DPAR: Dynamic Patchification for Efficient Autoregressive Visual Generation

Divyansh Srivastava[1], Akshay Mehra[2,*], Pranav Maneriker[2,*], Debopam Sanyal[3], Vishnu Raj[2], Vijay Kamarshi[2], Fan Du[2], Joshua Kimbal[2]

[1]University of California San Diego, [2]Dolby Laboratories, [3]Georgia Tech, [*]Equal Contribution

Code

CVPR JUNE 3-7, 2026 DENVER COLORADO

Background

High-Resolution Autoregressive (AR) Image Generation with Transformer backbone suffers from two major challenges:

- Quadratic complexity of Attention
- Quadratic increase in token count with increasing resolution: HD images require 4,024 tokens and 4K images need 65K tokens for generation!!

While prior works[1,2,3] tried to alleviate computational demands through architectural improvements, these approaches suffer from efficiency-performance tradeoffs. In contrast, direct strategies aimed at reducing token count remain underexplored.

Motivation & Approach

We hypothesize that images can be dynamically represented using a variable number of tokens based on their information content, motivated by two observations:

- Images often contain redundant low-information regions, such as sky or background, which can be represented with fewer tokens.
- Low-detail image regions should be generated with lower computational cost — an idea recently explored in language models like BLT[4].

Our method estimates next-token prediction entropy using a lightweight, unsupervised AR model as a proxy for local image information content, then merges tokens into patches for efficient transformer-based AR generation.

Method

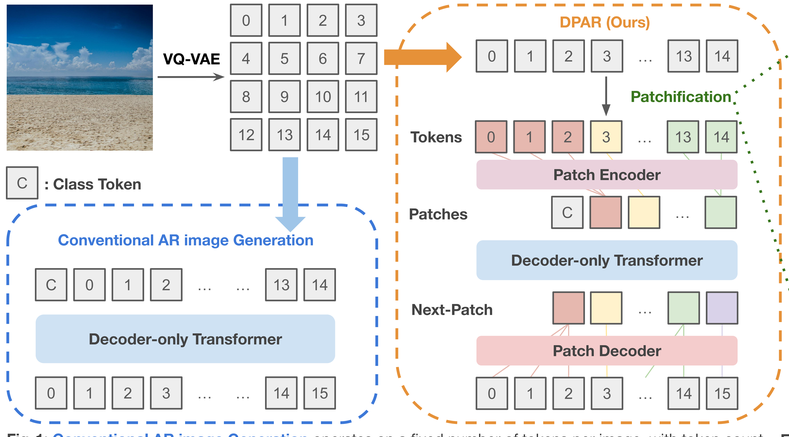


Fig-1: **Conventional AR image Generation** operates on a fixed number of tokens per image, with token count increasing quadratically with image resolution. (b) **DPAR (Ours)** dynamically aggregates tokens based on information content, generating reduced and variable number of patches for efficient AR generation.

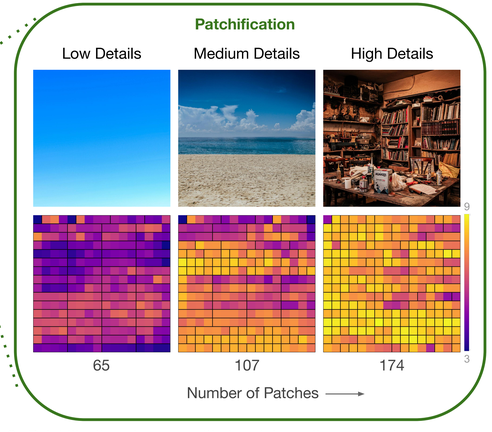


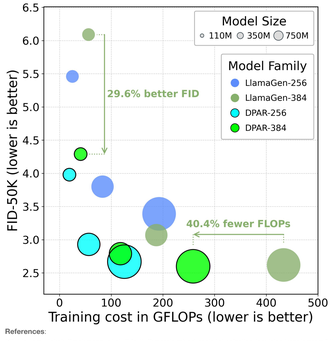
Fig-2: Patchification: Images with low information content, such as homogeneous regions, produce fewer high-entropy tokens, allowing the model to merge them into larger patches.

