Multi-Scale Representation Alignment for Visual Autoregressive Modeling with Mixture of Experts

Nuoyan Zhou[1,2†], Zhijun Tu[2], Lei Yu[2], Kun Cheng[1], Jie Hu[2*], Nannan Wang[1*], and Xinghao Chen[2]

[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
nuoyanzhou@stu.xidian.edu.cn, kuncheng.xidian@gmail.com, nnwang@xidian.edu.cn
[2] Huawei Technologies Co., Ltd.
{zhijun.tu, yulei96, hujie23, xinghao.chen}@huawei.com

Abstract. Visual AutoRegressive modeling (VAR) has pioneered a coarse-to-fine multi-scale autoregressive generative paradigm, demonstrating strong capabilities in image generation. However, VAR still suffers from inherent deficiencies in multi-scale representation learning. Specifically, lower scales primarily capture global semantics, while higher scales focus on fine-grained details. Employing a shared architecture across scales induces optimization conflicts. Moreover, due to the causal autoregressive process, inaccurate semantics at early scales can propagate and significantly degrade the final output. To address these issues, we introduce a scale-aware token-routed Mixture of Experts (MoE) architecture, allowing scale-adaptive expert selection, thereby facilitating decoupled representation learning across scales. In addition, we enhance semantic modeling at early scales by incorporating external self-supervised features. Unlike naive alignment, we analyse and design a residual feature aggregation scheme tailored to the VAR paradigm. Extensive experiments show that our method significantly improves both training efficiency and generation quality. On the ImageNet 256×256 benchmark, our model achieves a superior FID compared to the dense baseline while requiring only half of the default training epochs and a smaller parameter budget, with a merely marginal increase in training cost. Moreover, the performance gap further widens with larger training epochs.

1 Introduction

Autoregressive (AR) models, empowered by their remarkable scaling laws [19,30], have achieved tremendous success in natural language processing (NLP) [1, 2, 48, 50]. This success has inspired their adaptation to the computer vision [11, 51, 53, 56, 65], where researchers aim not only to replicate the breakthroughs witnessed in NLP, but also to explore AR modeling as a unified paradigm for both image generation and understanding tasks [47, 53]. A typical AR paradigm for image generation involves two stages: it first discretizes images into tokens via a tokenizer [6, 11, 26, 49, 52, 55, 61] and then utilizes an AR network to generate samples in a sequential manner. Built on this foundational paradigm, recent works [23, 35, 40, 42, 46, 49, 54, 60] have made great progress in image generation, demonstrating comparable performance to diffusion models.

† This work was done during an internship at Huawei Technologies Co., Ltd.
* Corresponding authors

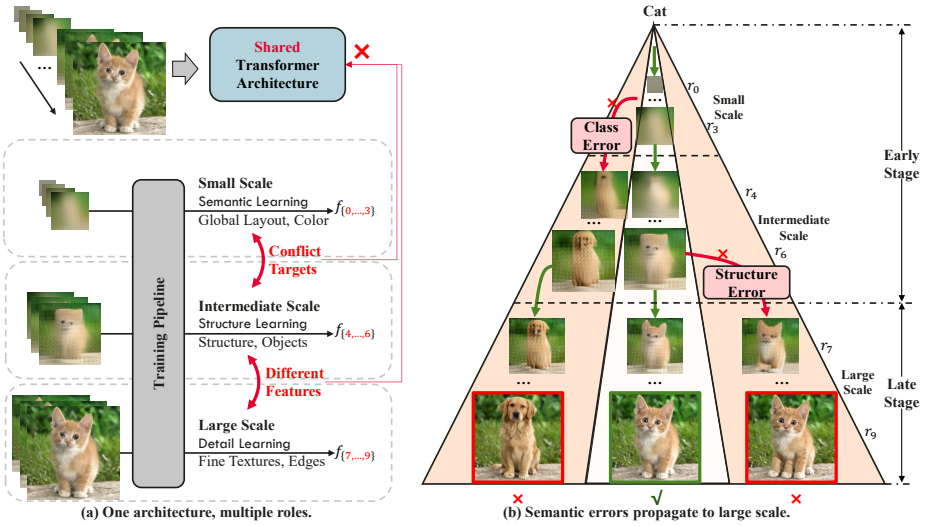

Fig. 1: Empirical analysis of representation space and expert load. (a) VAR learns different features across scales with a shared architecture, resulting in conflicts of optimization objectives. (b) Semantic errors at small and intermediate scales can propagate across scales and induce unpleasing results. Correct semantics at early stage are critical for predictions at later stage.

Specifically, Visual AutoRegressive modeling (VAR) [49] stands out among AR methods for its high-quality and fast image generation. It has pioneered a coarse-to-fine autoregressive modeling via next scale prediction, which decomposes the image data into multi-scale residual representations and builds causality across scales. This design preserves the 2D structure of images and supports efficient parallel decoding, leading to excellent performances in both generation quality and inference speed. However, it suffers from difficulties in multi-scale representation learning. In this paper, we argue that a Mixture of Experts (MoE) [25] architecture and semantic guidance in the representation space can substantially mitigate these issues.

We start by analysing the feature distribution of VAR representations. As shown in Fig.1a, different scales focus on different learning targets and generate distinct feature spaces, which requires the model to play roles with multiple capabilities across scales. Since all the scales share a single Transformer architecture, these differences induce conflicts in optimization and difficulties in multi-scale representation learning. Besides, due to causal nature across scales, predictions at higher scale depends heavily on the feature map produced at earlier scales. As shown in Fig.1b, inaccurate semantics at early scales can propagate and adversely affect later scales, leading to unpleasing outputs.

