Revisiting Autoregressive Models for Generative Image Classification

Ilia Sudakov[1,2] Artem Babenko[2] Dmitry Baranchuk[2]

[1] HSE University, Moscow, Russia
[2] Yandex Research, Moscow, Russia

Abstract. Class-conditional generative models have emerged as accurate and robust classifiers, with diffusion models demonstrating clear advantages over other visual generative paradigms, including autoregressive (AR) models. In this work, we revisit visual AR-based generative classifiers and identify an important limitation of prior approaches: their reliance on a fixed token order, which imposes a restrictive inductive bias for image understanding. We observe that single-order predictions rely more on partial discriminative cues, while averaging over multiple token orders provides a more comprehensive signal. Based on this insight, we leverage recent any-order AR models to estimate order-marginalized predictions, unlocking the high classification potential of AR models. Our approach consistently outperforms diffusion-based classifiers across diverse image classification benchmarks, while being up to 25× more efficient. Compared to state-of-the-art self-supervised discriminative models, our method delivers competitive classification performance – a notable achievement for generative classifiers. The code and models are available at: https://github.com/yandex-research/ar-classifier.

Keywords: Generative classification · Autoregressive models

1 Introduction

Generative models (GMs) have recently shown a remarkable ability to approximate highly complex visual data distributions [3, 43, 44, 55]. This progress raises the question of whether they can also succeed in well-studied discriminative setups. GMs have already demonstrated strong representation learning capabilities [2, 6, 26, 64], achieving comparable performance to state-of-the-art self-supervised discriminative approaches [7, 41] on multiple downstream benchmarks.

Another line of research investigates whether GMs can be used directly as *generative classifiers* (GCs). Likelihood-based GMs, such as diffusion models (DMs) and autoregressive (AR) models, can approximate class-conditional likelihoods $p(\mathbf{x}|\mathbf{y})$ and then derive posterior predictions $p(\mathbf{y}|\mathbf{x})$ via Bayes' rule. Interestingly, recent works [24, 32] reveal several intriguing properties of GCs. Unlike discriminative classifiers, which often rely on spurious correlations, GCs have been shown to avoid shortcut solutions [32]. Moreover, [24] found that GCs exhibit shape-biased predictions that align more closely with human perception, whereas standard discriminative models are typically texture-biased.

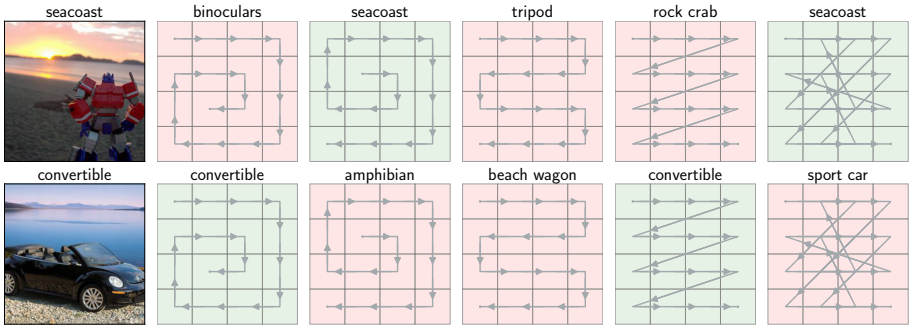

Fig. 1: Effect of different token orders on AR-based image classification. Each row presents an image and five different token orders used by the any-order AR model to predict its class. Token orders can largely affect the final classification outcome.

Motivated by the strong image generation quality of DMs, prior work [24,31–33] primarily focused on diffusion classifiers (DCs), leaving AR models relatively underexplored. However, recent advances in visual AR model design [34,35,38,43,54,55,63,71] have dramatically improved their generative performance, now achieving results competitive with DMs. These developments motivate a timely re-examination of visual AR-based GCs.

An important design choice affecting the performance of various iterative visual GMs is *how they generate images*. For example, classical next-token prediction AR models [54,68] typically generate images using a *raster-scan order*. In contrast, DMs generate images in a coarse-to-fine manner, which can be interpreted as a form of *spectral autoregression* [10,48,65]. From this perspective, DMs behave as implicit AR models that follow a low-to-high frequency order, better reflecting the hierarchical structure of natural images. We therefore hypothesize that this design choice is also critical for generative classification.

Contributions. In our work, we highlight the role of token ordering in AR-based generative classification. In Figure 1, we provide an illustrative example by classifying several images using an AR model evaluated under different token orders. We observe that some images are correctly classified under certain orders but misclassified under others, indicating that token orders may affect how the model interprets the image. This observation is further analyzed in Section 5.

To address this limitation, we investigate recent *any-order* AR (AO-AR) models that support generation under arbitrary token orders. Although a single AR prediction for a fixed token order can be less accurate than that of DMs due to the weaker visual inductive bias, any-order AR models can improve classification accuracy by marginalizing over multiple token orders. Intuitively, averaging predictions across different orders may capture contextual information from multiple image regions, leading to more discriminative predictions.

A practical advantage of AR models is that they can evaluate *order-conditional* log-likelihoods in a single forward pass, while DCs typically require $100-250$ model evaluations for a single likelihood estimate [31]. Therefore, even after

marginalizing over multiple token orders, AR models remain an order of magnitude more efficient than DMs.

Building on this insight, we propose an *order-marginalized* AR-based generative classifier that achieves both higher accuracy and improved efficiency compared to diffusion-based classifiers. The experiments show that our AR-based approach outperforms diffusion classifiers in both in-domain accuracy and robustness to distribution shifts, while offering up to 25× faster inference.

Furthermore, we compare GCs with state-of-the-art self-supervised discriminative models such as DINOv2 [41]. We find that our AR-based approach offers competitive classification performance on both in-domain and out-of-distribution benchmarks. To our knowledge, this is the first work that compares GCs with state-of-the-art SSL approaches which are typically stronger baselines than standard supervised classifiers.

2 Related work

Autoregressive models for image generation. Autoregressive (AR) models [12,35,43,54,69,70] were a promising paradigm for image generation prior to the wide adoption of diffusion models (DMs). They typically employ image tokenizers based on variants of VQ-VAE [3,12,58,68,72] to map images into discrete token sequences. Then, early AR models [12,30,40,57] performed *next-token prediction*, generating images token by token, typically in a fixed raster order (left-to-right, top-to-bottom), similarly to language models [46,56,66]. However, these models were later largely surpassed by DMs in generation quality, until more recent work introduced competitive next-token prediction [43,47,54,71,74] and parallel decoding [33,43,47,55,63] variants of AR models.

