Parallelized Autoregressive Visual Generation

Supplementary Material

Appendix

The supplementary material includes the following additional information:

- Sec. A provides more implementation details for PAR.
- Sec. B demonstrates the compatibility of our approach with typical LLM engineering optimizations.
- Sec. C provides more visualization results, including zero-shot high-resolution generation and long-range dependency examples.
- Sec. D provides the analysis of visual token dependencies.

A. Implementation details for PAR

Image Generation. For image generation, we train our models on the ImageNet-1K [10] training set, consisting of 1,281,167 images across 1,000 object classes. Following the setting in LlamaGen [51], we pre-tokenize the entire training set using their VQGAN [12] tokenizer and enhance data diversity through ten-crop transformation. For inference, we adopt classifier-free guidance [16] to improve generation quality. The detailed training and sampling hyper-parameters are listed in Tab. 5.

config	value
training hyper-params	
optimizer	AdamW [30]
learning rate	1e-4(L,XL)/2e-4(XXL,3B)
weight decay	5e-2
optimizer momentum	(0.9, 0.95)
batch size	256(L,XL)/ 512(XXL,3B)
learning rate schedule	cosine decay
ending learning rate	0
total epochs	300
warmup epochs	15
precision	bfloat16
max grad norm	1.0
dropout rate	0.1
attn dropout rate	0.1
class label dropout rate	0.1
sampling hyper-params	
temperature	1.0
guidance scale	1.60 (L) / 1.50 (XL) / 1.435 (XXL) / 1.345 (3B)

Table 5. **Detailed Hyper-parameters for Image Generation.**

Video Generation. For video generation, we train our models on the UCF-101 [48] training set, which contains 9.5K training videos spanning 101 action categories. Videos are processed as 8fps random clips and tokenized by our reimplementation of MAGVIT-v2 [71] (as their code is not publicly available), achieving a reconstruction FVD score of 32 on UCF-101. For inference, we use classifier-free guidance [16] with top-k sampling to improve generation quality. The detailed training and sampling hyper-parameters are listed in Tab. 6.

config	value
training hyper-params	
optimizer	AdamW [30]
learning rate	1e-4
weight decay	5e-2
optimizer momentum	(0.9, 0.95)
batch size	256
learning rate schedule	cosine decay
ending learning rate	0
total epochs	3000
warmup epochs	150
precision	bfloat16
max grad norm	1.0
dropout rate	0.1
attn dropout rate	0.1
class label dropout rate	0.1
sampling hyper-params	
temperature	1.0
guidance scale	1.15
top-k	8000

Table 6. **Detailed Hyper-parameters for Video Generation.**

B. Compatibility with Typical LLM Engineering Optimizations

We investigate whether our algorithmic parallel generation approach can complement typical engineering optimizations used in LLM inference. All experiments were conducted on a single NVIDIA A100 GPU with batch size 1, generating 384×384 resolution images. For simplicity, we only implemented PyTorch's compile feature (leveraging CUDA graph optimization) in our PAR model. As a comparison point, we tested LlamaGen [51] with vLLM [23] optimizations, which includes both page attention and CUDA graph optimizations.

As shown in Tab. 7, our algorithmic improvements and engineering optimizations are orthogonal and provide com-

Model	Resolution	Optimization	Latency
LlamaGen-3B	384	none	12.41s
LlamaGen-3B	384	vLLM	4.12s
PAR-3B-4x	384	none	3.46s
PAR-3B-4x	384	compile	1.15s
PAR-3B-16x	384	compile	0.43s

Table 7. **Compatibility with LLM engineering optimizations.** Even with just PyTorch compile optimization, our PAR approach achieves substantial speedups compared to LlamaGen with more comprehensive vLLM optimizations.

Figure 6. **Zero-shot generation at 512×512 resolution.** Our model successfully generates coherent high-resolution images despite being trained at 384×384 resolution.

plementary benefits. Even without any engineering optimization, PAR-3B-4x (3.46s) outperforms LlamaGen-3B with comprehensive vLLM optimizations (4.12s). When implementing just the simple CUDA graph optimization through PyTorch compile, PAR-3B-4x achieves 1.15s latency, a 3.6× improvement over optimized LlamaGen. With more aggressive parallelization, PAR-3B-16x with compile further reduces latency to 0.43s, demonstrating our approach's flexibility in speed-quality trade-offs. These results confirm that algorithm-level optimizations (reducing sequential steps) and engineering-level optimizations (improving computational efficiency) are orthogonal approaches that, when combined, maximize generation efficiency beyond what either can achieve alone.