To address these challenges, we propose multi-scale representation alignment (MEPA), a token-routed Mixture of Experts (MoE) framework for guiding efficient representation learning with semantically-enriched visual representations. First, we introduce a MoE architecture that adaptively decouples model capacity across scales, thereby implicitly reducing the adverse effect of conflicting optimization and promoting specialized feature learning. Our approach evolves from a scale-based routing scheme to a size-aware

token routing scheme, due to the latter preventing routing homogenization within the same scale caused by the former and improving load balancing of expert training(see Sec. 4.2). Next, we incorporate self-supervised representations to guide the semantic information at early stage in VAR models, mitigating the propagation of semantic errors throughout the autoregressive generation process. However, there exists an inherent feature gap between VAR models and self-supervised models, which prevents direct alignment [58, 62]. Thus, we conduct a series of empirical analyses on transforming VAR features to bridge the representation gap and inject accurate semantic information (see Sec. 4.3). To the best of our knowledge, this is the first systematic study of representation alignment in the VAR paradigm. Our experiments show that MEPA significantly improves both training efficiency and generation quality. It achieves superior performances compared to the baseline with a half of the default training epochs and a smaller parameter budget. Moreover, the performance advantages further widens with larger epochs. We highlight our main contributions below:

- We highlight the challenges of multi-scale representation learning in VAR, systematically analysing the inherent difficulties caused by the shared architecture and semantic errors propagation.
- We propose MEPA, a scale-aware MoE framework that promotes decoupled representation learning across scales and strengthens semantic information.
- MEPA achieves impressive performances on both training efficiency and generation quality. It reaches up to a 2× training speedup over the VAR training process.

2 Related Works

Autoregressive Model in Image Generation. The autoregressive (AR) methods mostly adopt a two-stage paradigm, which discretizes images into quantized tokens by a pretrained tokenizer [11,26,49,52,55,61] and then generate samples in a sequential manner. Built on this foundational paradigm, recent works [23,35,40,42,46,49,54,60] have made great progress in the performance of image generation. The basic unit generated at a single AR step can take various forms (such as token [40,46,60], scale [16,23,27,49,63], patch [39, 42], frequency [12, 24, 59], etc [22, 35, 37, 54]), enabling the AR model to simultaneously improve the quality of generated images and the efficiency of parallel inference. Besides, there are numerous efforts in exploring high-efficiency schemes of visual generation models. A few works [3, 6, 61] focus on reduce compressed token numbers of a image. [15] propose a pivotal token selection strategy to generate only pivotal tokens for fast inference. [63] apply a scale-Markov trajectory to cut off redundant KV cache. [8] use a large VAR model to generate the tokens at small scales and a small VAR model to efficiently predict the remaining tokens. However, few studies have focused on the training efficiency of AR models, which is precisely what our work aims to emphasize and address.

Self-supervised Representation Learning. Recent advances in self-supervised representation learning can be broadly categorized into contrastive learning and masked image reconstruction. The methods trained by contrastive learning, such as including DINO [38], SimCLR [7], MoCo [18], and BYOL [14], learn informative features

without explicitly modeling the data distribution, producing highly separable representations that excel in classification, retrieval, and dense prediction, but often discard fine grained generative information, limiting their utility for synthesis or reconstruction. The approaches trained by masked image reconstruction, including VAE [31], masked autoencoders (MAE) [17], masked image modeling (MIM) [57], diffusion models [20], and autoregressive (AR) models [2] capture rich contextual and perceptual information by reconstructing inputs or modeling the underlying data distribution, providing embeddings beneficial for downstream tasks; however, they are computationally intensive and may produce representations that are less discriminative than contrastive methods. The success of [36, 58, 62] has demonstrated that generation tasks benefit from pretrained self-supervised representations, which inspires us to transfer it to the AR paradigm for the first time.

3 Preliminaries

VAR employs next-scale prediction to generate multiple tokens simultaneously. Its tokenizer first compresses the image x into a feature map $f_0 \in \mathbb{R}^{H\times W\times c}$ and decomposes it as a set of multi-scale residual features $\{f_1, f_2, \ldots, f_K\}$. Specifically, it quantizes the downsampled feature into tokens $\{r_1, r_2, \ldots, r_K\}$ in ascending order of scale, and then obtains the next-scale feature by subtracting the quantized feature from the current-scale feature. The process can be formulated as

$$Q(z) = \arg\min_{v_i \in V} \|v_i - z\|, \tag{1}$$

$$r_k = Q(Down(f_{k-1}, sz_k)), \tag{2}$$

$$f_k = f_{k-1} - \sum_{i=1}^{k-1} Up(Res_i(LU(r_i)), sz_K), \tag{3}$$

where $Q(\cdot)$ denotes the quantization operation, V denotes the codebook, v_i denotes the ith feature in the codebook, $Down(\cdot)$ and $Up(\cdot)$ denotes the down and up sample operation, $Res_k(\cdot)$ denotes the ResNet block for the kth scale, sz_k denotes the feature size (H_k, W_k) at the kth scale, $LU(\cdot)$ means retrieving features from the codebook based on indices. VAR builds a causal relationship upon multi-scale features with conditions c:

$$p(r_1, r_2, \ldots, r_K) = \sum_{k=1}^{K} p(r_k | r'_1, r'_2, \ldots, r'_{k-1}, c), \tag{4}$$

where $r'_i = Down(\sum_{j=1}^{i} Up(Res_j(LU(r_j)), sz_K), sz_{i+1})$ means merging all preceding residual features as an accumulated input. In the training process, all scales except the last one are concatenated along the sequence dimension to predict the probability distribution. During the inference, VAR can generate tokens in parallel from small to large scales in a autoregressive method. The training loss can be formulated as

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}_x \left[CE\big(p(r_1, r_2, \ldots, r_K);\ r_1, r_2, \ldots, r_K\big) \right], \tag{5}$$

where $CE(\cdot;\cdot)$ denotes the cross-entropy loss.

