feature collapse? To answer this further question, we inquisitively add a contrastive objective to encourage the diversity of $\mathbf{w}$ as:

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{\exp\left(\mathrm{sim}\left(\mathbf{w}_i, \mathbf{w}_i\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\mathrm{sim}\left(\mathbf{w}_i, \mathbf{w}_j\right)/\tau\right)}, \tag{6}$$

where τ is a learnable temperature and $\mathrm{sim}\left(\cdot,\cdot\right)$ measures the cosine similarity between two vectors. We then define the total loss to be $\mathcal{L} = \mathcal{L}_r + 0.01\mathcal{L}_c$ and retrain CoMAE.

This ablation, detailed in Sec. 4.1, shows that CoMAE plateaus at a significantly higher $\mathcal{L}_r$ without $\mathcal{L}_c$, assigning similar $\mathbf{w}$ to all patches. With $\mathcal{L}_c$, $\mathcal{L}_r$ is significantly reduced by diverse masks across patches and escapes local minima. Thus, it's reasonable to assume that the one-to-many dependency of each patch during collapse is diverse.

3.2. Patch Ranking

N instances of $\mathbf{w}$ can be encoded for all patches in an image. Together they form a patch dependency graph with a $N \times N$ adjacency matrix $\mathbf{A}$ where $\mathbf{A}_{ij} = \mathbf{w}_{ij}$. To rank the independency of patches, we compute their PageRank scores from $\mathbf{A}$. A patch with high PageRank score has more influence on other patches. Please see Sec. 6 for a formal proof linking this ranking mechanism to the objective optimal dependency ranking R.

3.3. Collapsing Autoregressive Image Generation

As shown in Fig. 3 (a), autoregressive (AR) image generators often follow random orders to generate patches sequentially. This stochastic arrangement assumes that patch dependencies follow a uniform distribution in an image, inducing inefficiency in image uncertainty reduction. We apply our patch sequencing learned from CoMAE as extra supervision to retrain a stochastic AR model, MAR [27], which we call **Collapsed Mask Autoregressive Model**

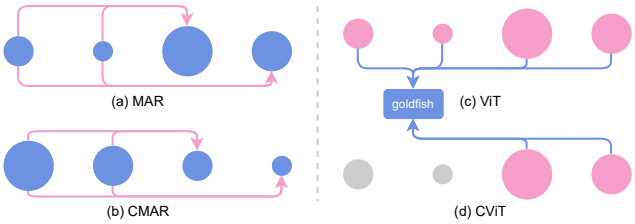

Figure 3. **Comparison of generators and classifiers.** Our generator (CMAR) and classifier (CViT) respect the collapse order.

Contrastive	Mask Entropy	Reconstruction Loss
✗	4.816	8.392
✓	**4.267**	**1.567**

Table 1. **Ablation of CoMAE's contrastive regularization.** Contrastive guidance aids reconstruction and mask polarization.

(CMAR) in Fig. 3 (b). CMAR learns to generate patches following the collapse order instead of stochastic orders.

For each training sample, we first pass it through the CoMAE encoder and obtain its patch ranks. We then mask a random amount of low-rank patches and learn to generate them from high-rank ones. Since this sampling order of patches is deterministic, CMAR is more likely to overfit than the original model. To compensate for this effect, we replace 10% sampling sequences with random ranks as a form of regularization. We also keep the random image flip data augmentation from MAR.

3.4. Collapsing Image Classification

Conventionally, an image classifier has access to the entire input image as in Fig. 3 (c). Since we show earlier that the realization of an image follows collapse order, it's intuitive to ask if such classifiers can maintain accurate when only high-rank (*i.e.* highly independent) patches in the collapse order are present.

Correspondingly, we train a ViT classifier on ImageNet [7] under two settings: full-image and masked. We mask $0 \sim 99\%$ low-rank patches with a cosine annealing schedule favoring lower mask rates, which results in a **Collapsed Vision Transformer (CViT)** depicted in Fig. 3 (d). Instead of replacing the masked patches with mask tokens, CViT directly drops them from the input sequence for efficiency.

4. Experiments and Results

We conduct all our experiments on center-cropped 256×256 images from ImageNet-1k [7]. These images are first processed by a KL-16 VAE [37] used by MAR [27] into 256 16-dim tokens. We use a RTX5090 GPU for all trainings.