C. More Visualization Results

Zero-shot Generation on Higher Resolutions. Fig. 6 demonstrates our model's capability for zero-shot generation at higher resolutions (512×512) using Rotary Position Embedding [50]. Despite being trained on 384×384 images, our approach effectively maintains coherent global structures and detailed local patterns in higher resolution generation. This shows the flexibility of our parallel generation framework and its compatibility with positional encoding methods that support resolution extrapolation.

Long-range Dependency Handling While our approach leverages the observation that spatially distant tokens typically have weaker dependencies, certain visual elements exhibit strong long-range dependencies. Fig. 7 showcases our model's ability to maintain consistency between distant but strongly dependent visual elements, such as symmetric features (deer antlers, vehicle wheels) and coherent structures across the image.

Figure 7. **Long-range dependency handling.** Our method successfully maintains consistency between distant but strongly related elements (highlighted regions), even when generating tokens from different spatial regions in parallel.

Addtional Image Generation Visualization. In Fig.8 and Fig.9, we provide additional visualization results of PAR-4× and PAR-16× image generation on ImageNet [10] dataset, respectively.

Addtional Video Generation Visualization. In Fig.10, we provide the visualization results of video generation using our model on the UCF-101[48] dataset. The results are sampled from 128×128 resolution videos with 17 frames. As shown in the figure, even with 16× parallelization (PAR-16×), our method shows no obvious quality degradation compared to single-token prediction (PAR-1×), producing smooth motion and stable backgrounds across frames.

D. Analysis of Visual Token Dependencies

In Sec.3.1, we demonstrated through pilot studies that parallel generation of adjacent tokens leads to quality degradation due to strong dependencies, while tokens from distant regions can be generated simultaneously. In this section, we provide a theoretical perspective of conditional entropy to explain this observation and our design. We use conditional entropy to measure the token dependencies quantitatively - lower conditional entropy between tokens indicates stronger dependency, while higher conditional entropy suggests weaker dependency and thus potential for parallel generation. We further validate our PAR design from the perspective of conditional entropy - In AR-based generation, each step predicts a conditional distribution of the next tokens given all previous tokens. Higher conditional entropy indicates higher difficulty for the model to predict the next tokens. In this section, we first introduce the estimation of conditional entropy in Sec.D.1, and then validate our proposed approach by analyzing the relationship between to-