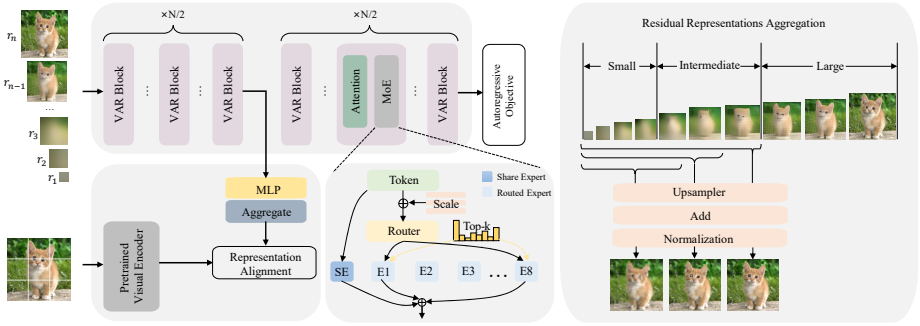

Fig. 2: An overview of our proposed MEPA framework. We follow the next-scale prediction paradigm of VAR [49], and introduce MoE layers to route tokens across scales, allowing them to adaptively select experts. During training, the outputs of the middle block in small and intermediate scales are aggregated to align with a pretrained visual encoder's features.

4 Method

4.1 Overview of MEPA

We introduce MEPA, a scale-aware MoE framework for guiding efficient representation learning with semantically-enriched visual representations, as shown in Fig. 2. MEPA extends the model to a multi-experts architecture adapting to specialization at every scale, as detailed in Sec. 4.2. It decouples model capacity across scales and induce scale-specialized representation learning. Besides, MEPA heuristically aggregates residual latents for further alignment, as detailed in Sec. 4.3. This allows the VAR model to leverage semantically-rich external representations for generation, preventing semantic errors at early scales. Our key contributions lie in systematically evaluating different MoE architectures and aggregating strategies for multi-scale residual representations from the perspectives of semantic level, training efficiency, and generation performance.

4.2 Scale-Aware Token-Routed Mixture of Experts

Mixture of Experts. There is a fundamental insight of MoE that different components of a model can specialize in handling distinct tasks. By dynamically activating only relevant modules, MoE enables efficient scaling of model capacity while maintaining computational efficiency. Let's formally define a general MoE layer with M experts, which are implemented as Feed-Forward Network (FFN) with the same architecture, denoted by $E_1, E_2, ..., E_M$. The MoE layer employs a router to select a part of experts to process input tokens and replaces vanilla FFN layer in the transformer block. Consider input $x \in \mathbb{R}^{B\times N\times d}$ where B is the batch size, N is token length of one sample, and d is the hidden dimension. The output of the MoE layer is the average weighted result of multi experts, defined as follows:

$$y_i = \sum_{j=1}^{m} G_{i,j} E_j(x_i), x_i \in \mathbb{R}^{B\times 1\times d}, \tag{6}$$

Table 1: Mean and variance of routed tokens by different MoE. We separately train two VAR-d16 models based on SMoE and STMoE. Then we compute the average activated times of each expert across all layers. For clarity, we normalize the total number of processed tokens to 64.

Expert	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	Variance
SMoE	11.35	5.61	9.13	5.18	5.71	9.23	11.32	6.46	6.62
STMoE	7.92	7.64	11.56	6.91	6.44	7.54	8.31	7.67	2.41

where i is the token indice along the sequence length, j is the expert indice, $G_{i,j}$ indicates whether the jth expert receives the ith token. Obtaining $G_{i,j}$ involves two components: a router to compute the token-to-expert affinity score $S_{i,j}$ and a gating function to achieve sparse expert activation. The routing scheme is evidently crucial for tailoring the specialized learning and preference relationship of experts, which will be detailed below. Here, a common gating function is defined as

$$G_{i,j} = \begin{cases} S_{i,j}, & S_{i,j} \in topK(\{S_{i,j}\}_{j=1}^{M}, m), \\ 0, & otherwise. \end{cases} \tag{7}$$

Scale-routed MoE. We first design a Scale-routed MoE (SMoE) structure for specialized representation learning at different scales. A routing matrix $W \in \mathbb{R}^{d \times M}$ and learnable scale embeddings $SE \in \mathbb{R}^{K \times d}$ are utilized to calculate the scale-to-expert affinity matrix $S \in \mathbb{R}^{B \times N \times M}$:

$$S = Softmax_E(SE(scale(x))W), \tag{8}$$

where $Softmax_E$ denotes the softmax operation along the expert axis, $scale(\cdot)$ denotes finding the scale indice corresponding to the input token. SMoE framework sets up a scale-based router clearly assigning tokens at different scales to distinct experts, thereby decoupling model capabilities across scales and boosting multi-scale representation learning.

However, SMoE can only provide the same activation scheme for all tokens at the same scale. The routing homogenization within the same scale will cause new problems. Due to the significant difference in the number of tokens across scales, simple scale-based routing leads to an imbalanced load for experts, severely affecting training efficiency. As shown in Tab.1, the variance of the token numbers processed by different experts in the SMoE model are relatively large. Moreover, different tokens within the same scale may have varying semantic richness and importance due to their positions. SMoE applies an identical expert activation pattern to all tokens within the same scale, which limits the model's capability and causes routing homogenization within the same scale. Due to the above shortcomings, SMoE only achieves the suboptimal generation performance, as illustrated in Tab.4.

Scale-aware Token-routed MoE. To handle these issues, we further propose a Scale-aware Token-routed MoE (STMoE) structure, which integrates the scale embeddings into tokens to perceive the scale information, and then employs a token-based router to adaptively enable specialized representation learning and enhance the flexibility of

routing. STMoE calculate the token-to-expert affinity matrix $S \in \mathbb{R}^{B\times N\times M}$ as follows:

$$S = Softmax_E((SE(scale(x)) + x)W). \tag{9}$$

In practice, the VAR transformer adds positional and scale embeddings to input tokens, naturally meeting the need for scale awareness. Benefiting from scale-aware tokens, STMoE can not only assign experts with distinct semantic granularity to different tokens within the same size, but also efficiently learn scale-specificity with superior expert load balancing (as demonstrated in Tab.1 and Tab.4).