Importantly, most of these AR approaches fixate a specific token order. As we show later, this detail imposes an important limitation for generative classification with AR models. Therefore, in our work, we focus on one of the recent next-token prediction models, RandAR [43], which, unlike most AR counterparts, explicitly conditions on token positions and supports generation in arbitrary token orders. We will discuss this model in more detail in Section 3.

Generative classifiers. Among diverse generative paradigms [12,14,21,27,29,29,53], diffusion classifiers (DCs) [4,5,31,32] have recently received increased attention. DCs achieve competitive in-domain performance with discriminative models while offering improved robustness under distribution shifts [32]. However, an important limitation of DCs is their notoriously slow inference process that typically requires hundreds of model forward passes per image to achieve high accuracy. Joint Energy-based Models (JEMs) [15,16,67] address this problem by combining discriminative and generative objectives within a unified framework.

Meanwhile, classical AR models have proven effective as GCs in text classification tasks [25,32], though they were outperformed by DCs in visual domains [24]. Recent work [8] explores the potential of modern visual AR models by introducing VAR-based GCs with contrastive alignment fine-tuning. In our

work, we investigate GCs built upon state-of-the-art next-token prediction models and focus on purely generative approaches, without any additional tuning.

3 Preliminaries

Generative Classifiers. In general, GCs rely on estimating class-conditional likelihoods $p(\mathbf{x}|c)$ and apply Bayes' rule to obtain posterior probabilities $p(c|\mathbf{x}) \propto p(\mathbf{x}|c)p(c)$. Then, classification can be performed as:

$$c_i = \arg\max_{c_i \in C} \left[\log p(\mathbf{x}|c_i) + \log p(c_i)\right], \tag{1}$$

where the $\log p(c_i)$ term is often ignored in practice, assuming a uniform prior $p(c)$ over classes.

Pretrained DMs ϵ_θ approximate $\log p(\mathbf{x}|c)$ by estimating ELBO:

$$\log p(\mathbf{x}|c) \geq ELBO(\mathbf{x}, c) \approx \sum_{t,\epsilon} \lambda_t \|\epsilon_\theta(\mathbf{x}_t, t, c) - \epsilon\|^2, \tag{2}$$

where $\mathbf{x}_t$ is obtained using the forward diffusion process $q(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\mathbf{x}_t|\alpha_t\mathbf{x}, \sigma_t^2\mathbf{I})$. Previous works [31, 32] observed that using an objective with uniform weighting $\lambda_t = 1$ yields better classification performance in practice. Note that, Equation (2) requires $100-250$ model evaluations across different timesteps [31]. Since most state-of-the-art DMs are latent diffusion models [49], classification accuracy is computed in the continuous VAE latent space.

In contrast, AR models first encode images into discrete token sequences $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ of length N, typically using a VQ-VAE. Then, they provide explicit log-likelihood estimates in the latent space using a single forward pass:

$$\log p(\mathbf{x}|c) \approx \sum_{n=1}^{N} \log p_\theta(\mathbf{x}_n|\mathbf{x}_{<n}, c) \tag{3}$$

In our work, we highlight that both DMs and AR models implicitly impose an inductive bias through the way images are modeled, and that this choice can directly affect classification performance.

Any-order Autoregressive Modeling. In our work, we aim to explore the role of different AR token orders on image classification and therefore consider RandAR [43], a state-of-the-art AR model capable of generating images in arbitrary token orders. RandAR augments the image token sequence $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ with a corresponding set of *position instruction tokens* $P = \{\mathbf{p}_1, ...\mathbf{p}_N\}$. Each position token $\mathbf{p}_i$ is paired with its respective image token, forming an interleaved sequence $[\mathbf{p}_1, \mathbf{x}_1, ..., \mathbf{p}_N, \mathbf{x}_N]$.

To enable any-order generation, RandAR applies a random permutation π over token indices, producing the reordered sequence $[\mathbf{p}_1^{\pi(1)}, \mathbf{x}_1^{\pi(1)}, ..., \mathbf{p}_N^{\pi(N)}, \mathbf{x}_N^{\pi(N)}]$,

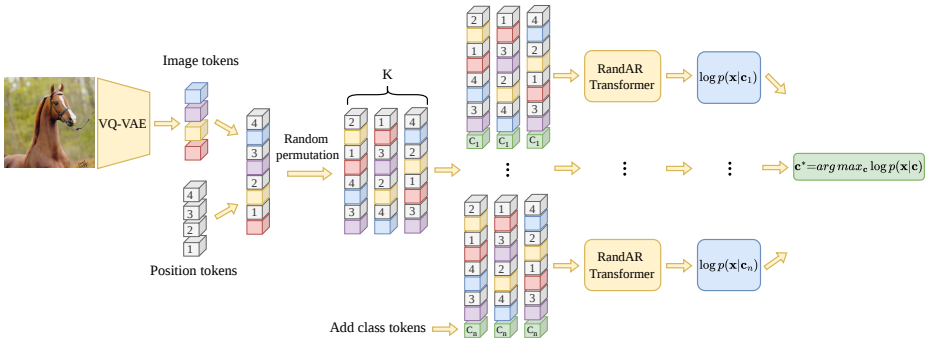

Fig. 2: Order-marginalized generative classification framework. i) Input image is tokenized with the VQ-VAE into a sequence of discrete tokens. ii) Position-instruction tokens are concatenated with the image tokens, and K randomly permuted sequences are constructed. iii) Class condition tokens c_i for each class are appended to every sequence. iv) The RandAR model predicts $\log p(\mathbf{x}|c_i)$ using Equation 5. v) The predicted class c^* is obtained as $\arg\max_{c_i} \log p(\mathbf{x}|c_i)$.

where $\pi(i)$ denotes the original position of the i-th token. Thus, RandAR *explicitly* defines the order-conditional likelihood $p(\mathbf{x}|\pi, c)$ as

$$p(\mathbf{x}|\pi, c) = \prod_{n=1}^{N} p(\mathbf{x}_n^{\pi(n)}|\mathbf{p}_1^{\pi(1)}, \mathbf{x}_1^{\pi(1)}, ..., \mathbf{x}_{n-1}^{\pi(n-1)}, \mathbf{p}_n^{\pi(n)}, c) \tag{4}$$

Note that RandAR can be readily reduced to a raster-order AR model by fixing the permutation to the identity mapping $\pi(i) = i$.

