中国科学院大学
University of Chinese Academy of Sciences
ISCAS

Autoregression with Self-Token Prediction

Dengsheng Chen, Yangming Shi, Enhua Wu

Core idea: predict multiple spatially causal tokens per AR step

Model: SAR, a spatially autoregressive image generator

Key outcome: strong ImageNet quality with far fewer AR steps

Motivation: next-token AR is a bottleneck for images/videos

LLM-style next-token prediction is simple and scalable, but dense visual modalities create very long token sequences.

Serializing 2D images into a rigid 1D raster order is unnatural and slow.

Prior causal AR models often lag diffusion or masked-prediction alternatives in latency and quality.

Self-token prediction: causal generation by groups

Partition tokens into ordered groups $G_1 \dots G_T$.

At step t, all tokens in G_t are predicted in parallel.

The group order defines causal dependencies: G_t can depend only on previous groups $G_{<t}$.

Classical next-token prediction is recovered when each group contains one token.

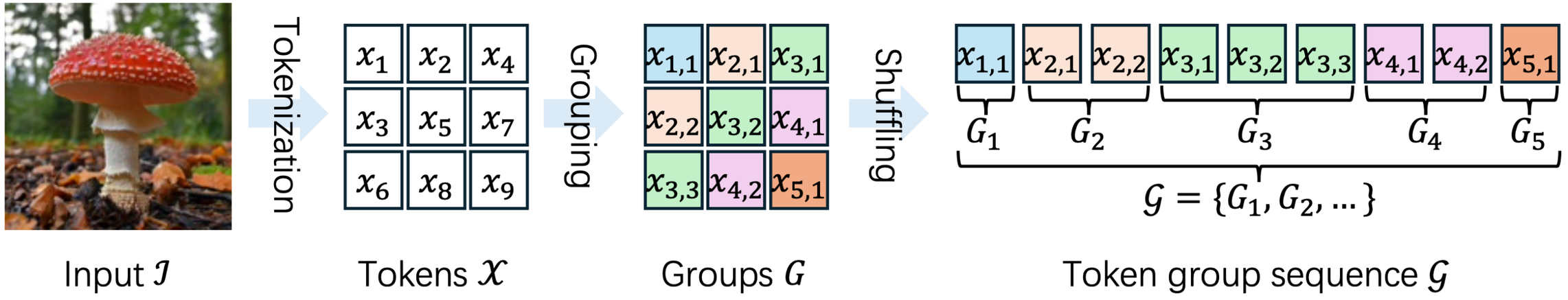

Figure 1. **Token grouping pipeline.** An input token sequence $\mathcal{X}$ is partitioned into an ordered set of groups $\mathcal{G} = \{G_t\}_{t=1}^{T}$, where each group G_t is generated in parallel within a single autoregressive step, and the group order defines the causal dependency across steps.