Load Balance Loss. Directly MoE strategies often encounter the issue of load imbalance. Existing study [66] has shown that evenly activated experts in an MoE layer can lead to better performance. Besides, expert load balancing is crucial for training efficiency in distributed training scenarios. To achieve balanced loading among different experts, we have incorporated a load-balance loss, $\mathcal{L}_b$, which is widely used in previous works [13, 33].

$$\mathcal{L}_b = M \sum_{j=1}^{m} (\frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(G_{i,j} \neq 0))(\frac{1}{N} \sum_{i=1}^{N} S_{i,j}), \tag{10}$$

where $\mathbb{I}(G_{i,j} \neq 0)$ denotes the indicator function that image token i selects expert j.

4.3 Residual Aggregation and Semantic Guidance

To prevent semantic errors at early stages of autoregressive generation, we introduce a Semantic Guidance (SG) mechanism that distills VAR representations toward pretrained self-supervised visual features, following the existing work of representation alignment [58, 62].

Unlike conventional alignment approaches that constrain the final representation, our design is motivated by the causal structure of VAR. Since generation proceeds from small to large scales, predictions at later stage are strictly conditioned on representations produced at earlier stage. Consequently, inaccurate semantics generated at early stage (small and intermediate scales) can propagate and be amplified in subsequent generation. Therefore, semantic regularization should be imposed at early scales rather than only at the final scale.

However, directly aligning individual residual representations is suboptimal. VAR constructs a multi-scale residual representation space, while self-supervised vision encoders learn patch-based complete representations. There is an inherent gap between VAR and self-supervised representations leading to mismatched alignment. Moreover, pretrained vision encoders are optimized for discriminative tasks rather than generation, and their representations may not preserve fine-grained details required at later stage (large scales). To better inject semantic information into the VAR feature space, we first transform the residual representation space into progressively enriched aggregated representations before alignment.

Specifically, residual features are sorted in ascending order of scale, and then they are upsampled to match the shape of the self-supervised representations. We perform cumulative summation on the features from the first scale to each subsequent scale to obtain a set of aggregated ones with increasing semantic richness and information

content. Consider an input image x and its output of pretrained encoder $g \in \mathbb{R}^{N\times D}$, where N is the patch numbers and D is the channel dimension. Let $m_\phi(h_\theta)_{\{1,2,\ldots,K\}}$ denote a projection of the VAR transformer output $h_\theta = f_\theta(c, r'_1, r'_2, \ldots, r'_{K-1})$ that through a multilayer perception (MLP) m_ϕ. In practice, we select the output of the intermediate layer in a VAR transformer, avoiding straightforward alignment and reaching the best generation performance. The group of aggregated representations can be formulated as

$$z_j = \sum_{i=1}^{j} Up(m_\phi(h_\theta)_i, sz_K). \tag{11}$$

Instead of aligning the fully aggregated representation, we selectively apply semantic guidance to aggregated features at small and intermediate scales. This design directly reflects the autoregressive dependency: by strengthening semantic correctness at early stage, we provide a reliable foundation for subsequent large-scale prediction, thereby reducing semantic error propagation.

The objective of semantic guidance maximizes patch-wise similarity between the pretrained representation $g \in \mathbb{R}^{N\times D}$ and the selected aggregated features:

$$\mathcal{L}_{SG}(\theta, \phi) = -\mathbb{E}_x \left[\frac{1}{N} \sum_{i=1}^{N} \sum_{j\in F} sim(g^i, z_j^i) \right], \tag{12}$$

where i denotes the patch index, $sim(\cdot,\cdot)$ is cosine similarity, and F represents the selected early-stage aggregation group (i.e., aggregated features at scales 1-5, scales 1-6, and scales 1-7).

The final training objective combines semantic guidance with the original autoregressive loss:

$$\mathcal{L} = \mathcal{L}_{SG} + \lambda \mathcal{L}_{CE}, \tag{13}$$

where $\lambda > 0$ controls the balance between autoregressive learning and semantic regularization.

By explicitly enhancing semantic modeling at small and intermediate scales, semantic guidance stabilizes the autoregressive generation chain and improves both convergence speed and final generation quality.

5 Experiments

5.1 Setup

Benchmarks and Baselines. In this paper, we evaluate the generative capability on the ImageNet 256×256 class-conditional benchmark and compare our quality results with state-of-the-art image generation models, the metrics include Fréchet inception distance (FID), inception score (IS), precision and recall. Our baselines cover GAN, Diffusion, AR (next-token prediction paradigm), and existing VAR methods. For a fair comparison, we report the number of parameters, the number of steps required to generate an image, and the number of training epochs for each approach.

Fig. 3: Generation results of MEPA on VAR-d16. We use classifier-free guidance with $w = 4.0$.

Implementation Details. We train MEPA models with depths 12 and 16 on ImageNet-1K dataset [9]. All models are trained under identical configurations. We use the AdamW optimizer with a batch size of 96 and a weight decay of 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.95$. The basic learning rate is scheduled to decay from $1e-4$ to $1e-5$ following a linear annealing strategy, which is the same as VAR [49]. The hyper-parameter $\lambda = 0.5$.

Comparative Method. We compare our method with several representative approaches across GANs, diffusion models, and autoregressive models, including BigGAN [4], GigaGAN [29], StyleGAN-XL [44], ADM [10], CDM [21], LDM [43], DiT [41], VQGAN [11], RQTran [32], RCG [34], and MaskGIT [5]. In addition, we compare our baseline VAR with other state-of-the-art autoregressive generative models, including MAR [35], FAR [59], LlamaGen [46], PAR [54], FlexVAR [27], and SpectralAR [23]. Compared with these methods, we emphasize that our approach is designed to achieve high training efficiency rather than state-of-the-art generative performance.