Implementation of CoMAE. Both the encoder and decoder of CoMAE have 12 attention blocks following ViT [10], with embedding dimensions 64 and 256 respectively. A four-layer residual MLP is appended to the decoder to output the final 16-dim target patch token. The encoder and decoder are alternatively optimized in a batch-wise manner. We train CoMAE for 512 epochs with a cosine annealing schedule decaying learning rate from 1e-4 to 0.

Implementation of CMAR. Training MAR from scratch is computationally daunting for our resources. Instead, we treat the pretrained MAR as an order-agnostic prior and fine-tune it for 24,000 steps (batch size 32) on the same ImageNet data with our collapse order. Due to restricted computational resources, we only experiment on the MAR-B variant. We linearly warmp up the learning rate to 1e-7 during the first 10% steps and then decay it to 0 with cosine annealing. The model weights are updated with a per-step estimated mean average (EMA) of rate 0.99999. During inference, we set CMAR's classifier-free guidance (CFG) [18] scale to 3.0 from our ablations in Tab. 3. We keep the original MAR at CFG 2.9 for its optimal performance.

Implementation of CViT. We fine-tune the ImageNet-21k pretrained ViT-Base [10] model on ImageNet-1k for classification of 1000 image classes. The training expands 3 epochs with a cosine annealing schedule decaying learning rate from 1e-4 to 0. The model weights are updated by per-step EMA with rate 0.9999. We apply a random horizontal flip of training images for data augmentation.

4.1. Properties of CoMAE

We provide experiments to corroborate our statements of CoMAE's behaviors. First, we show that CoMAE emerges polarized selections instead of uniformly distributing the weights in during reconstruction optimization. To quantify polarization, we define a Mask Entropy metric as:

$$H_{\text{mask}} = -\frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{N} \mathbf{w}_j^i \log \mathbf{w}_j^i, \tag{7}$$

where M is the number of samples and N is the number of patches. A lower H_{mask} reflects more polarized mask distribution in $\mathbf{w}$. As shown in Fig. 6, H_{mask} grows smaller as reconstruction loss optimizes, indicating that only a subset of patches is responsible for a target patch's collapse.

Next, we observe that different target patches depend on different patch subsets for their collapse. Our contrastive ablations in Tab. 1 reveal that this diverse selection further polarizes masks and significantly optimizes reconstruction. Additionally, we provide a visualization of the collapse orders illustrated in Fig. 4. The patches outlining major objects have higher ranks in each image, which aligns intu-

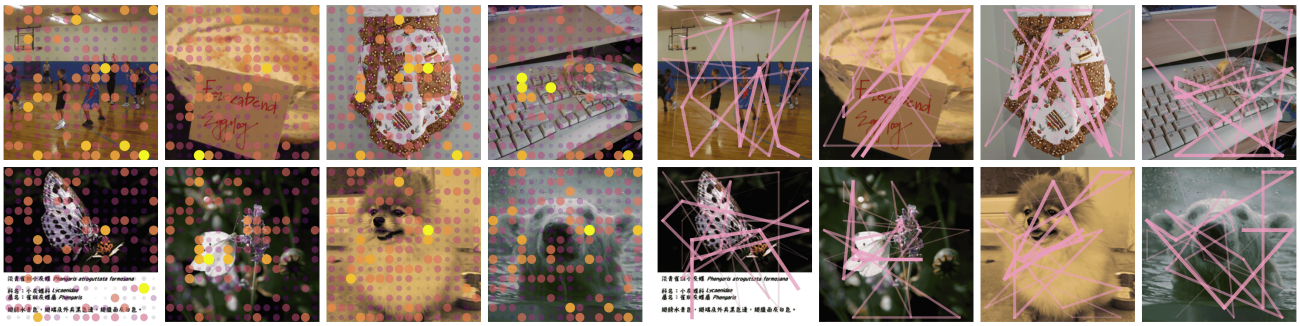

Figure 4. **Visualization of collapse order.** The left figure shows image patches with different collapse ranks indicated by circle sizes. The right figure connects the top-ranked 64 patches by collapse order. One can observe that top patches outline important shapes in each image.