4 Order-marginalized generative classification

Since RandAR explicitly conditions on token positions and models multiple $p(\mathbf{x}|\pi, c)$, it allows estimating the order-unconditional likelihood $p(\mathbf{x}|c)$, by marginalizing over all possible permutations as $\mathbb{E}_\pi [p(\mathbf{x}|\pi, c)]$. Then, this expectation can be estimated with Monte Carlo sampling using K random order samples. However, in practice, we find this estimate of $p(\mathbf{x}|c)$ significantly less effective than estimating the lower bound on the order-unconditional log-likelihood, derived using Jensen's inequality:

$$\log p(\mathbf{x}|c) = \log \mathop{\mathbb{E}}_{\pi} [p(\mathbf{x}|\pi, c)] \geq \mathop{\mathbb{E}}_{\pi} [\log p(\mathbf{x}|\pi, c)] \approx \frac{1}{K} \sum_{k=1}^{K} \log p(\mathbf{x}|\pi_k, c) \tag{5}$$

For classification, we then append class tokens $C = \{\mathbf{c}_1, ..., \mathbf{c}_M\}$ to each permuted sequence, producing $[M, K]$ input sequences of length $2N+1$. Importantly, the same sampled orders are shared across all classes to ensure consistency.

For each class token $\mathbf{c}_i$, the model produces $[\log p(\mathbf{x}|\pi_1, \mathbf{c}_i), ..., \log p(\mathbf{x}|\pi_K, \mathbf{c}_i)]$, which are then aggregated into $\log p(\mathbf{x}|\mathbf{c}_i)$ using Equation 5. Finally, the class

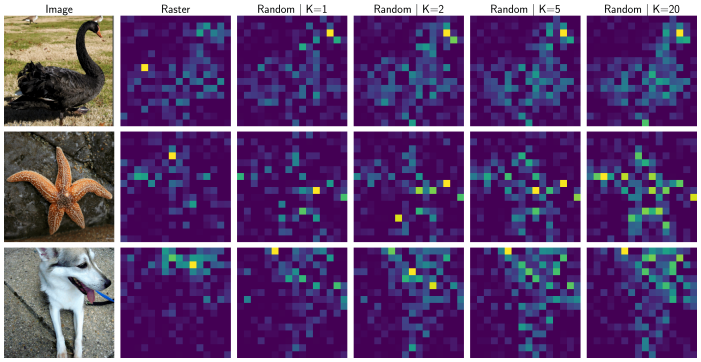

Fig. 3: Per-token "discriminative" log-likelihoods, computed as $clip[\log p(\mathbf{x}|c_{true}) - \log p(\mathbf{x}|c_{false}), 0]$, across different orders and K values. c_{true} denotes the correct class; c_{false} refers to a randomly selected incorrect one. Order-marginalized log-likelihood estimates ($K > 1$) capture class-specific attributes more accurately.

c^* is predicted according to Equation 1. The overall scheme of our classification procedure is presented in Figure 2.

Discussion. We hypothesize that the lower bound largely outperforms the direct likelihood estimate because RandAR was trained directly on Equation 5, making it better aligned with image understanding. Notably, classification with DMs also gets worse when evaluated using the true ELBO estimate, and instead show significantly better classification results when evaluated with their training objective Equation (2) with $\lambda_t = 1$ [31].

Another question is whether we can similarly marginalize DM predictions. According to Equation (2), DMs allow marginalization over multiple noise samples at each timestep. However, in Supplementary Materials, we show that additional averaging over noise samples provides no noticeable improvements and only increases computational cost. We hypothesize that this is because noise averaging is already included during timestep averaging.

Finally, AO-AR models are closely connected to *Masked Diffusion* (MD) models [22,42]. In Supplementary Materials, we show that the order-marginalized AO-AR objective in Equation (5) equals the MD evidence lower bound in expectation. This connection enables both AR- and MD-based evaluation of any-order AR-based GCs. In Section 6.7, we show that the AR formulation consistently outperforms the MD one.

5 Analysing order-marginalized AR classifiers

Setup. We pretrain two RandAR-L/16 variants under same training setups following [43]: a model using a fixed raster-order and another operating with random orders. The models are trained on ImageNet1K [50] using 256×256 images. We use the LlamaGen tokenizer [54], producing sequences of N=256 tokens.

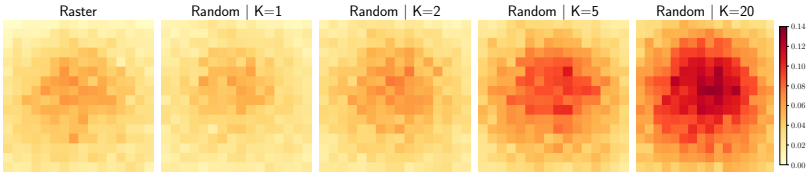

Fig. 4: Per-token accuracy of the RandAR classifier across different orders and K values. Center image tokens appear more discriminative, likely due to the center-object bias present in ImageNet. Increasing K consistently improves accuracy for all tokens.

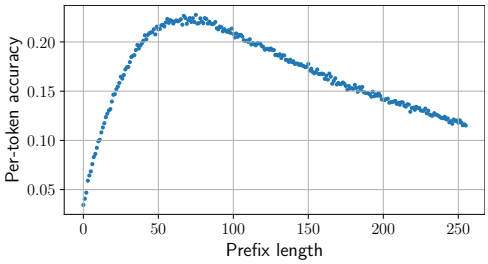

Fig. 5: Per-token accuracy of the RandAR classifier for $K = 20$ w.r.t. the prefix length.

Token order	IN-Val	IN-R	IN-S	IN-A
Raster	0.701	0.351	0.301	0.174
Random K=1	0.670	0.351	0.289	0.110
Random K=2	0.726	0.402	0.343	0.127
Random K=5	0.759	0.444	0.383	0.140
Random K=10	0.768	0.459	0.392	0.144
Random K=20	**0.769**	**0.469**	**0.409**	**0.146**

Table 1: Comparison of RandAR trained with random and raster orders in terms of top-1 classification accuracy on ImageNet1K.