Qualitative Samples

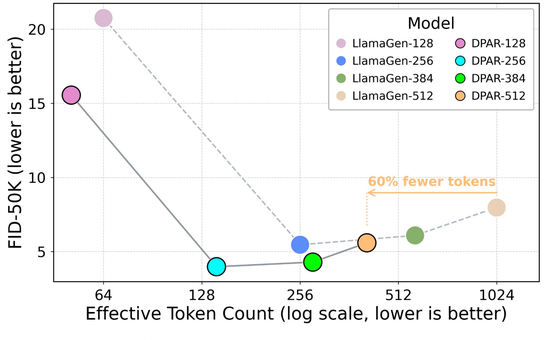
Fig-3: Selected samples from our class-conditional DPAR-384-XL model trained on ImageNet.

Key Results

1 Efficient Model Scaling



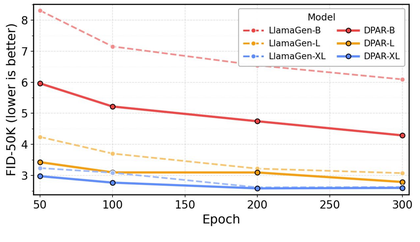
2 Efficient Resolution Scaling



3 Improved Generation Quality

Model	#Params	FID↓	IS↑	Prec.↑	Rec.↑	Steps
LlamaGen-B	111M	5.46	193.61	0.84	0.46	256
DPAR-B (cfg=2.1)	120M	3.98	250.62	0.83	0.49	142
LlamaGen-L	343M	3.80	248.30	0.83	0.52	256
DPAR-L (cfg=1.9)	352M	2.93	269.34	0.81	0.56	142
LlamaGen-XL	775M	3.39	227.10	0.81	0.54	256
DPAR-XL (cfg=2.0)	789M	2.67	281.65	0.82	0.56	142
LlamaGen-384-B	111M	6.09	182.53	0.84	0.42	576
DPAR-384-B (cfg=2.10)	120M	4.29	254.54	0.83	0.47	280
LlamaGen-384-L	343M	3.07	256.06	0.83	0.52	576
DPAR-384-L (cfg=1.90)	352M	2.79	283.84	0.81	0.55	280
LlamaGen-384-XL	775M	2.62	244.08	0.80	0.57	576
DPAR-384-XL (cfg=1.90)	789M	2.60	285.43	0.81	0.57	280
LlamaGen-384-XXL	1.4B	2.34	253.91	0.81	0.60	576
DPAR-384-XXL (cfg=1.75)	1.4B	2.30	287.37	0.82	0.59	280

4 Faster Training Convergence



5 Adaptive Patching at Inference

$\mathcal{H}_{\text{Th}}^{\text{inf}}$	7.8	7.9	8.0	8.1
P_{avg}	1.81	1.92	2.03	2.16
Static	3.58	7.85	17.91	25.59
DPAR-L	3.31	3.32	3.38	3.39

Discussion

Our work is the first to show that entropy-based patching can dynamically merge 2D image tokens based on information content, reducing sequence length while preserving the spatial structure and controllability of 2D token grids. While our results establish the promise of dynamic patchification for efficient AR generation, several directions remain open for future work:

- Flexible patching: Extend beyond row-limited merging to irregular and variable-shaped patches.
- Positional embeddings: Design better position encodings to support dynamic patch layouts.
- Efficient inference: Develop KV-Cache mechanisms for AR decoding with dynamically sized patches.
- Extension to specialized AR training strategies, such as Random-order Generation[5].
- Scaling Laws for dynamic patch DPAR architecture.
- Large-scale text-to-image generation.

References:

[1]. Hiervtorn, Jakob Drachmann, et al. "MeVit: Dynamic mixed-scale tokenization for vision transformers." ICCV 2023. **[2].** Shen, Junhong, et al. "Cat: Content-adaptive image tokenization." arXiv preprint arXiv:2501.03120 (2025).

[3]. Ma, Xu, e t al. "Token-shuffle: Towards high-resolution image generation with autoregressive models." arXiv preprint arXiv:2504.17789 (2025). **[4].** Pagnoni, Artidoro, et al. "Byte latent transformer: Patches scale better than tokens." arXiv preprint arXiv:2412.09871 (2024).

[5] Pang, Ziqi, et al. "Randar: Decoder-only autoregressive visual generation in random orders." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.