Training trick: context stream + generation stream

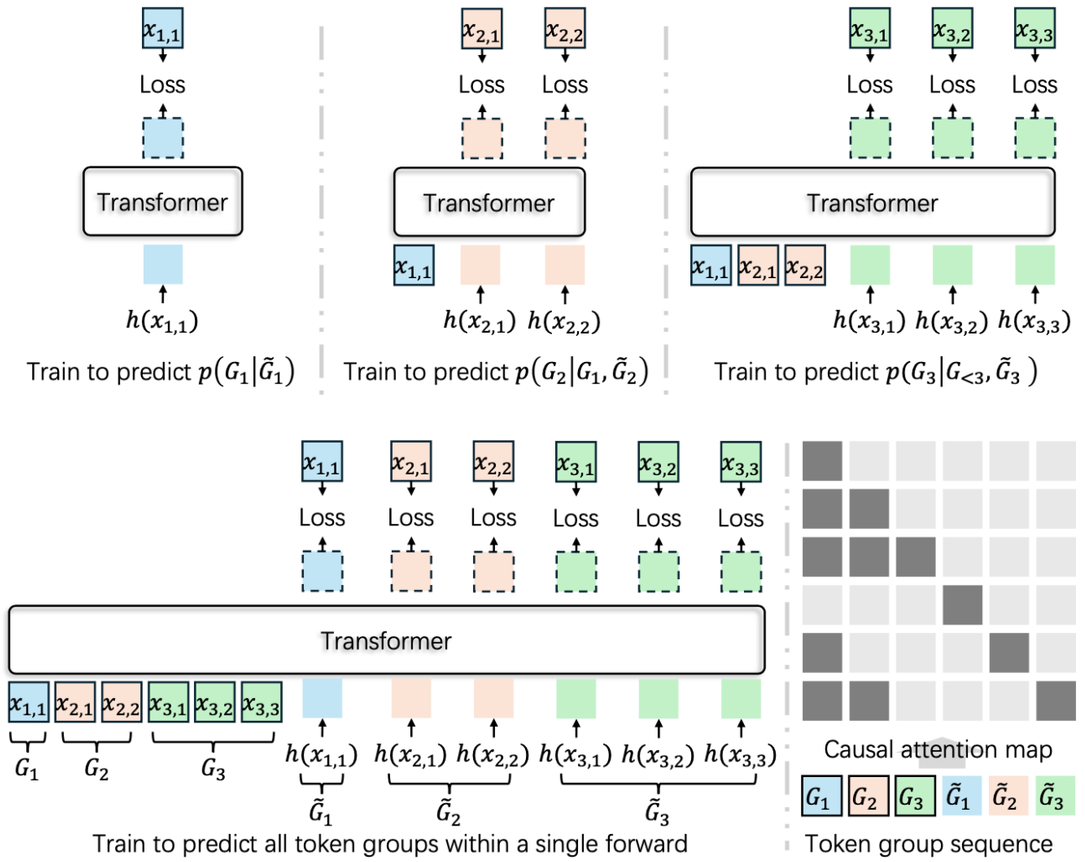

Figure 2. **Naïve vs. efficient training pipelines. Top:** a naive implementation predicts one token group per forward pass, which prevents parallel training across groups. **Bottom:** our efficient pipeline predicts all groups in a single forward pass by duplicating each step into a teacher-forced *context* stream and a supervised *generation* stream, with a structured group-causal attention mask that enforces the intended dependencies.

Naïve training would require one forward pass per group.

The paper duplicates each group into: a context copy using ground-truth tokens, and a generation copy using constructed causal input.

A structured group-causal attention mask trains all groups in one forward pass without leaking current/future ground truth.

Spatial dependency injection matters

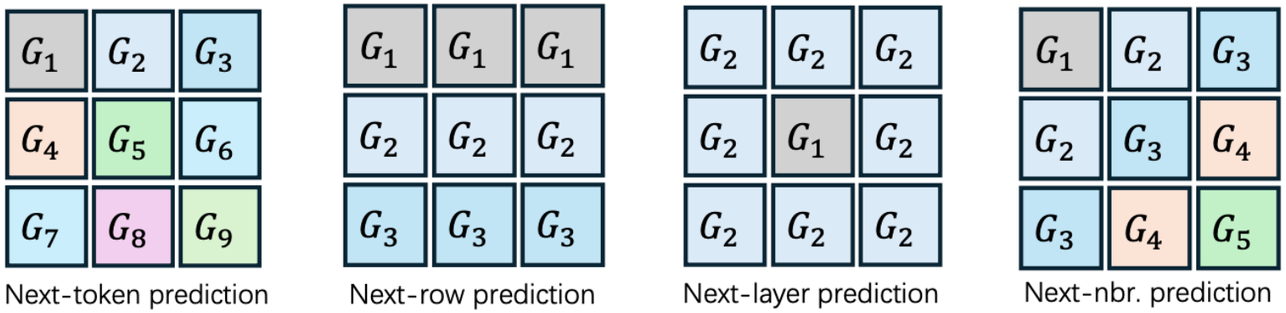

Figure 8. **Grouping strategy.** Note that next-token prediction 8a and next-row prediction 8b satisfy the shift-alignment condition in (12), and therefore require no additional training overhead (no context/generation duplication) and no inference-time cache reconciliation.