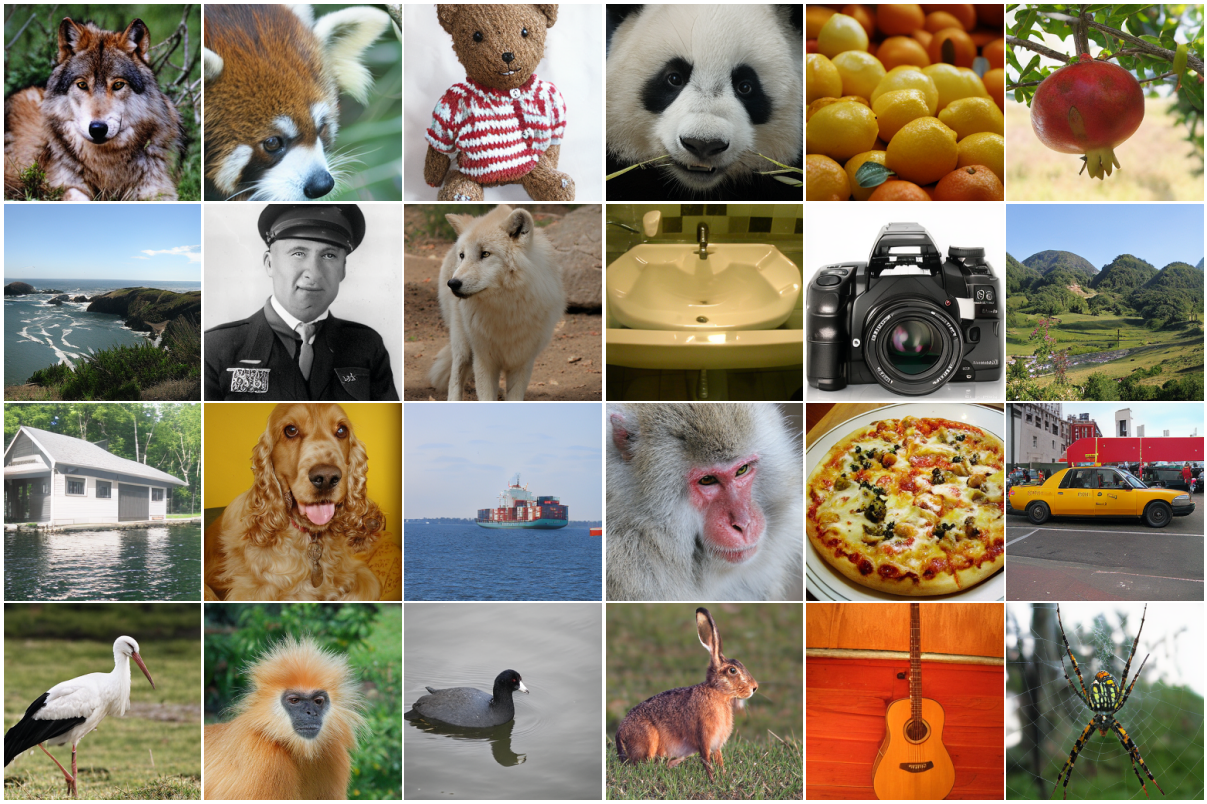
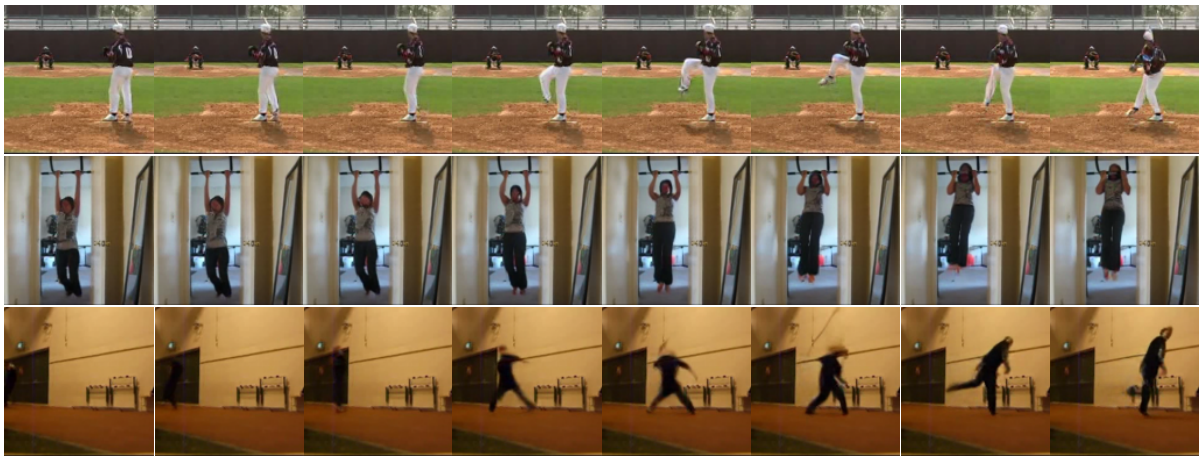


Figure 8. **Additional image generation results of PAR-4× across different ImageNet [10] categories.**

Figure 9. **Additional image generation results of PAR-16× across different ImageNet [10] categories.**

PAR-1x



PAR-4x

PAR-16x

Figure 10. **Video generation results on UCF-101 [48].** Each row shows sampled frames from a 17-frame sequence at 128×128 resolution, generated by PAR-1×, PAR-4×, and PAR-16× respectively across different action categories.

ken dependencies and spatial distances in Sec. D.2.

D.1. Conditional Entropy Estimation

Given a visual token sequence $\{v_1, v_2, ..., v_n\}$, our goal is to estimate the conditional entropy $H(v_k|\{v_j\}_{j<k})$ where the token feature $v_i \in \mathbb{R}^d$ and $\{v_j\}_{j<k}$ is the set of (all) visual tokens that precede v_k in the sequence. This conditional entropy measures the uncertainty of the current token v_k given the previously occurring visual tokens, thereby characterizing the dependency between v_k and the set $\{v_j\}_{j<k}$. It is important to emphasize that we do not require the exact value of $H(v_k|\{v_j\}_{j<k})$. Instead, we aim to reflect the trends in $H(v_k|\{v_j\}_{j<k})$ under different scenarios, such as given different sets of $\{v_j\}_{j<k}$ and considering different positions of v_k given the same set of $\{v_j\}_{j<k}$.