Method	w/o CFG					w/ CFG				
	FID↓	tFID↓	IS↑	Pre.↑	Rec.↑	FID↓	tFID↓	IS↑	Pre.↑	Rec.↑
MAR	**7.114**	**3.498**	**194.50**	0.784	0.571	5.997	2.330	281.48	0.822	0.571
MAR+C (Ours)	7.173	3.563	193.80	**0.788**	0.568	5.956	2.321	**284.78**	**0.826**	0.566
CMAR (Ours)	7.213	3.600	190.92	0.781	**0.572**	**5.928**	**2.238**	280.55	0.818	**0.576**

Table 2. **Generation performance of different AR methods.** The first place is **bolded**. MAR+C denotes the unfine-tuned MAR results following our collapse order. CMAR tests MAR fine-tuned with collapse order.

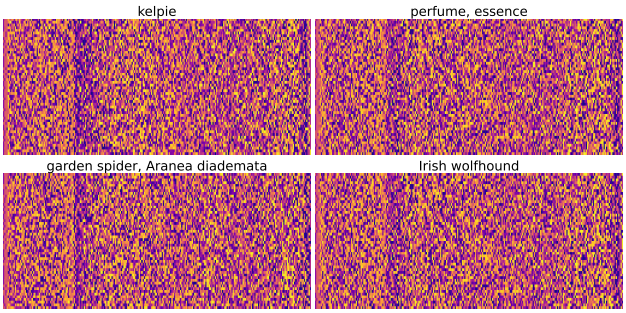

Figure 5. **Class-wise collapse order patterns.** These heatmaps show sample patch indices sorted in collapse order for each class.

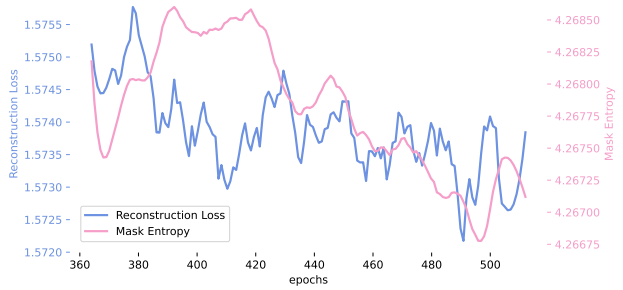

Figure 6. **Training of CoMAE (last 140 epochs).** Mask entropy drops together with reconstruction loss.

itively with human's visual scanning order [20]. This emergent similarity between image synthesis and recognition is particularly intriguing, as it might suggest a convergence of optimal scanning order between these opposite tasks.

Finally, we visualize class-wise collapse order patterns in Fig. 5. One can see that similar collapse orders emerge among instances from the same class, since there are multiple closely aligned patch indices indicated by the vertical heat lines. The inter-class collapse order patterns are also similar judged by the locations of their heat lines. Therefore, it's reasonable to assume that the identified collapse order exhibit moderate consistency over classes and image instances, which suggests the existence of a common structure behind different images' realization.

4.2. Benchmarking CMAR

We compare our collapse order's effect by quantitative metrics in Tab. 2. MAR denotes the original model with random-order generation, and MAR+C is the same model inferred with collapse order. We measure generation performance by Frèchet Inception Distance (tFID) [17] and Inception Score (IS) [39] following the original work [27] on 50,000 augmented training samples. Additionally, we measure the original FID, precision, and recall against the 10,000 samples in standard reference batch [9].

Our CMAR achieves a significant 4% gain in tFID despite of a very minor degradation in IS (0.3%). The fine-tuning-less MAR+C achieves the highest IS score and has a tFID behind CMAR. Our two methods slightly degrades