Method Configuration		Efficiency & Cost		Generation Quality	
Grouping Strategy	$h(\cdot)$	**#Steps↓**	**Train Cost↓**	**FID50k↓**	**Latency↓**
Next-token (Fig. 8a)	G_{t-1}	$h \times w$	1×	4.92	622ms
Next-row (Fig. 8b)	G_{t-1}	h	1×	22.78	66ms (9.4×)
Next-layer (Fig. 8c)	$\text{Agg}_{\text{NN}}(\cdot)$	$\frac{\max(h,w)}{2}$	2×	29.33	45ms (13.8×)
Next-nbr. (Fig. 8d)	$\text{Agg}_{\text{SE}}(\cdot)$	$h+w-1$	2×	58.49	106ms (5.8×)
Next-nbr. (Fig. 8d)	$\text{Agg}_{\text{AVG}}(\cdot)$	$h+w-1$	2×	**4.40**	106ms (5.8×)

SAR architecture and performance on different space

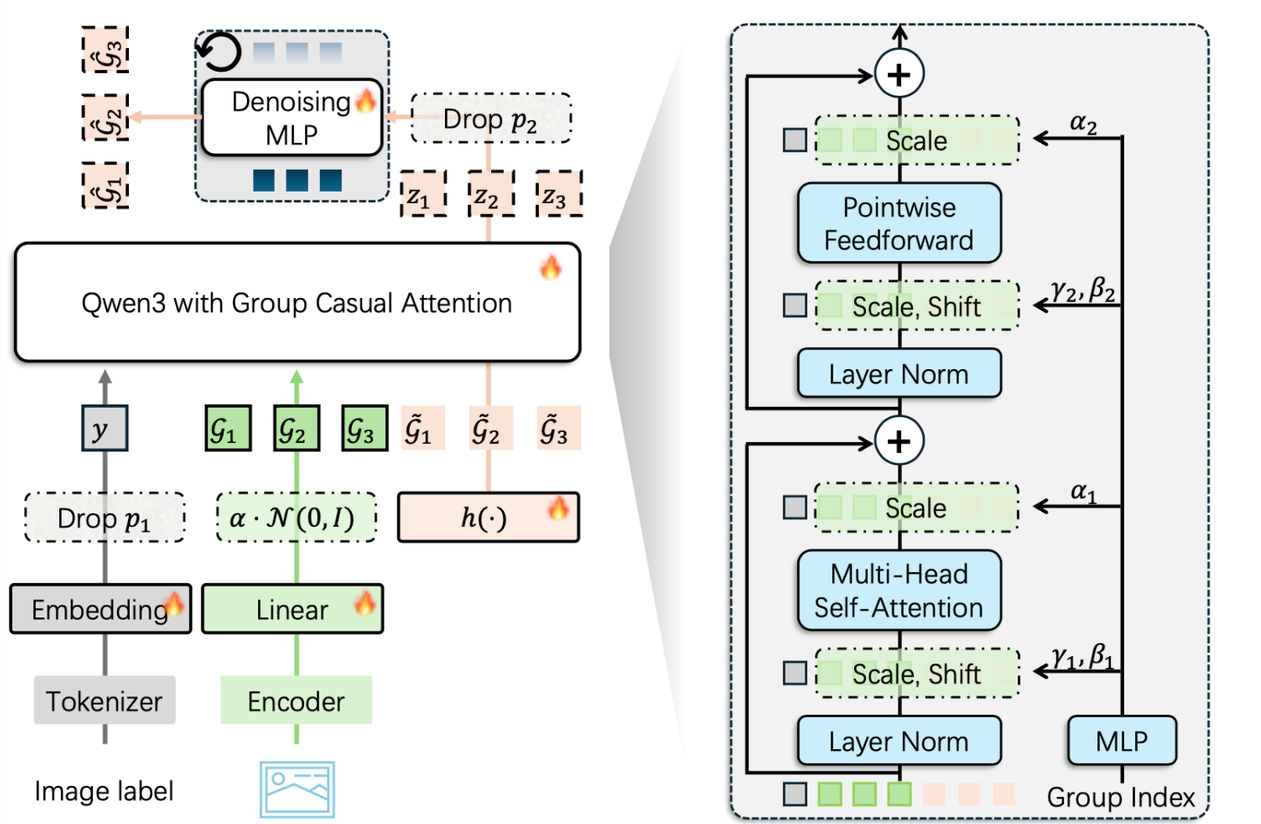

Figure 4. **Overall architecture of SAR.**

Guidance under causal context accumulation

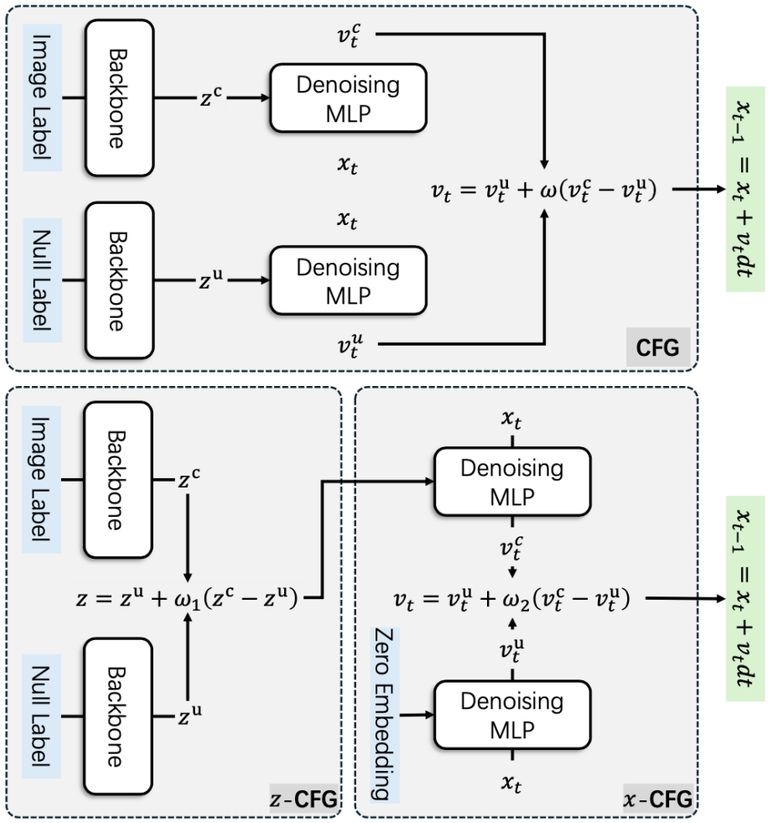

Figure 12. **Factorizing standard classifier-free guidance.** Under v-prediction, we decompose conventional CFG into two complementary forms: z-**CFG**, which applies guidance by interpolating backbone features, and x-**CFG**, which applies guidance at the head output by constructing an unconditional branch with $z = 0$.

Table 2. **Guidance ablation under spatially causal decoding.** We compare standard CFG, x-CFG, and z-CFG in terms of computation, latency, and sample quality. Guidance is essential for SAR: without guidance, performance is poor (FID 39.68), while standard CFG at scale 3.50 reaches FID 4.40. x-CFG provides a strong quality gain with only a minor latency increase, and combining x-CFG with z-CFG achieves the best FID (3.56) under the same latency as standard CFG.