In particular, we characterize the relationship between the token v_k and the previous ones as the following model

$$v_k = f(\{v_j\}_{j<k}) + \boldsymbol{\epsilon}_k \tag{4}$$

where v_k is the next token we focus on and $\{v_j\}_{j<k}$ is the conditioning token(s), $f(\cdot)$ is a deterministic function, and $\boldsymbol{\epsilon}_k$ is the random additive error term. Then the conditional entropy $H(v_k|\{v_j\}_{j<k})$ satisfies

$$\begin{aligned} H(v_k|\{v_j\}_{j<k}) &= H(f(\{v_j\}_{j<k}) + \boldsymbol{\epsilon}_k|\{v_j\}_{j<k}) \\ &= H(\boldsymbol{\epsilon}_k|\{v_j\}_{j<k}), \end{aligned} \tag{5}$$

where the second equation holds since $f(\cdot)$ is a deterministic function. However, exactly calculating $H(\boldsymbol{\epsilon}_k|\{v_j\}_{j<k})$ is intractable as we cannot access the entire data distribution. To this end, inspired by prior research on bounding techniques for entropy and mutual information estimation [1, 2, 32, 33, 54, 62], we seek their upper bound as a proxy for showing the trends of the conditional entropy for different tokens. In particular, we have

$$H(\boldsymbol{\epsilon}_k|\{v_j\}_{j<k}) \leq H(\boldsymbol{\epsilon}_k) \leq \frac{1}{2}\log((2\pi e)^d|\boldsymbol{\Sigma}|), \tag{6}$$

where $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\boldsymbol{\epsilon}_k$. Notably, the first inequality naturally holds and the second inequality follows from the maximum entropy theory [9, 19], which is achievable when $\boldsymbol{\epsilon}_k$ follows a Gaussian distribution.

Based on Eq. 6, we can estimate the trend of conditional entropy changes by calculating the determinant of the residual covariance matrix, i.e., $|\boldsymbol{\Sigma}|$. In order to obtain the additive errors $\boldsymbol{\epsilon}$, we consider training a parameterized model $f_\theta(\cdot)$ to get the function f and characterize $\boldsymbol{\epsilon}$ as the residual errors. The detailed algorithm is provided in Algorithm 1.

D.2. Entropy Analysis on ImageNet Data and PAR

Based on the conditional entropy estimation method introduced above, we conduct experiments on ImageNet to analyze token dependencies and validate our parallel generation strategy. We randomly sample 10,000 images from ImageNet [10] and extract their features using VQGAN [12] encoder, followed by vector quantization to obtain continuous features from the codebook.

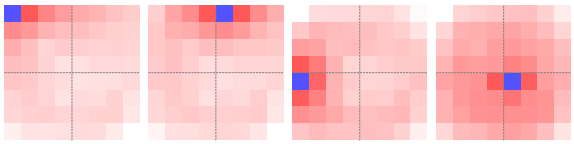


Figure 11. Visualization of token conditional entropy maps. Each map shows the conditional entropy of all tokens when conditioned on a reference token (blue square). Darker red indicates lower conditional entropy and thus stronger dependency with the reference token. The visualization shows that **tokens exhibit strong dependencies with their spatial neighbors and weak dependencies with distant regions.**

Algorithm 1 Conditional Entropy Estimation

Input:
m: number of data points
$\{v_{i,1}, v_{i,2}, ..., v_{i,n}\}_{i=1}^m$: visual token sequences, where each $v_{i,j} \in \mathbb{R}^d$
k: index of the target token
f_θ: parameterized model
Output: Estimated conditional entropy $\hat{H}(v_k|\{v_j\}_{j<k})$
Initialize empty lists $\mathcal{X}$ and $\mathcal{Y}$
for $i = 1$ to m **do**
  $X_i \leftarrow \{v_{i,j}\}_{j<k}$
  $Y_i \leftarrow v_{i,k}$
  Append (X_i, Y_i) to $(\mathcal{X}, \mathcal{Y})$
end for
Train a model f_θ to estimate Y given X using $(\mathcal{X}, \mathcal{Y})$
Initialize empty list $\mathcal{E}$ for residuals
for (X, Y) in $(\mathcal{X}, \mathcal{Y})$ **do**
  $Y_{pred} \leftarrow f_\theta(X)$
  $\boldsymbol{\epsilon}_k \leftarrow Y - Y_{pred}$
  Append $\boldsymbol{\epsilon}_k$ to $\mathcal{E}_k$
end for
Compute residual covariance matrix $\hat{\boldsymbol{\Sigma}}$ of $\mathcal{E}_k$
$\hat{H}(v_k|\{v_j\}_{j<k}) \leftarrow \frac{1}{2}\log((2\pi e)^d|\hat{\boldsymbol{\Sigma}}|)$
return $\hat{H}(v_k|\{v_j\}_{j<k})$