Per-token likelihoods. First, we estimate per-token "discriminative" log-likelihoods $clip[\log p(\mathbf{x}|c_{true}) - \log p(\mathbf{x}|c_{false}), 0]$, where c_{true} denotes the correct class and c_{false} a random incorrect one. These log-likelihoods indicate which tokens are more discriminative for a particular class prediction.

Figure 3 shows the results for several images, with more examples in Appendix **??**. We observe that the raster-order model focuses more on specific class-object parts. Similarly, the random-order variant with K=1, a single-sample Monte Carlo estimate of Equation 5, also focuses on localized object parts. In contrast, for K=20, the object shapes appear more clearly, indicating that the model leverages broader class-level information for classification.

Per-token accuracy. Figure 4 presents per-token top-1 accuracy averaged over 1000 images, shown at their original token positions. The token accuracy consistently increases with K, resulting in more discriminative tokens. Central image tokens achieve higher accuracy, reflecting ImageNet's center-object bias.

Also, we explore how token accuracy varies with prefix length. Figure 5 reports per-token accuracy for K=20, averaged over 30,000 images. Token accuracy initially increases with prefix length, peaking ∼70 tokens, after which it declines. We hypothesize that the tokens with $[50 - 80]$ prefix lengths are more discriminative because such prefixes capture only high-level image information, forcing the model to generate class-defining details. In contrast, later tokens may contribute less, as they can rely more heavily on nearby observable tokens.

Overall performance. In Table 1, we compare the models in terms of their overall classification performance on the ImageNet validation set and several out-of-distribution (OOD) benchmarks.

We notice that the raster-order variant provides better results compared to the random-order one with K=1. We attribute this to the limited model capability to fit arbitrary orders similarly well as a single one. This intuition is supported by slightly worse generative performance of random-order models compared to the raster-order counterpart [43]. However, for K>1, the random-order model significantly outperforms the raster one across all validation sets.

Note that the higher accuracy of the random-order model for K>1 comes with additional computational cost, as each order requires an extra model forward pass. Nevertheless, as shown in Section 6.2, this overhead is fully justified compared to diffusion-based classifiers, which typically require hundreds of model forward passes [31, 32].

6 Experiments

In this section, we compare order-marginalized AR classifiers with state-of-the-art discriminative and generative approaches.

6.1 Main results

Baselines. We evaluate several baseline families known for their strong classification performance:

- **Discriminative approaches.** We compare with supervised pretrained ViT [11] models and the state-of-the-art self-supervised learning (SSL) method DINOv2 [41] with linear probing. To our knowledge, prior works have not compared against leading discriminative SSL methods, despite their largely superior performance over standard supervised classifiers.
- **Diffusion classifiers.** For DM-based baselines, we consider DiT [44] and SiT [39]. Since SiT is based on flow matching [36], we observed that classification using the v-prediction led to suboptimal results. Therefore, we estimate ELBO for SiT using the ϵ-prediction objective, which consistently yields better performance. In Supplementary Materials, we ablate various ELBO loss weighting schemes for both DiT and SiT to ensure a fair comparison. Also, we evaluate SiT equipped with representation alignment (REPA) [73], which leverages SSL representations to improve generative performance. In Supplementary Materials, we further compare against MeanFlow [13], a recent few-step DM-based model.
- **Joint generative-discriminative models.** We also compare to the most recent state-of-the-art joint energy-based model, DAT [67].
- **Autoregressive models.** We compare our method with LlamaGen [54], VAR [55] and A-VARC+ [8]. For VAR, likelihoods are estimated across all spatial scales. A-VARC+ additionally employs likelihood smoothing for more accurate classification. Also, we pretrain RandAR with a fixed raster-order to take into account the model design differences with LlamaGen.

Model size	Model type	Method	IN-Val	IN-R	IN-S	IN-C Gauss	IN-C JPEG	IN-A
L/16	Discriminative	ViT [11]	0.803	0.409	0.291	**0.620**	0.694	0.166
		DINOv2 [41]	**0.819**	**0.476**	0.358	0.497	**0.736**	**0.363**
	Joint model	DAT [67]	0.764	0.379	0.364	0.569	0.709	0.034
	Diffusion	DiT [44]	0.771	0.393	0.361	0.467	0.642	0.133
	Autoregressive	LlamaGen [54]	0.640	0.298	0.232	0.358	0.452	0.143
		VAR [55]	0.656	0.255	0.177	0.137	0.390	0.083
		A-VARC+ [8]	0.717	0.277	0.175	0.161	0.391	0.072
		RandAR raster [43]	0.701	0.351	0.301	0.476	0.588	0.174
		RandAR [43] **(Ours)**	0.780	0.463	**0.406**	0.589	0.687	0.145
XL/16	Discriminative	DINOv2 [41]	**0.827**	0.486	0.354	0.500	**0.754**	**0.345**
	Diffusion	SiT [39]	0.697	0.257	0.223	0.311	0.537	0.104
		SiT + REPA [73]	0.733	0.296	0.262	0.416	0.554	0.169
		DiT [44]	0.772	0.402	0.367	0.505	0.661	0.153
	Autoregressive	LlamaGen [54]	0.559	0.248	0.200	0.359	0.391	0.210
		VAR [55]	0.630	0.222	0.153	0.146	0.336	0.090
		A-VARC+ [8]	0.719	0.259	0.143	0.051	0.353	0.083
		RandAR [43] **(Ours)**	0.813	**0.530**	**0.459**	**0.652**	**0.751**	0.233

Table 2: Top-1 accuracy on various ImageNet benchmarks for different model types.

All baselines are pretrained on ImageNet1K [50]. For methods lacking corresponding publicly available checkpoints, we reproduced the pretraining ourselves, strictly following their official implementations. We provide more details in Supplementary Materials.

To ensure a fair comparison, all models share similar transformer-based architectures, comparable model sizes, same image resolution 256×256 and the number of tokens per image (256), thereby minimizing the factors unrelated to model types. The only exception is DAT, which uses a ConvNext-L backbone.