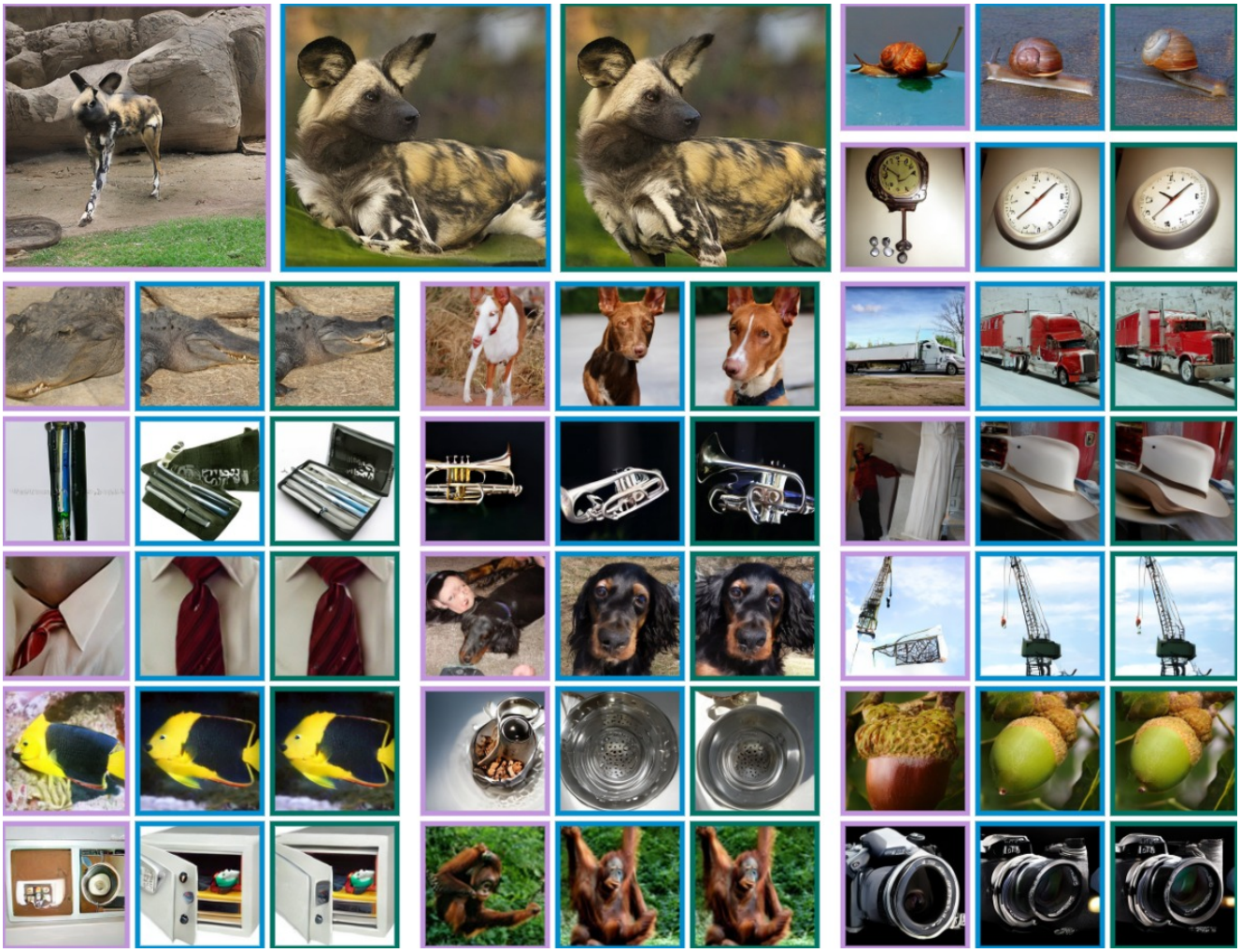

Figure 7. **Qualitative comparison of ARs.** MAR is framed in purple. Our MAR+C and CMAR results are in blue and green respectively.

CFG	FID↓	tFID↓	IS↑	Pre.↑	Rec.↑
2.9	**5.913**	**2.219**	276.375	0.816	**0.582**
3.0	5.928	2.238	280.55	0.818	0.576
3.1	5.933	2.240	**284.50**	**0.819**	0.572

Table 3. **Ablation of CMAR's CFG.** First place is **bolded**.

Order	FID↓	tFID↓	IS↑	Pre.↑	Rec.↑
Ascend	6.005	2.267	269.27	0.782	0.556
Descend	**5.928**	**2.238**	**280.55**	**0.818**	**0.576**

Table 4. **Ablation of CMAR's synthesis order.** CMAR trained with descending collapse ranks has the best performance.

in metrics during generation without CFG, possibly due to their higher reliance on the conditioning from class-specific labels and orders which requires a stronger conditioning guidance provided by higher CFG scales.

The qualitative samples of MAR+C and CMAR are shown in Fig. 7. It can be observed that CMAR synthesizes slightly more realistic images than MAR+C, while there are significant artifacts in the baseline MAR such as object compositional defects or content confusion.