We first analyze the dependencies between tokens at different positions. For each position j in the feature map, we calculate the conditional entropy $H(v_i|v_j)$ where $i \neq j$, given the token v_j at the j-th position and considering all tokens v_i at other positions. It should be noted that Algorithm 1 is not limited to $H(v_k|\{v_j\}_{j<k})$ where the given visual tokens $\{v_j\}$ must satisfy $j < k$. This is because any given tokens v_j and v_i can be considered to satisfy Eq. 4, making the proposed method applicable in calculating $H(v_i|v_j)$. Fig. 11 presents the experimental results. We observe that given different token positions v_i, the adjacent tokens typically exhibit lower conditional entropy (shown in redder

Figure 12. **Conditional entropy differences between parallel and sequential generation in different orders.** (a)(d) show parallel (4 tokens) generation strategies and (b)(e) show sequential generation strategies for our proposed order and raster scan order respectively. Numbers indicate generation step in each order. (c)(f) visualize the conditional entropy increase when switching from sequential to parallel generation for each order, where darker red indicates larger entropy increase and thus higher prediction difficulty. Both orders generate the first four tokens sequentially (shown as white regions in entropy maps). Our proposed order that generates tokens from different spatial blocks in parallel shows smaller entropy increases compared to raster scan order that generates consecutive tokens simultaneously, indicating parallel generation across spatial blocks introduces less prediction difficulty than generating adjacent tokens simultaneously.

colors). This indicates that the dependencies between adjacent tokens are stronger compared to the dependencies between tokens that are farther apart in position. This observation aligns with the spatial locality in visual data, where nearby regions have stronger correlations than distant ones.

Next, we analyze how different token ordering strategies affect the difficulty of parallel generation in Fig. 12. To simulate the prediction difficulty during generation, we compute each token's conditional entropy given all its previous tokens - higher conditional entropy indicates more uncertainty and thus higher prediction difficulty at that position. By comparing the conditional entropy difference between sequential (one token at a time) and parallel generation (predicting multiple tokens simultaneously), we can quantify the increased difficulty introduced by parallel generation at each position. We conduct experiments with 4-token parallel prediction under two ordering strategies: our proposed generation order that first generates the initial four tokens sequentially to establish global structure, then generates tokens from different spatial blocks in parallel, and the raster scan ordering that directly predicts consecutive tokens simultaneously after the initial four tokens.

For our proposed order, we aim to characterize the entropy increase caused by the parallel generation, when compared to the entirely sequential generation methods. In particular, let $v_k^{(r)}$ be the token at position k in region r, we define $\mathcal{V}_{k,r}^{\text{seq}}$ and $\mathcal{V}_{k,r}^{\text{par}}$ by the sets of the previous tokens of $v_k^{(r)}$ for sequential and parallel generations (see Fig. 12(a)(b)). Then the conditional entropy of the sequential generation (single-token) and parallel generation (multi-token) are defined as $H(v_k^{(r)}|\mathcal{V}_{k,r}^{\text{seq}})$ and $H(v_k^{(r)}|\mathcal{V}_{k,r}^{\text{par}})$. We characterize the entropy increase caused by the parallel generation, i.e.,

$$H(v_k^{(r)}|\mathcal{V}_{k,r}^{\text{par}}) - H(v_k^{(r)}|\mathcal{V}_{k,r}^{\text{seq}}). \tag{7}$$