Datasets. For evaluation, we use the ImageNet-Val set and several OOD benchmarks such as ImageNet-R/S/A [17, 19, 61] and ImageNet-C [18] for Gaussian noise and JPEG corruptions applied with a strength 3 out of 5. For all validation sets, we report top-1 accuracy. Since GCs require substantial compute for 1000-class predictions, we follow the protocol of [31] and report results on 2K subsets of IN-Val, IN-S and IN-C. We verified that increasing the subset size yields only minor variations in accuracy. 200-class benchmarks, IN-R and IN-A, are evaluated on the entire sets.

Evaluation setup. For diffusion-based classifiers and our RandAR-based ones, we provide the results using sufficiently large K for RandAR and the number of timesteps for DiT/SiT to ensure near-optimal performance. Specifically, we use K=20 and 250 timesteps, following the setup in [31]. The RandAR models are trained with the LlamaGen tokenizer [54] with latent noise augmentation, which is discussed and ablated in Section 6.4.

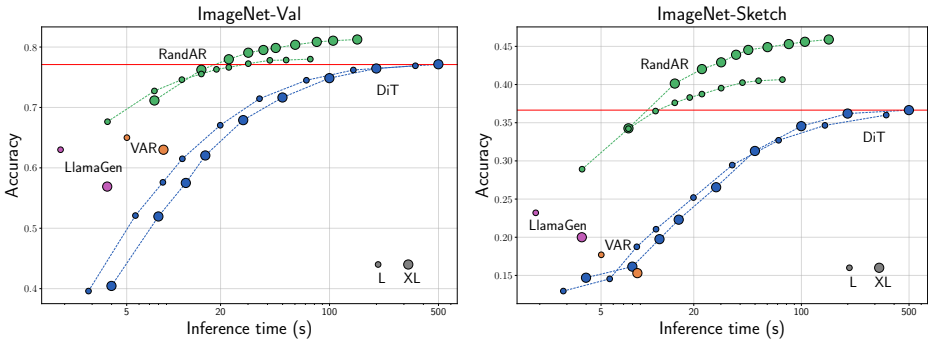

Fig. 6: Comparison of the generative classifiers in terms of top-1 classification accuracy and efficiency across different model sizes. Efficiency is reported in seconds per single-image classification.

Results. The results are presented in Table 2. First, RandAR demonstrates strong scaling potential, showing significant performance gains as model size increases from L to XL across all validation sets. Also, RandAR significantly surpasses fixed-order counterparts such as LlamaGen, VAR and the raster-order RandAR variant. Moreover, RandAR outperforms recent A-VARC, which averages predictions over multiple noise samples, resulting in comparable compute.

Compared to diffusion-based classifiers, RandAR achieves significantly superior performance on each validation set, with more pronounced gains on the OOD benchmarks. Although it is beyond the scope of this work, we would like to note that SiT, while underperforming DiT, may benefit from future exploration of alternative weighting functions and timestep schedules.

Comparing to DINOv2, RandAR shows lower in-domain accuracy by 1.4% for the XL model. Nevertheless, RandAR-XL surpasses DINOv2 on 3 of 5 OOD sets (IN-R, IN-S, IN-C Gauss), matches performance on IN-C JPEG, and only trails on IN-A.

Overall, RandAR establishes a new state-of-the-art among GCs and competes with DINOv2 — the result that has not been previously achieved with GCs.

6.2 Efficiency evaluation

Here, we evaluate various generative classifiers of different model sizes in terms of their classification accuracy with respect to the total runtime measured in seconds per a single image classification task. Both RandAR and DiT classifiers are each evaluated with different numbers of NFEs in order to approximate the respective objectives, namely $p(\mathbf{x}|c)$ for RandAR and the ELBO for DiT. We provide the corresponding experimental results for both IN-Val and IN-S datasets in Figure 6. All of the runtimes are measured on a single A100 GPU. We observe that RandAR consistently shows better classification accuracy than DiT across all the different operating points, while at the same time offering up to 25× faster inference.

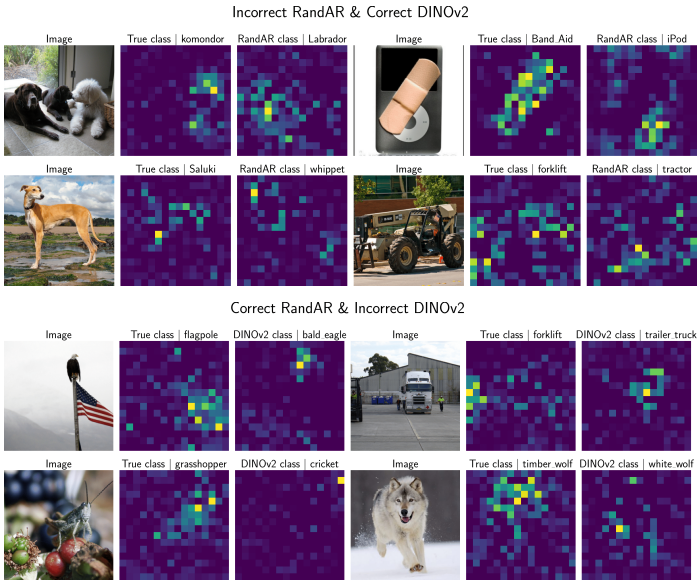

Fig. 7: Error analysis. The examples where RandAR-XL is incorrect while DINOv2-XL is correct (Top), and vice versa (Bottom). Each example is accompanied by discriminative log-likelihoods produced by RandAR for the correct and incorrect classes. Both models tend to fail on multi-object images or visually similar object categories.

6.3 Error analysis

We analyze image examples where RandAR-XL fails while DINOv2-XL succeeds, and vice versa. In Figure 7, we present representative examples illustrating typical failure cases of both models, along with the corresponding RandAR discriminative log-likelihoods computed for the true and predicted class labels.

We observe that both models exhibit similar error types, which are common among image classifiers and can be broadly grouped into two categories:

- **Similar object classes** – the models misclassify the objects that appear visually similar to another class, e.g., similar dog breeds. In such cases, the discriminative log-likelihoods are often less interpretable, as the model may assign high likelihoods to the same object.
- **Multi-object images** – the input images contain multiple distinct objects belonging to different classes, and the model simply selects one of them as the primary prediction.

Notably, we observe that even when predicting an incorrect overall class, RandAR often still assigns high likelihood scores to the correct object when it is explicitly conditioned on the true class label.