5.2 Main Results

We compare MEPA with existing generative methods on the ImageNet-1K benchmark [9], including GAN, diffusion models, AR models and VAR models. To ensure a fair comparison, we report the total number of parameters activated by STMOE in MEPA and compare MEPA with the dense baseline with similar parameters. As shown in Tab 2, MEPA achieves state-of-the-art performance among all methods with the default setting of training epochs, indicating great enhancement in generative quality. To

Table 2: Qualitative comparison on class-conditional ImageNet 256×256. ↑ and ↓ denote that higher and lower values are better, respectively. Param represents the number of parameters, while Step indicates the number of steps required to generate an image. ‡ means we reproduce this baseline with our setting.

Type	Model	FID↓	IS↑	Precision↑	Recall↑	Param	Step	Epoch
Generative Adversarial Networks (GANs)								
GAN	BigGAN [4]	6.95	224.5	0.89	0.38	112M	1	-
GAN	GigaGAN [29]	3.45	225.5	0.84	0.61	569M	1	-
GAN	StyleGan-XL [44]	2.30	265.1	0.78	0.53	166M	1	-
Diffusion Models								
Diff	ADM [10]	10.94	101.0	0.69	0.63	554M	250	-
Diff	CDM [21]	4.88	158.7	-	-	-	8100	-
Diff	LDM-4-G [43]	3.60	247.7	0.87	0.48	400M	250	-
Diff	DiT-L/2 [41]	5.02	167.2	0.75	0.57	458M	250	-
Diff	DiT-XL/2 [41]	2.27	278.2	0.83	0.57	675M	250	-
Autoregressive Models								
AR	VQGAN-re [11]	18.65	80.4	0.78	0.26	227M	256	-
AR	RQTran [32].	13.11	119.3	-	-	821M	68	-
AR	RCG [34]	3.49	215.5	-	-	502M	20	-
AR	MaskGIT [5]	6.18	182.1	0.80	0.51	227M	8	300
AR	MAR-B [35]	2.31	281.7	0.82	0.57	208M	64	800
AR	FAR-B [59]	4.26	248.9	0.79	0.51	208M	10	400
AR	LlamaGen-L [46]	3.80	248.3	0.83	0.51	343M	256	300
AR	LlamaGen-XL [46]	3.39	227.1	0.81	0.54	775M	256	300
AR	PAR-L [54]	3.76	218.9	0.84	0.50	343M	147	300
AR	PAR-XL [54]	2.61	259.2	0.82	0.56	775M	147	300
Visual Autoregressive Models								
VAR	VAR-d16 [49]	3.55	280.4	0.84	0.51	310M	10	200
VAR	VAR-d20 [49]	2.95	302.6	0.83	0.56	600M	10	250
VAR	FlexVAR-d16 [27]	3.05	291.3	0.83	0.52	310M	10	180
VAR	FlexVAR-d20 [27]	2.41	299.3	0.85	0.58	600M	10	250
VAR	SpectralAR-d16 [23]	3.02	282.2	0.81	0.55	310M	64	200
VAR	SpectralAR-d20 [23]	2.49	305.4	-	-	600M	64	250
VAR	**MEPA-d12 (Ours)**	2.86	288.44	0.82	0.55	252M	10	200
VAR	**MEPA-d16 (Ours)**	2.32	311.26	0.82	0.57	585M	10	200
VAR	VAR-d16‡ [49]	4.10	241.6	0.85	0.47	310M	10	100
VAR	FlexVAR-d16‡ [27]	7.68	176.63	0.77	0.49	310M	10	100
VAR	FlexVAR-d20‡ [27]	5.77	215.04	0.77	0.52	600M	10	100
VAR	**MEPA-d12 (Ours)**	3.27	260.97	0.81	0.53	252M	10	100
VAR	**MEPA-d16 (Ours)**	2.65	304.60	0.82	0.56	585M	10	100

show its advantages of training efficiency, we report the performance of MEPA with 100 epochs. As shown in Tab 2, MEPA-{d12,d16} easily outperform other VAR methods under 100 epochs, and slightly surpass VAR-{d16,d20} under the default training epochs by −0.28 and −0.30 FID decrease. It gains comparable FID/IS performance to other VAR methods while requiring only about a half of training epochs, revealing remarkable enhancement in training efficiency. To validate the performance under fully converged settings, we further train VAR models with MEPA for 200 epochs, matching the default VAR training schedule to ensure sufficient convergence. As shown in Tab 2, MEPA further outperforms the corresponding VAR baseline, extending the advantage on FID to −0.69 and −0.63. These empirical results demonstrate that MEPA consistently outperforms baselines both before and after full convergence, demonstrating it indeed makes progress in both generation performance and training efficiency.

Table 3: Ablation Results of STMoE and semantic guidance. We evaluate the impact of STMoE and SG on generation performance on VAR-d16 models trained by 100 epochs.

STMoE	SG	IS↑	sFID↓	FID↓	Precision↑	Recall↑
✖	✖	241.6	8.75	4.10	**0.85**	0.47
✔	✖	284.2	**8.04**	2.99	0.84	0.54
✖	✔	269.0	8.58	3.59	0.84	0.50
✔	✔	**304.6**	8.25	**2.65**	0.82	**0.56**

Table 4: Ablation Results on MoE types. We evaluate the impacts of different types of MoE on generation performance on VAR-d16 models **without semantic guidance**.

Model	FID ↓	sFID ↓	IS ↑	Precision	Recall
VAR	4.10	8.75	241.6	**0.85**	0.47
+MoEEC [64]	4.06	8.81	234.9	0.84	0.48
+MoE++ [28]	3.39	8.13	271.7	0.84	0.51
+SMoE	3.11	**7.77**	280.7	0.84	0.52
+STMoE	**2.99**	8.04	**284.2**	0.84	**0.54**

5.3 Ablation Studies

In this study, we verify the proposed MEPA through detailed ablation experiments, with a focus on evaluating the effectiveness of its core mechanisms and the contribution of each component.