4.3. Properties of CMAR

We conduct an ablation of CFG scales for CMAR in Tab. 3. A lower CFG achieves lower FIDs at the expense of IS, while higher CFG degrades in FIDs with an increase of IS. Therefore, we choose 3.0 as our optimal CFG scale to balance performance, different from the original setting at 2.9. Intuitively, this shift in CFG scale can be explained by CMAR's stronger reliance on conditioning labels in order to address diverse collapse orders in different classes.

Should CMAR follow ascending or descending patch

Method	T1@0%	T5@0%	T1@30%	T1@30%	T1@50%	T5@50%	T1@78%	T5@78%	AuC
ViT	**82.91**	**96.28**	80.03	94.80	74.38	91.48	22.38	36.76	57.16
DynamicViT [34]	81.74	95.64	**81.44**	**95.46**	**77.54**	**93.30**	20.66	37.09	56.32
ViT+C (Ours)	82.84	96.23	78.09	93.82	71.67	89.58	**31.04**	**49.45**	**57.27**
RViT	83.10	96.46	81.33	95.60	78.94	94.50	67.27	87.23	70.86
CViT(Ours)	**83.11**	**96.50**	**81.37**	**95.69**	**79.39**	**94.63**	**70.57**	**88.94**	**72.19**

Table 5. **Classification accuracy of ViT under different settings.** First place is **bolded**. CViT achieves superior classification performance throughout mask rates. ViT+C degrades slightly without masking but outperforms ViT at 78% masking.

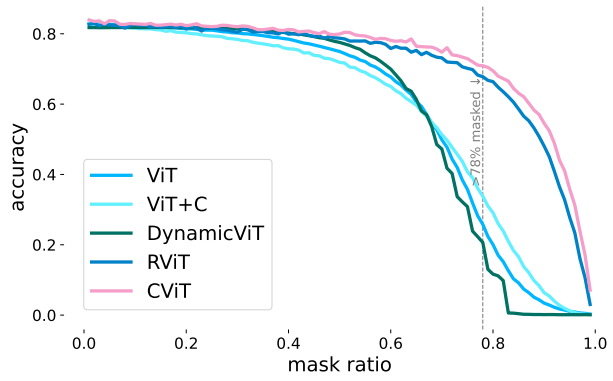

Figure 8. **ViT accuracy curves.** Our CViT outperforms baselines consistently along different mask ratios.

ranks for generation? We train a separate model in each direction in Tab. 4 for ablation. The ascending model performs significantly worse than the descending one. This observation is consistent with the intuition that patches most relied upon by others should be generated first.

4.4. Benchmarking CViT

We compare CViT with four baselines under different patch mask ratios in Tab. 5. The ViT model sees full images during training and is randomly masked during inference. ViT+C is a ViT variant that infers from collapse masks instead of random masking without modifying ViT's training. RViT is trained in the same way as CViT but with random masks. Although DynamicViT [34] is a token-pruning method that operates on model instead of data space, we include it to compare the effects of these separate pruning perspectives. Specifically, we employ the DeiT-B variant of DynamicViT which has similar parameter sizes as our ViT. We test these models' performance with top-k classification accuracy and an Area under Curve (AuC) metric that integrates over top-1 accuracies along the $0 \sim 99\%$ mask rates.

CViT leads in almost all metrics, demonstrating our collapse order's efficiency in capturing salient visual information from sparse high-rank patches. This effect is also observed in the training-less ViT+C, whose accuracy is higher than ViT and DynamicViT when extremely sparse 78% patches are masked. Our AuC superiority under both scenarios suggests that the overall classification performance benefits from patch collapse modeling.

4.5. Properties of CViT

Our CViT maintains high classification accuracy with a significant portion of low-collapse-rank patches masked, as visualized in Fig. 8. To analyze the accuracy decay, we find the knee of this concave accuracy curve with the Kneedle algorithm [40]. CViT's performance maintains until 78% patches are omitted, at which point its accuracy drops to 70.6%. As we apply masking by dropping out patches from the input sequence instead of replacing them with mask tokens, the computational cost is reduced by 95.16% from the $O\left(n^2\right)$ complexity of attention process. Furthermore, CViT also achieves the highest accuracy without masking. This result suggests that our patch collapse modeling can also benefit full-image classification.

It's also worth noticing that RViT, although following random order during training, also experiences a performance gain at lower mask ratios. However, its accuracy is lower than CViT at each level, suggesting our collapse order's superiority over stochastic modeling.