CFG Configuration			Efficiency & Cost		Generation Quality	
CFG	x-CFG	z-CFG	Computation	Latency↓	FID50k↓	IS↑
	no-CFG		$C + c$	73 ms	39.68	50.09
-	1.25	-	$C + 2 \times c$	81 ms	18.53	77.98
3.50	-	-	$2 \times C + 2 \times c$	100 ms	4.40	**307.34**
-	-	3.50	$2 \times C + c$	95 ms	6.90	287.16
-	1.25	2.70	$2 \times C + 2 \times c$	100 ms	**3.56**	275.13

Representation universality and robustness

Table 3. **Representation universality across continuous image spaces.** SAR is evaluated on raw RGB patches and VAE latent spaces, including joint image–video latents (WanX (2025)) and KL-VAE latents (2022), under a fixed 16×16 effective token grid for 256×256 images. We report FID50k/IS and compare guidance variants. **S&P** denotes the VAE stride (if applicable) and the patch size. Flow matching remains effective across spaces and consistently outperforms a diffusion-style (DDPM) objective under comparable settings.

Method Configuration			CFG Configuration			Generation Quality	
Denoising Head	**Denoising Space**	**S&P**	**CFG**	**x-CFG**	**z-CFG**	**FID50k↓**	**IS↑**
FM (2022)	Raw (2025)	(-,16)	- 6.50	1.50 -	4.50 -	8.36 8.42	240.36 319.88
FM (2022)	Joint-IV (2025)	(8,2)	- 4.50	1.15 -	3.50 -	4.84 5.31	314.32 342.39
FM (2022)	Latent (2022)	(16,1)	- 3.50	1.25 -	2.70 -	3.56 4.40	275.13 307.34
DDPM (2021)	Latent (2022)	(16,1)	- 12.50	1.80 -	3.90 -	29.48 16.85	140.33 301.25

Flow matching works across raw RGB patches, joint image-video latents, and KL-VAE latents.