STMoE and Semantic Guidance. We evaluate the impact of STMoE and semantic guidance on generation performance. As shown in Tab. 3, both STMOE and semantic guidance can achieve significant improvement on generation performance, and their combination gains the best overall performance. This indicates that the two strategies not only function independently but also complement each other in alleviating the challenges associated with multi-scale residual representation learning.

Routing Policy of MoE. We evaluate the impacts of different MoE strategies on the VAR model without semantic guidance. As shown in Tab. 4, MoEEC [64] and +MoE++ [28] offers no appreciable benefits in practice, as they does not incorporate scale-aware design. This makes them struggle to perceive scale changes and decouple models' learning capabilities. Although SMoE best aligns with the principle that "features of different scales require independent modeling," it still performs slightly worse than STMoE. This is caused by its routing homogenization within the same scale, consistent with our analysis in Sec. 4.2.

Besdeis, We further visualize the routing heatmaps of SMoE and STMoE in Fig.4. Although both SMoE and STMoE can decouple experts' preferences across scales,

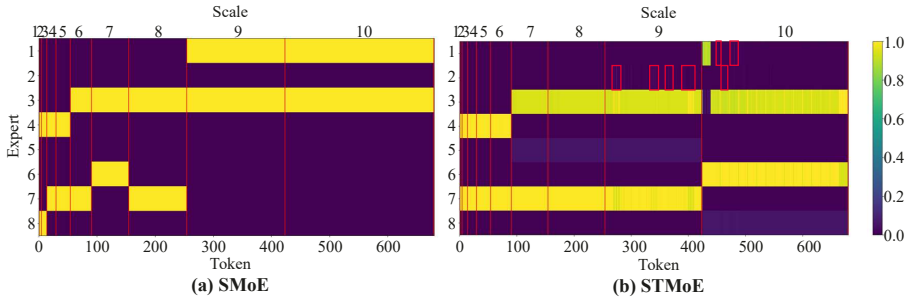

Fig. 4: Routing heatmaps of SMoE and STMoE. We visualize the routing heatmap of VAR-d16 models based on SMoE and STMoE in the penultimate layer. All the models are trained **without semantic guidance**. For better visualization, we highlight several points with relatively low activation values using red boxes.

STMoE exhibits two clear advantages. First, it activates eight experts, whereas SMoE utilizes only six, resulting in higher expert utilization. Second, STMoE enables adaptive routing based on token semantics and spatial positions. For example, at scale 10, Expert 3 prefers tokens at the boundary of the feature map, Expert 6 favors non-boundary tokens, and Expert 1 primarily attends to tokens in the first row of the feature map. Such a complex routing pattern emerges from adaptive learning conditioned on token-level semantic and positional information. By avoiding homogeneous routing within the same scale, STMoE provides greater flexibility and more fine-grained expert specialization, which promotes decoupled representation learning in the multi-scale feature space.

Target Representation for Semantic Guidance. We employ different pretrained self-supervised visual encoders for semantic guidance, as listed in Tab.5. Specifically, we align the features from the intermediate layer in the VAR model with those extracted from pre-trained encoders. Our experimental results demonstrate that MoCov3 [18] provides no substantial benefit for promoting model convergence, while DINOv3 [45] achieves slightly superior performance compared to DINOv2 [38]. This is because the training pipeline of DINOv3 is more complex, and its prediction of structure and semantics is more accurate. Therefore, we adopt DINOv3 for semantic guidance in MEPA.

Residual Representations Aggregation Policy. Following the division strategy illustrated in Fig.2, we aggregate residual features across different scales in the VAR model and then align these aggregated features with those extracted by DINOv3. The corresponding results are reported in Tab.6. Notably, the naive feature alignment strategy (denoted as "+SG-Last"), which applies alignment only after aggregating residual features from all scales into a complete representation, yields limited performance gains for the VAR model. We attribute this to the causal nature of the autoregressive generation process in VAR. Since generation proceeds in a coarse-to-fine manner, predictions at larger scales are conditioned on feature maps produced at smaller scales. Directly constraining the complete feature representation does not explicitly strengthen the semantic correctness of early-scale predictions, and therefore cannot effectively prevent semantic errors from propagating to later scales.

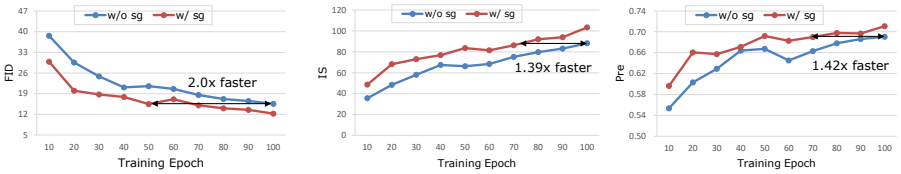

Fig. 5: Comparison of VAR with/without semantic guidance. We conduct a comprehensive evaluation of generation performance for models trained with and without semantic guidance. All the models are dense transformers **without MoE layers**.

Table 5: Ablation Results on External Self-supervised Encoders for Semantic Guidance. We evaluate the impact of different self-supervised vision models on generation performance on VAR-d16 models trained with semantic guidance. "+SG-Mocov3", "+SG-DINOv2", and "+SG-DINOv3" denote aligning with external representations from Mocov3, DINOv2, and DINOv3, respectively. All the models are dense transformers **without MoE layers**.