VQ-VAE	Vocab size	Flipped tokens	rFID
MaskGIT	1024	0%	**1.67**
		10%	1.71
		20%	1.85
		40%	2.49
LlamaGen	16384	0%	**1.68**
		18%	**1.67**
		27%	**1.69**
		42%	1.73

Table 3: Reconstruction quality of the MaskGIT and LlamaGen VQ-VAE tokenizers for varying levels of latent noise corruption, which alters a portion of tokens in the original token sequences.

Model size	Tokenizer	IN-Val	IN-R	IN-S	IN-A
L/16	MaskGIT	**0.780**	0.448	0.379	0.101
	LlamaGen	0.769	**0.469**	**0.409**	**0.146**
	LlamaGen + Noise aug.	**0.780**	0.463	**0.406**	**0.145**
XL/16	LlamaGen	**0.812**	0.512	0.452	**0.241**
	LlamaGen + Noise aug.	**0.813**	**0.530**	**0.459**	0.233

Table 4: Comparison of the RandAR models pretrained with different tokenizers and the proposed noise augmentation in terms of top-1 classification accuracy.

6.4 The effect of VQ-VAE tokenizers

In this experiment, we evaluate the RandAR-L models for two VQ-VAE tokenizers, MaskGIT [3] and LlamaGen [54], widely used by the recent discrete generative models. These tokenizers have different codebook sizes of 1024 in MaskGIT and 16384 in LlamaGen.

Prior work [60,74] has noticed that tokenizers producing significantly different token sequences for almost identical images may bring additional challenges for AR model training. Therefore, we first investigate the robustness of the MaskGIT and LlamaGen tokenizers to minor noise perturbations.

We evaluate the tokenizers in terms of their reconstruction quality while corrupting the continuous image latents (prior to quantization) with flow matching [36] noising process: $(1-t)z+t\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. We apply noising across different t and track the ratio of "flipped" tokens in a discrete sequence and rFID [20] on IN-Val. Table 3 shows that the original tokenizers provide similar reconstruction quality and thus can be fairly compared to each other. Then, we observe that the larger codebook is resistant to $\sim$27% flipped tokens while the small one starts degrading with $\sim$10%.

The results suggest applying the noise corruption technique as an augmentation during RandAR training with the LlamaGen tokenizer to make the model invariant to slight perturbations, not leading to noticeable image differences. Meanwhile, this augmentation is not used for the MaskGIT tokenizer since only few tokens can be changed without compromising reconstruction quality.

Finally, we compare RandAR models pretrained with different VQ-VAE tokenizers, including the one using the noise augmentation. Table 4 shows that RandAR with MaskGIT provides higher in-domain accuracy while LlamaGen shows slightly better robustness across all OOD benchmarks. Noise augmentation improves LlamaGen's in-domain accuracy while maintaining comparable

OOD performance. Therefore, in all our experiments, we use the LlamaGen tokenizer with the suggested noise augmentation.

6.5 Likelihood estimation strategy ablation

Strategy ╲ K	1	2	5	10	20
Lower bound	0.712	0.762	0.795	0.804	0.813
Log-likelihood	0.712	0.724	0.742	0.747	0.760

Table 5: Ablation of likelihood estimation strategies for order-marginalized AR GCs on IN-val.

Here, we ablate the order-marginalized likelihood estimation approach, discussed in Section 4. In Table 5, we compare direct log-likelihood estimation, $\log \mathbb{E}_\pi \left[p(\mathbf{x}|\pi, c)\right]$, with its lower bound, $\mathbb{E}_\pi \left[\log p(\mathbf{x}|\pi, c)\right]$. We observe that the lower-bound estimate achieves significantly higher classification accuracy as K increases.

6.6 Generalization beyond RandAR

In our work, we primarily focus on RandAR, as it is the only publicly available model that directly supports random-order generation. However, to demonstrate the generalization of the order-marginalized approach, we also consider another order-conditioned architecture, RAR [71].

RAR is not natively any-order, since its random permutations are gradually annealed to raster order during training. We therefore train RAR-B without order annealing to enable any-order sampling. As shown in Table 6 (left), RAR-B matches RandAR-B and outperforms DiT-B, suggesting that the benefits of order marginalization are not specific to RandAR.

6.7 Any-order AR as masked diffusion

As discussed in Section 4, any-order AR (AO-AR) models are closely related to masked diffusion (MD) models [22, 42, 52]. This connection allows RandAR to be evaluated in the MD manner, where tokens beyond the observed prefix are treated as masked and scored conditionally on the prefix. Unlike AR evaluation, which estimates the full $\log p(\mathbf{x}, |, \pi, c)$ in a single forward pass, MD scores token log-likelihoods only for a fixed prefix length per pass, requiring multiple forward passes to cover all terms.

First, we compare RandAR-B under both evaluation regimes and include MaskGIT [3] as an open-source MD baseline. Since the official MaskGIT checkpoints and training code are unavailable, we evaluate the only publicly available MaskGIT-B checkpoint[3] we found. In Table 6 (Left), RandAR-B significantly outperforms both MD counterparts under the same compute budget (K=20).

Then, we evaluate RandAR-B under both regimes as a function of the number of evaluations K. Table 6 (Right) shows that the MD-based classification approaches the AR variant only for large K, while AR evaluation remains significantly more accurate for small K.

[3] https://huggingface.co/llvictorll/Maskgit-pytorch

Model (K=20)	IN-Val	IN-R	IN-S	IN-A
DiT-B (250 steps)	0.654	0.280	0.216	0.040
MaskGIT-B	0.605	0.315	0.234	0.035
RandAR-B as MD	0.662	0.329	0.239	0.051
RAR-B	**0.741**	**0.350**	**0.303**	**0.068**
RandAR-B	0.708	**0.354**	**0.305**	**0.058**

Model ＼ K	1	2	5	10	20	60
RandAR-B as MD	0.348	0.422	0.550	0.593	0.662	0.701
RandAR-B	**0.586**	**0.640**	**0.677**	**0.692**	**0.708**	**0.711**

Table 6: Left: B-scale comparisons on ImageNet benchmarks for K=20 including MaskGIT, RAR, and RandAR evaluated as masked diffusion. **Right:** ImageNet-Val accuracy as a function of the number of evaluations K for RandAR-B scored as either an any-order AR model or masked diffusion.