5. Conclusion

In this work, we introduce a novel modeling of the local feature uncertainty reduction process in images as patch collapse. By training a Collapse Masked Autoencoder to reconstruct a target patch relying on other tiles, and analyzing its resultant patch dependency graph with PageRank, we are able to identify an optimal ordering of patches during image realization that maximally reduces uncertainty, *i.e.*, the collapse order of patches. Experiments show that respecting this order benefits masked image modeling methods in: (1) autoregressive image generation, where the state-of-the-art model MAR is boosted in FID and IS, and (2) image classification, which leads to a ViT that can maintain high accuracy despite seeing only 22% image patches. We hope our patch collapse modeling will encourage a new perspective on salient visual structures, benefiting future exploration on efficient and scalable vision methods.

Supplementary Material

6. Proof of Optimal Collapse Ranking

We give a formal proof below to show that computing PageRank scores of the adjacency matrix $\mathbf{A}$ in Section 3.2 from CoMAE gives the optimal ranking R in Eq. (1) that minimizes image uncertainty (cumulative patch entropy).

Definition 1 (Cumulative conditional entropy)**.** *Define the cumulative conditional entropy of an ordered prefix S by:*

$$H_c(S) := \sum_{n=1}^{N} H\big(e_n \mid \{e_j : j \in S\}\big), \tag{8}$$

where H measures the distribution entropy.

Assumption 1 (Linear influence model)**.** *There exists $\beta \in \mathbb{R}^N_{\geq 0}$ such that the expected marginal reduction in H_c by observing patch j is approximately given by:*

$$\Delta(j \mid S) \approx \big((I - cP)^{-1}\beta\big)_j. \tag{9}$$

Assumption 2 (Submodularity)**.** *The entropy H_c is monotone and approximately submodular in S.*

Theorem 1 (Optimal collapse ranking)**.** *Let $\mathbf{A} \in [0,1]^{N\times N}$ be the learned dependency matrix with $\mathbf{A}_{ij}$ indicating the influence of patch j on patch i. Let P be the corresponding column-stochastic matrix, and let $c \in (0,1)$. Ordering patches in descending order of:*

$$r := (I - cP)^{-1}\beta, \tag{10}$$

minimizes the linearized proxy of H_c at each prefix. If β is constant or interpreted as a personalized teleport vector, this ordering corresponds to PageRank on P.

Proof. By the Neumann series expansion of Eq. (10):

$$r = (I - cP)^{-1}\beta = \sum_{t=0}^{\infty} c^t P^t \beta, \tag{11}$$

we derive the following recurrence for r:

$$cPr = cP\sum_{t=0}^{\infty} c^t P^t \beta = \sum_{t=0}^{\infty} c^{t+1} P^{t+1}\beta = \sum_{s=1}^{\infty} c^s P^s \beta, \tag{12}$$

where $s = t + 1$. Therefore,

$$\beta + cPr = \beta + \sum_{s=1}^{\infty} c^s P^s \beta = \sum_{s=0}^{\infty} c^s P^s \beta = r. \tag{13}$$

Thus, r satisfies the fixed-point equation:

$$r = \beta + cPr. \tag{14}$$

Under Assumption of Submodularity, greedy maximization of the marginal gain $\Delta(j \mid S)$ gives a $(1 - 1/e)$-approximation to the minimization of H_c [31]. Since the linear model assumes $\Delta(j \mid S) \approx r_j$, selecting nodes in descending order of r_j yields an optimal prefix sequence for the cumulative entropy minimization objective.

Finally, note that the fixed-point equation above matches the PageRank formulation:

$$r = (1 - c)\beta + cPr, \tag{15}$$

after normalizing β so that it sums to one. Hence, r is precisely the PageRank (or personalized PageRank) vector for P, completing the proof. □

7. Additional Qualitative Results

We provide more qualitative samples generated by our CMAR in Fig. 9. It can be observed that our method synthesizes high-fidelity images across a broad range of classes.

8. Model Architectures

The architectures of MAR [27] and Vision Transformer [10] are well-documented in their original papers. We elaborate on our novel CoMAE architecture in Fig. 10 below.