Model	FID ↓	sFID ↓	IS ↑	Precision	Recall
VAR	4.10	8.75	241.6	**0.85**	0.47
+SG-Mocov3 [18]	4.11	8.76	250.4	**0.85**	0.46
+SG-DINOv2 [38]	3.81	8.43	260.8	**0.85**	0.48
+SG-DINOv3 [45]	**3.59**	**8.58**	**269.0**	0.84	**0.49**

Table 6: Ablation Results on Residual Representation Aggregation. We evaluate generation performance of VAR-d16 models under different residual representation aggregation strategies trained with semantic guidance. "+SG-Last", "+SG-L", "+SG-S", and "+SG-M" denote selecting the last/top-3 largest/top-3 smallest/3 intermediate accumulated features, respectively. All the models are dense transformers **without MoE layers**.

Model	FID ↓	sFID ↓	IS ↑	Precision	Recall
VAR	4.10	8.75	241.6	**0.85**	0.47
+SG-Last	3.81	8.43	260.8	**0.85**	0.48
+SG-L	3.68	8.71	252.2	0.84	0.49
+SG-S	3.64	8.77	254.6	0.84	**0.50**
+SG-M	**3.59**	**8.58**	**269.0**	0.84	0.49

To better align with the autoregressive dependency structure, we instead apply semantic alignment to selectively aggregated features at earlier stage. As discussed in Sec. 4.3, we set three groups of optional aggregation scales. Then we search the best scale group by generation performances. As shown in Tab.6, the intermediate scale group (i.e., aggregated features at scales 1-5, scales 1-6, and scales 1-7) reach the best generation quality. This result is consistent with our motivation: strengthening semantic regularization at small and intermediate scales improves the quality of the coarse semantic sketches generated at early stage. Since large-scale predictions are conditioned on these representations of earlier scales, a more accurate semantic foundation reduces error propagation and facilitates more reliable fine-grained detail generation, ultimately improving overall generation quality.

To further demonstrate the effect of semantic guidance, we compare VAR-d16 models trained with and without semantic guidance across training epochs. As shown in Fig. 5, semantic guidance accelerates convergence, achieving up to a 2× speedup when reaching the same performance level as the baseline model. This validates that explicitly enhancing early-scale semantic modeling effectively improves both representation learning and training efficiency under the autoregressive framework.

5.4 Discussion about representation alignment

A key distinction is that VAR introduces residual-based, multi-scale feature space, which is absent in diffusion models studied by REPresentation Alignment (REPA) [62]. We analyze these VAR-specific challenges (feature gap between residual and complete features, selecting specific semantic granularity for alignment) and propose tailored solutions. To our knowledge, this is the first systematic study of representation alignment in the VAR paradigm, rather than a direct extension of REPA. Moreover, MEPA consists of STMoE and semantic guidance, where our contributions to analyzing and designing the MOE architecture are also critical.

As shown in the Tab. 7, the improvement of MEPA outperforms REPA significantly (21.36% vs. 12.62% gain), proving its suitability for VAR transformers. Although SiT+REPA shows lower FID, this is expected since the VAR has weaker generation quality but higher inference efficiency than SiT. MEPA significantly boosts VAR's generation quality to approach diffusion levels while retaining VAR's distinct advantage in inference speed. In terms of time cost, MEPA increases inference time by 0.7× and training time per epoch by 4.50%. This moderate overhead is offset by faster convergence and consistent quality gains, supporting our efficiency claim.

Table 7: Qualitative comparison with REPA. We conduct a qualitative comparison between MEPA and REPA. Param represents the number of parameters. Inference denotes the inference time per image, while Train indicates the training time of one epoch.

Model	Param	Epoch	IS↑	FID↓	Inference	Train
VAR-d20	600M	250	302.6	2.95	0.50s	1h51m
MEPA-d16	585M	100	299.8(**−0.93%**)	2.79(**−5.42%**)	0.85s	1h56m
MEPA-d16	585M	200	311.3(+2.88%)	2.32(**−21.36%**)	0.85s	1h56m
SiT-XL/2	675M	1400	270.3	2.06	45s	_
+REPA	675M	200	264.0(-2.33%)	1.96(-4.85%)	45s	_
+ REPA	675M	800	284.0(**+5.07%**)	1.80(-12.62%)	45s	_

Table 8: Qualitative comparison on class-conditional ImageNet 512×512..

Model	BigGAN	ADM	DiT-XL/2	MaskGIT	VQGAN	VAR-d36-s	MEPA-d20
FID↓	8.43	23.24	3.04	7.32	26.52	2.63	7.29
IS↑	177.9	101.0	240.8	156.0	66.8	303.2	199.2
Epoch	_	_	_	_	_	350	66

6 Limitations and Future Work

Due to limited computational resources, we were unable to conduct additional experiments within the current timeline. We report the results of MEPA-d20 trained at a

512×512 resolution for 66 epochs in Tab. 8. Even with this restricted training budget, MEPA already outperforms several existing methods despite not reaching state-of-the-art performance. It indicates the potential of our approach for further scaling up. For future work, we intend to finish the complete training pipeline under the 512×512 setting and explore novel architectures with higher input resolutions.

7 Conclusion

In this work, we address the inherent challenges of multi-scale representation learning in Visual AutoRegressive (VAR) modeling. We identify two key limitations: optimization conflicts caused by sharing a single architecture across scales, and semantic error propagation induced by the causal coarse-to-fine generation process. To overcome these issues, we propose multi-scale representation alignment (MEPA), a scale-aware token-routed MoE framework that promotes decoupled representation learning across scales and strengthens semantic information. Extensive experiments demonstrate that our method consistently improves both training efficiency and generation quality.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096, in part by Scientific and Technological Innovation Teams in Shaanxi Province under Grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under Grant 2024QY2-GJHX-11, in part by the Establishment Fund of the State Key Laboratory of Wakefullness/Sleep and Cognition(Chongqing Institute for Brain and Intelligence), in part by the Fundamental Research Funds for the Central Universities under Grant QTZX26138, in part by the Innovation Fund of Xidian University under Grant YJSJ26022.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alaparthi, S., Mishra, M.: Bidirectional encoder representations from transformers (bert): A sentiment analysis odyssey. arXiv preprint arXiv:2007.01127 (2020)
3. Bachmann, R., Allardice, J., Mizrahi, D., Fini, E., Kar, O.F., Amirloo, E., El-Nouby, A., Zamir, A., Dehghan, A.: Flextok: Resampling images into 1d token sequences of flexible length. In: Forty-second International Conference on Machine Learning (2025)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. 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)
6. Chen, H., Wang, Z., Li, X., Sun, X., Chen, F., Liu, J., Wang, J., Raj, B., Liu, Z., Barsoum, E.: Softvq-vae: Efficient 1-dimensional continuous tokenizer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28358–28370 (2025)