6.8 Real-world distribution shifts

Finally, we evaluate RandAR in terms of robustness to real-world distribution shifts. RandAR is pretrained on each dataset for B/16 and L/16 model settings using the LlamaGen tokenizer with the proposed noise augmentation.

Model size	Method	Camelyon17		CelebA		FMoW	
		ID	OOD	ID	ID WG	ID	OOD WG
B/16	ERM [59]	0.946	0.581	**0.939**	0.342	0.310	0.137
	RWY [51]	0.945	0.578	0.923	0.426	0.291	0.139
	σ−stitching [45]	0.938	0.667	0.926	0.461	–	–
	DiT [44]	**0.992**	0.586	0.899	**0.601**	**0.571**	**0.281**
	RandAR raster [43]	**0.990**	0.647	0.897	0.572	0.423	0.205
	RandAR [43] **(Ours)**	**0.992**	**0.783**	0.908	**0.600**	0.563	**0.276**
L/16	ERM [59]	0.954	0.538	**0.939**	0.359	0.304	0.140
	RWY [51]	0.952	0.543	0.914	0.453	0.279	0.130
	σ−stitching [45]	0.942	0.677	0.922	0.419	–	–
	DiT [44]	**0.995**	0.525	0.904	0.596	**0.584**	**0.274**
	RandAR raster [43]	**0.996**	0.501	0.893	0.594	0.346	0.189
	RandAR [43] **(Ours)**	**0.996**	**0.763**	0.920	**0.639**	0.567	**0.273**

Table 7: Classification performance comparison on the datasets with subpopulation or distribution shifts.

Datasets. These experiments are conducted on the WILDS benchmark [28], which contains a collection of datasets for evaluating robustness to distribution shifts. Specifically, we consider the CelebA [37], FMoW [9], and Camelyon [1] datasets that represent a diverse range of shift types and visual domains. CelebA consists of natural face images and exhibits a subpopulation shift, and we therefore report worst-group (WG) accuracy. FMoW includes satellite images and involves both subpopulation shifts and a domain shift, for which we report OOD worst-group accuracy. Camelyon contains tissue patches from different hospitals.

Baselines. Among discriminative baselines, we consider ERM [59], RWY [23,51] and ERM with the $\sigma-$stitching loss [45]. As a backbone, we use ViT [11] and train the models from scratch following [32]. We train 3 models with different random seeds and average the results. As a generative classifier baseline, we consider DiT [44], training the B/16 and L/16 variants at an image resolution of 256×256.

Model selection. We follow the setup from [32]. Specifically, checkpoints are selected based on the class-balanced accuracy on in-domain validation set.

Results. Table 7 provides the results for both ID and OOD test sets. On Camelyon17, RandAR outperforms the supervised baselines and matches DiT on the ID set, while substantially surpassing all baselines on the OOD set. On CelebA, ERM achieves higher in-domain accuracy but performs poorly on the OOD set. RandAR exceeds DiT on the ID set and outperforms all models in WG accuracy in most cases. On FMoW, generative classifiers largely outperform discriminative methods on the ID set, although DiT shows slightly higher accuracy than RandAR. We attribute this to the relatively small training set of $\sim 77K$, where DiT may be more sample-efficient than RandAR due to stronger noise regularization. In terms of OOD WG accuracy, RandAR and DiT perform similarly and both largely surpass ERM and RWY.

In total, RandAR demonstrates high in-domain accuracy and strong robustness to subpopulation and distribution shifts, integrally outperforming DiT.

7 Conclusion

In this work, we investigate recent visual AR models as GCs and introduce a framework that leverages AR models generating images in random token orders. We show that order-marginalized AR classifiers significantly outperform diffusion-based and prior AR counterparts, achieving state-of-the-art performance in generative image classification. These findings suggest that enabling AR models to represent images using different token orders opens new perspectives, and we look forward to further development and scaling of any-order AR models to better understand the limits of order-marginalized AR classifiers.

A promising direction for future research is to leverage advances in self-supervised learning, which have already shown strong potential in raster-order AR models [74] and diffusion models [73]. Another interesting update would be to equip RandAR classifiers with image-adaptive token-order predictors, inspired by recent results in diffusion-based classifiers [62].

Moreover, since one of the key limitations of GCs is their significant computational cost, particularly for the problems with a large number of classes, it would be valuable to explore whether these classifiers can be effectively distilled into discriminative models. Such an approach could potentially combine the best of both worlds: the high inference efficiency of discriminative models and the strong classification performance of generative classifiers.

References

1. Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B.E., Lee, B., Paeng, K., Zhong, A., 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)
2. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
3. Chang, H., Zhang, H., Jiang, L., Liu, C., Freeman, W.T.: Maskgit: Masked generative image transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11315–11325 (2022)
4. Chen, H., Dong, Y., Shao, S., Zhongkai, H., Yang, X., Su, H., Zhu, J.: Diffusion models are certifiably robust classifiers. Advances in Neural Information Processing Systems **37**, 50062–50097 (2024)
5. Chen, H., Dong, Y., Wang, Z., Yang, X., Duan, C., Su, H., Zhu, J.: Robust classification via a single diffusion model. arXiv preprint arXiv:2305.15241 (2023)
6. Chen, X., Liu, Z., Xie, S., He, K.: Deconstructing denoising diffusion models for self-supervised learning. arXiv preprint arXiv:2401.14404 (2024)
7. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9640–9649 (2021)
8. Chen, Y.C., Inouye, D.I., Gao, J.: Your var model is secretly an efficient and explainable generative classifier. arXiv preprint arXiv:2510.12060 (2025)
9. Christie, G., Fendley, N., Wilson, J., Mukherjee, R.: Functional map of the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6172–6180 (2018)
10. Dieleman, S.: Diffusion is spectral autoregression (2024), https://sander.ai/2024/09/02/spectral-autoregression.html
11. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
12. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
13. Geng, Z., Deng, M., Bai, X., Kolter, Z., He, K.: Mean flows for one-step generative modeling. Advances in Neural Information Processing Systems **38**, 75460–75482 (2026)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
15. Grathwohl, W., Wang, K.C., Jacobsen, J.H., Duvenaud, D., Norouzi, M., Swersky, K.: Your classifier is secretly an energy based model and you should treat it like one. arXiv preprint arXiv:1912.03263 (2019)
16. Guo, Q., Ma, C., Jiang, Y., Yuan, Z., Yu, Y., Luo, P.: Egc: Image generation and classification via a diffusion energy-based model. arXiv preprint arXiv:2304.02012 (2023)
17. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)

18. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
19. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15262–15271 (2021)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
22. Hoogeboom, E., Gritsenko, A.A., Bastings, J., Poole, B., van den Berg, R., Salimans, T.: Autoregressive diffusion models. In: International Conference on Learning Representations (2022)
23. Idrissi, B.Y., Arjovsky, M., Pezeshki, M., Lopez-Paz, D.: Simple data balancing achieves competitive worst-group-accuracy. In: Conference on Causal Learning and Reasoning. pp. 336–351. PMLR (2022)
24. Jaini, P., Clark, K., Geirhos, R.: Intriguing properties of generative classifiers. arXiv preprint arXiv:2309.16779 (2023)
25. Kasa, S.R., Gupta, K., Roychowdhury, S., Kumar, A., Biruduraju, Y., Kasa, S.K., Priyatam, P.N., Bhattacharya, A., Agarwal, S., Huddar, V.: Generative or discriminative? revisiting text classification in the era of transformers. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 9615–9637 (2025)
26. Kim, D., Thomas, X., Ghadiyaram, D.: Revelio: Interpreting and leveraging semantic information in diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4659–4669 (2025)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
28. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International conference on machine learning. pp. 5637–5664. PMLR (2021)
29. Lai, C.H., Song, Y., Kim, D., Mitsufuji, Y., Ermon, S.: The principles of diffusion models. arXiv preprint arXiv:2510.21890 (2025)
30. Larochelle, H., Murray, I.: The neural autoregressive distribution estimator. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 29–37. JMLR Workshop and Conference Proceedings (2011)
31. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2206–2217 (2023)
32. Li, A.C., Kumar, A., Pathak, D.: Generative classifiers avoid shortcut solutions. In: The Thirteenth International Conference on Learning Representations (2024)
33. Li, H., Yang, J., Li, G., Wang, H.: Autoregressive image generation with randomized parallel decoding. arXiv preprint arXiv:2503.10568 (2025)
34. Li, T., Sun, Q., Fan, L., He, K.: Fractal generative models. arXiv preprint arXiv:2502.17437 (2025)
35. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. Advances in Neural Information Processing Systems **37**, 56424–56445 (2024)
36. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)

37. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
38. Luo, Z., Shi, F., Ge, Y., Yang, Y., Wang, L., Shan, Y.: Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. arXiv preprint arXiv:2409.04410 (2024)
39. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
40. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. Advances in neural information processing systems **29** (2016)
41. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
42. Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., Li, C.: Your absorbing discrete diffusion secretly models the conditional distributions of clean data. arXiv preprint arXiv:2406.03736 (2024)
43. Pang, Z., Zhang, T., Luan, F., Man, Y., Tan, H., Zhang, K., Freeman, W.T., Wang, Y.X.: Randar: Decoder-only autoregressive visual generation in random orders. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 45–55 (2025)
44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
45. Puli, A., Zhang, L., Wald, Y., Ranganath, R.: Don't blame dataset shift! shortcut learning due to gradients and cross entropy. Advances in Neural Information Processing Systems **36**, 71874–71910 (2023)
46. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
47. Ren, S., Yu, Q., He, J., Shen, X., Yuille, A., Chen, L.C.: Beyond next-token: Next-x prediction for autoregressive visual generation. arXiv preprint arXiv:2502.20388 (2025)
48. Rissanen, S., Heinonen, M., Solin, A.: Generative modelling with inverse heat dissipation. In: International Conference on Learning Representations (ICLR) (2023)
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
50. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
51. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. arXiv preprint arXiv:1911.08731 (2019)
52. Shih, A., Sadigh, D., Ermon, S.: Training and inference on any-order autoregressive models the right way. Advances in Neural Information Processing Systems **35**, 2762–2775 (2022)
53. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)

54. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024)
55. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024)
56. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
57. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. pp. 1747–1756. PMLR (2016)
58. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
59. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)
60. Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., Baranchuk, D.: Switti: Designing scale-wise transformers for text-to-image synthesis. arXiv preprint arXiv:2412.01819 (2024)
61. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019)
62. Wang, Y., Chen, L.: Noise matters: Optimizing matching noise for diffusion classifiers. arXiv preprint arXiv:2508.11330 (2025)
63. Wang, Y., Ren, S., Lin, Z., Han, Y., Guo, H., Yang, Z., Zou, D., Feng, J., Liu, X.: Parallelized autoregressive visual generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12955–12965 (2025)
64. Yang, X., Wang, X.: Diffusion model as representation learner. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18938–18949 (2023)
65. Yang, X., Zhou, D., Feng, J., Wang, X.: Diffusion probabilistic model made slim. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 22552–22562 (2023)
66. Yenduri, G., Srivastava, G., Maddikunta, P.K.R., Jhaveri, R.H., Wang, W., Vasilakos, A.V., Gadekallu, T.R., et al.: Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv preprint arXiv:2305.10435 (2023)
67. Yin, X., Zhang, C., Steele, J., Shavit, N., Wang, T.T.: Joint discriminative-generative modeling via dual adversarial training. arXiv preprint arXiv:2510.13872 (2025)
68. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved vqgan. arXiv preprint arXiv:2110.04627 (2021)
69. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation (2022), https://arxiv.org/abs/2206.10789
70. Yu, L., Lezama, J., Gundavarapu, N.B., Versari, L., Sohn, K., Minnen, D., Cheng, Y., Gupta, A., Gu, X., Hauptmann, A.G., Gong, B., Yang, M.H., Essa, I., Ross,

D.A., Jiang, L.: Language model beats diffusion - tokenizer is key to visual generation. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=gzqrANCF4g
71. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Randomized autoregressive visual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18431–18441 (2025)
72. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. Advances in Neural Information Processing Systems **37**, 128940–128966 (2024)
73. Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., Xie, S.: Representation alignment for generation: Training diffusion transformers is easier than you think. arXiv preprint arXiv:2410.06940 (2024)
74. Yue, X., Wang, Z., Wang, Y., Zhang, W., Liu, X., Ouyang, W., Bai, L., Zhou, L.: Understand before you generate: Self-guided training for autoregressive image generation. arXiv preprint arXiv:2509.15185 (2025)