Given the patch token sequence, the CoMAE encoder concatenates a learned [cls] token in front of it. The target patch is dropped to avoid information leakage during reconstruction. This sequence is then transformed by Rotary Position Embedding (RoPE) [43] to inform the encoder of each patch's relative image location. Next, we pass the sequence through 12 self-attention blocks with an embedding dimension of 256, retaining the processed [cls] token as output. Finally, a selection weight vector is obtained by passing this [cls] token through a linear projection, followed by a $\tanh$ normalization to keep each entry between $[0, 1]$.

We take the original input patch token sequence (without encoder processing) and inject them with Gaussian noise following Eq. (3) in Section 3.1. Again, the target patch is dropped for reconstruction. To inform the decoder which patch is currently under reconstruction, we append a [tgt cls] token in front of the entire sequence. The decoder then processes this sequence with 12 self-attention blocks. The embedding dimension is set as 64. Finally, we retrieve the first [tgt] token and pass it through a feedforward head to obtain the reconstructed target patch.

9. Limitations and Future Improvements

As our collapse order describes local image units as patches, it's subject to the constraints of this representation: (1) each region has the same fixed size and (2) the shape of each

Figure 9. **Additional qualitative samples of CMAR.** Our method generates high-fidelity images across classes.

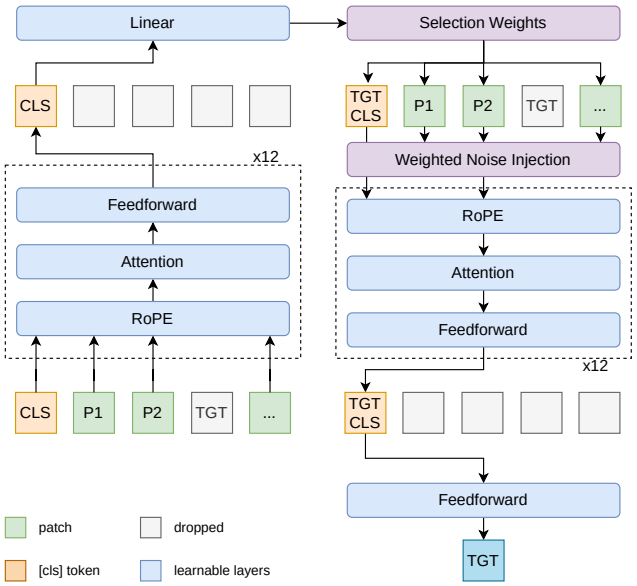

Figure 10. **Architecture of CoMAE.** Both the encoder and decoder follow ViT to pool information into a [cls] token.

patch doesn't reflect object saliency. In a more generalized setup, the image units can be expressed with salient local features such as segmentation maps or class activation maps. This extra layer of saliency fits more closely with our assumption of image locality during the collapse process. For now, we keep the modeling simple to study the preliminary feasibility of formalizing image collapse and leave these explorations to future works.

Additionally, we were unable to train more variants for CoMAE, CMAR, and CViT in this work due to limited computation. While we show that the current collapse order can already boost MIM methods in image generation and classification, scaling up training should achieve even higher performance gains. For instance, CoMAE can identify collapse orders more accurately with a deeper encoder, and larger MAR models can be converted into CMARs with longer training. More vision tasks can also be tested with collapse order for optimization, such as segmentation and detection. We will conduct these experiments if more computational resources become available in the future.

Finally, an alternative CoMAE design remains unexplored. Instead of training a decoder from scratch, we could directly employ MAR or similar autoregressive image generators for decoding. These pretrained decoders decouple encoder learning from the reconstruction objective, making it more efficient. Furthermore, CoMAE can also be deployed on the representation space of Representation Autoencoders (RAEs) [57] to identify the process of **representation collapse** instead of patch collapse. We deem these directions as having high potentials in improving CoMAE and will study them soon.

10. Ethical Statement

Our method studies the optimal scanning order of patches in image synthesis and classification. Since the identified high-rank patches in our collapse sequence have shown greater influence on these downstream tasks, visual patch attacks [4] could leverage them for more concentrated prompt engineering. However, the same information can also be adopted by visual models to enhance their robustness on high-rank patches to guard against such attacks. We will release our implementation code and model weights to aid the design of these defenses.