Best reported SAR-L ablation in KL-VAE latent space: FID 3.56 with x-CFG + z-CFG.

Raw-pixel generation remains competitive, showing the method is not tied to a specific tokenizer.

A DDPM-style objective is much worse than flow matching in this framework.

Main ImageNet results: quality and scaling

Table 4. **Comparison with state-of-the-art models on ImageNet-1K at** 256×256**.** We report FID-50K (lower is better) and Inception Score (IS; higher is better). **Spatial Causality** indicates whether the model enforces spatially causal modeling (*e.g.*, disallowing full attention). Notably, although VAR (Tian et al., 2024) is autoregressive, it uses full attention within each resolution stage and thus does not satisfy strict spatial causality.

Model	Spatial Causality	#Params	#Steps↓	FID50k↓	IS↑
GAN					
BigGAN (Brock, 2018)	✗	112M	1	6.95	224.5
GigaGAN (Kang et al., 2023)	✗	569M	1	3.45	225.5
StyleGAN-XL (Sauer et al., 2022)	✗	166M	1	2.30	265.1
Diffusion					
LDM-4-G (Rombach et al., 2022)	✗	400M	250	3.60	247.7
DiT-XL/2 (Peebles & Xie, 2023)	✗	675M	250	2.27	278.2
JiT-L/16 (Li & He, 2025)	✗	459M	50	2.36	298.5
ADM (Dhariwal & Nichol, 2021)	✗	554M	250	10.94	101.0
VDM++ (Kingma & Gao, 2023)	✗	2.0B	512	2.12	267.7
Mask Prediction					
MAGVIT-v2 (Yu et al., 2023)	✗	307M	64	1.78	319.4
MaskGIT (Chang et al., 2022)	✗	227M	8	6.18	182.1
MAR-L (Li et al., 2024)	✗	479M	256	1.78	296.0
Autoregressive (Next-Token/Scale Prediction)					
VAR-d20 (Tian et al., 2024)	✗	600M	10	2.57	302.6
VQGAN (Esser et al., 2021)	✓	1.4B	256	15.78	74.3
ViT-VQGAN (Yu et al., 2021)	✓	1.7B	1024	4.17	175.1
RQTran. (Lee et al., 2022)	✓	3.8B	68	7.55	134.0
LlamaGen-XXL (Sun et al., 2024a)	✓	1.4B	256	2.34	253.9
Autoregressive (Self-Token Prediction)					
SAR-B (ω_x=1.35, ω_z=3.0)	✓	141M	31	4.23	231.1
SAR-L (ω_x=1.25, ω_z=2.7)	✓	714M	31	2.52	288.5
SAR-H (ω_x=1.20, ω_z=2.5)	✓	1.3B	31	2.26	290.0

Dataset: ImageNet-1K at 256×256; evaluation uses 50K generated samples.

SAR-B: 141M params, 31 steps, FID 4.23.

SAR-L: 714M params, 31 steps, FID 2.52.

SAR-H: 1.3B params, 31 steps, FID 2.26, comparable to DiT-XL/2 FID 2.27 and better than LlamaGen-XXL FID 2.34.

Qualitative samples and visual fidelity

Figure 6. **Curated samples from SAR-L on ImageNet-1K (256 × 256).** Images are generated with strictly spatially causal group-wise autoregressive decoding, demonstrating high fidelity and diversity under the proposed self-token prediction paradigm.

Limitations and open questions

- Designing token grouping and dependency injection $h(\cdot)$ remains modality-dependent and empirical.
- Experiments focus mainly on ImageNet image generation; generalization to editing, audio, video, and multimodal fusion needs further study.
- The broader hypothesis—that strict spatial causality is essential for later-stage capability emergence—remains open.
- Potential misuse risks include photorealistic deceptive content and inherited data bias.

Takeaways

- Self-token prediction decouples causality from one-token decoding.
- SAR demonstrates that strictly causal visual AR can be both efficient and high quality.
- The decisive design choice is structured spatial dependency injection, not just larger token groups.
- The method opens a path toward unified multimodal AR models that retain causal learning signals while reducing dense-modality latency.