7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
8. Chen, Z., Ma, X., Fang, G., Wang, X.: Collaborative decoding makes visual auto-regressive modeling efficient. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23334–23344 (2025)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
11. 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)
12. Esteves, C., Suhail, M., Makadia, A.: Spectral image tokenizer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17181–17190 (2025)
13. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
14. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems **33**, 21271–21284 (2020)
15. Guo, H., Li, Y., Zhang, T., Wang, J., Dai, T., Xia, S.T., Benini, L.: Fastvar: Linear visual autoregressive modeling via cached token pruning. arXiv preprint arXiv:2503.23367 (2025)
16. Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., Liu, X.: Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15733–15744 (2025)
17. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
19. Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T.B., Dhariwal, P., Gray, S., et al.: Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701 (2020)
20. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
21. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. Journal of Machine Learning Research **23**(47), 1–33 (2022)
22. Hu, T., Zhang, J., Yi, R., Weng, J., Wang, Y., Zeng, X., Xue, Z., Ma, L.: Improving autoregressive visual generation with cluster-oriented token prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9351–9360 (2025)
23. Huang, Y., Chen, W., Zheng, W., Duan, Y., Zhou, J., Lu, J.: Spectralar: Spectral autoregressive visual generation. arXiv preprint arXiv:2506.10962 (2025)
24. Huang, Z., Qiu, X., Ma, Y., Zhou, Y., Chen, J., Zhang, H., Zhang, C., Li, X.: Nfig: Autoregressive image generation with next-frequency prediction. arXiv preprint arXiv:2503.07076 (2025)

25. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
26. Jia, M., Yin, W., Hu, X., Guo, J., Guo, X., Zhang, Q., Long, X.X., Tan, P.: Mgvq: Could vq-vae beat vae? a generalizable tokenizer with multi-group quantization. arXiv preprint arXiv:2507.07997 (2025)
27. Jiao, S., Zhang, G., Qian, Y., Huang, J., Zhao, Y., Shi, H., Ma, L., Wei, Y., Jie, Z.: Flexvar: Flexible visual autoregressive modeling without residual prediction. arXiv preprint arXiv:2502.20313 (2025)
28. Jin, P., Zhu, B., Yuan, L., Yan, S.: Moe++: Accelerating mixture-of-experts methods with zero-computation experts. arXiv preprint arXiv:2410.07348 (2024)
29. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10124–10134 (2023)
30. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
31. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
32. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11523–11532 (2022)
33. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
34. Li, T., Katabi, D., He, K.: Self-conditioned image generation via generating representations. CoRR (2023)
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. Li, X., Qiu, K., Chen, H., Kuen, J., Gu, J., Raj, B., Lin, Z.: Imagefolder: Autoregressive image generation with folded tokens. arXiv preprint arXiv:2410.01756 (2024)
37. Liu, W., Zhuo, L., Xin, Y., Xia, S., Gao, P., Yue, X.: Customize your visual autoregressive recipe with set autoregressive modeling. arXiv preprint arXiv:2410.10511 (2024)
38. 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)
39. Pang, Y., Jin, P., Yang, S., Lin, B., Zhu, B., Tang, Z., Chen, L., Tay, F.E., Lim, S.N., Yang, H., et al.: Next patch prediction for autoregressive visual generation. arXiv preprint arXiv:2412.15321 (2024)
40. 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)
41. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
42. 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)
43. 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. pp. 10684–10695 (2022)
44. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)

45. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
46. 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)
47. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv preprint arXiv:2405.09818 (2024)
48. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
49. 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)
50. 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)
51. Van Den Oord, A., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. In: International conference on machine learning. pp. 1747–1756. PMLR (2016)
52. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)
53. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
54. 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)
55. Weber, M., Yu, L., Yu, Q., Deng, X., Shen, X., Cremers, D., Chen, L.C.: Maskbit: Embedding-free image generation via bit tokens. arXiv preprint arXiv:2409.16211 (2024)
56. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. arXiv preprint arXiv:2408.12528 (2024)
57. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022)
58. Yao, J., Yang, B., Wang, X.: Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15703–15712 (2025)
59. Yu, H., Luo, H., Yuan, H., Rong, Y., Zhao, F.: Frequency autoregressive image generation with continuous tokens. arXiv preprint arXiv:2503.05305 (2025)
60. 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)
61. 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)
62. 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)
63. Zhang, J., Long, W., Han, M., You, W., Gu, S.: Mvar: Visual autoregressive modeling with scale and spatial markovian conditioning. arXiv preprint arXiv:2505.12742 (2025)

64. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems **35**, 7103–7114 (2022)
65. Zhuang, X., Xie, Y., Deng, Y., Liang, L., Ru, J., Yin, Y., Zou, Y.: Vargpt: Unified understanding and generation in a visual autoregressive multimodal large language model. arXiv preprint arXiv:2501.12327 (2025)
66. Zuo, S., Liu, X., Jiao, J., Kim, Y.J., Hassan, H., Zhang, R., Zhao, T., Gao, J.: Taming sparsely activated transformer with stochastic experts. arXiv preprint arXiv:2110.04260 (2021)