References

[1] Not all patches are what you need: Expediting vision transformers via token reorganizations. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *ICLR*, 2021. 3

[2] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. AdaMAE: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *CVPR*, pages 14507–14517, 2023. 2

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998. 2, 3

[4] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 3

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, pages 11315–11325, 2022. 2, 3

[6] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latent patches for improved masked image modeling. *arXiv preprint arXiv:2502.08769*, 2025. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 2, 5

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 2

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6

[10] 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 *ICLR*, 2021. 2, 3, 5, 1

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 2, 3

[12] Judith E Fan, Wilma A Bainbridge, Rebecca Chamberlain, and Jeffrey D Wammes. Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, 2(9):556–568, 2023. 1

[13] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *ECCV*, pages 396–414. Springer, 2022. 3

[14] Richard Phillips Feynman. Space-time approach to non-relativistic quantum mechanics. *Reviews of Modern Physics*, 20(2):367, 1948. 1

[15] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *CVPR*, pages 10406–10417, 2023. 2

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 6

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 5

[19] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *IEEE TPAMI*, 46(4):2506–2517, 2023. 2

[20] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 2002. 6

[21] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. In *ECCV*, pages 300–318, 2022. 2

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[23] Zhenglun Kong, Peiyan Dong, Xiaolong Ma, Xin Meng, Wei Niu, Mengshu Sun, Xuan Shen, Geng Yuan, Bin Ren, Hao Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *ECCV*, pages 620–640. Springer, 2022. 3

[24] Hermann Kumbong, Xian Liu, Tsung-Yi Lin, Ming-Yu Liu, Xihui Liu, Ziwei Liu, Daniel Y Fu, Christopher Re, and David W Romero. Hmar: Efficient hierarchical masked autoregressive image generation. In *CVPR*, pages 2535–2544, 2025. 2, 3

[25] Gang Li, Heliang Zheng, Daqing Liu, Chaoyue Wang, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *NeurIPS*, 35:14290–14302, 2022. 2

[26] Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *CVPR*, pages 2142–2152, 2023. 3

[27] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *NeurIPS*, pages 56424–56445, 2024. 1, 2, 3, 4, 5, 6

[28] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *CVPR*, pages 6252–6261, 2023. 2

[29] Neelu Madan, Nicolae-Cătălin Ristea, Kamal Nasrollahi, Thomas B Moeslund, and Radu Tudor Ionescu. Cl-mae: Curriculum-learned masked autoencoders. In *WACV*, pages 2492–2502, 2024. 2

[30] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *CVPR*, pages 12309–12318, 2022. 3

[31] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14 (1):265–294, 1978. 1

[32] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007. 2

[33] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 2

[34] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NeurIPS*, 34:13937–13949, 2021. 3, 8

[35] Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. In *ICCV*, pages 15781–15791, 2025. 2, 3

[36] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, pages 1278–1286. PMLR, 2014. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 5

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3

[39] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016. 6

[40] Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a" kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE, 2011. 8

[41] Jeongwoo Shin, Inseo Lee, Junho Lee, and Joonseok Lee. Self-guided masked autoencoder. *NeurIPS*, 37:58929–58954, 2024. 3

[42] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001. 1

[43] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1

[44] Quan Tang, Bowen Zhang, Jiajun Liu, Fagui Liu, and Yifan Liu. Dynamic token pruning in plain vision transformers for semantic segmentation. In *ICCV*, pages 777–786, 2023. 3

[45] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 37:84839–84865, 2024. 2, 3

[46] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, pages 10078–10093, 2022. 2

[47] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *NeurIPS*, 29, 2016. 2, 3

[48] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, pages 1747–1756. PMLR, 2016. 2, 3

[49] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103, 2008. 2

[50] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2

[51] 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 *CVPR*, pages 9653–9663, 2022. 2

[52] Xiaoyun Xu, Shujian Yu, Zhuoran Liu, and Stjepan Picek. MIMIR: Masked image modeling for mutual information-based adversarial robustness. *arXiv preprint arXiv:2312.04960*, 2023. 2

[53] Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at what you see: Masked image modeling without reconstruction. In *CVPR*, pages 22732–22741, 2023. 2

[54] Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, pages 10809–10818, 2022. 3

[55] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, pages 10459–10469, 2023. 3

[56] Zilong Zhang, Chujie Qin, Chunle Guo, Yong Zhang, Chao Xue, Ming-Ming Cheng, and Chongyi Li. Ram++: Robust representation learning via adaptive mask for all-in-one image restoration. *arXiv preprint arXiv:2509.12039*, 2025. 3

[57] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 3