As a comparison, we also consider the raster scan order, where the tokens are exactly arranged based on their positions, denoted as $v_1, v_2, \ldots$. In this setting, given the current token v_k, we define $\mathcal{V}_k^{\text{seq}}$ and $\mathcal{V}_k^{\text{par}}$ by the sets of the previous tokens of v_k for sequential and parallel generations (see Fig. 12(d)(e)). Then, we will also characterize the entropy increase caused by the parallel generation in the raster scan order, i.e.,

$$H(v_k|\mathcal{V}_k^{\text{par}}) - H(v_k|\mathcal{V}_k^{\text{seq}}). \tag{8}$$

The numerical results of (7) and (8) are presented in Fig. 12(c) and (f). It can be seen that both orderings maintain identical conditional entropy for the first four tokens due to the sequential generation. For subsequent tokens, our proposed order leads to significantly smaller conditional entropy increases compared to the raster scan order. This indicates that when switching from sequential to parallel generation, generating tokens from different spatial blocks introduces less prediction difficulty than generating consecutive tokens in raster scan order. The result quantitatively validates our design.

References

[1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 5

[2] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 5

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 6, 7

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Sastry, et al. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *CVPR*, 2022. 1, 3, 6, 7

[6] Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023. 1, 3

[7] Liang Chen, Zekun Wang, Shuhuai Ren, Lei Li, Haozhe Zhao, Yunshui Li, Zefan Cai, Hongcheng Guo, Lei Zhang, Yizhe Xiong, et al. Next token prediction towards multimodal intelligence: A comprehensive survey. *arXiv preprint arXiv:2412.18619*, 2024. 1

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, pages 1691–1703, 2020. 1

[9] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3, 5, 1, 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 6, 7

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 1, 3, 5, 6, 7

[13] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *ECCV*, pages 102–118, 2022. 7

[14] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020. 1

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 5

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[17] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 6, 7

[18] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2022. 7

[19] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003. 5

[20] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *CVPR*, pages 10124–10134, 2023. 6, 7

[21] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3

[22] Siqi Kou, Lanxiang Hu, Zhezhi He, Zhijie Deng, and Hao Zhang. Cllms: Consistency large language models. *arXiv preprint arXiv:2403.00835*, 2024. 1, 3

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 1

[24] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 32, 2019. 7

[25] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022. 1, 3, 6

[26] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *ICML*, 2023. 1, 3

[27] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024. 6, 7

[28] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle: Speculative sampling requires rethinking feature uncertainty. *arXiv preprint arXiv:2401.15077*, 2024. 1, 3

[29] Dongyang Liu, Shitian Zhao, Le Zhuo, Weifeng Lin, Yu Qiao, Hongsheng Li, and Peng Gao. Lumina-mgpt: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024. 1

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 1

[31] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. *arXiv preprint arXiv:2303.08320*, 2023. 7

[32] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. 5

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[34] OpenLM-Research. Openllama 3b. https://huggingface.co/openlm-research/open_llama_3b, 2023. 5

[35] Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. Randar: Decoder-only autoregressive visual generation in random orders. *arXiv preprint arXiv:2412.01827*, 2024. 1

[36] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *ICML*, pages 4055–4064, 2018. 1

[37] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 6, 7

[38] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1, 5

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021. 1, 3

[41] Shuhuai Ren, Shuming Ma, Xu Sun, and Furu Wei. Next block prediction: Video generation via semi-autoregressive modeling. *arXiv preprint arXiv:2502.07737*, 2025. 1, 7

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6, 7

[43] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017. 1, 3

[44] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022. 6, 7

[45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2022. 7

[46] Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. Hierarchical patch diffusion models for high-resolution video generation. In *CVPR*, 2024. 7

[47] Yang Song, Chenlin Meng, Renjie Liao, and Stefano Ermon. Accelerating feedforward computation via parallel nonlinear equation solving. In *ICML*, 2021. 1, 3

[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3, 7, 1, 2, 4

[49] Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise parallel decoding for deep autoregressive models. *Advances in Neural Information Processing Systems*, 31, 2018. 3

[50] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 5, 2

[51] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 1, 2, 5, 6, 7

[52] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 1

[53] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 1, 3, 6, 7

[54] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE, 2015. 5

[55] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 5

[56] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1

[57] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[58] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *NeurIPS*, 2016. 1, 3

[59] Aäron Van Den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016. 1, 3

[60] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017. 1, 3

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 3

[62] Sergio Verdú. α-mutual information. In *2015 Information Theory and Applications Workshop (ITA)*, pages 1–6. IEEE, 2015. 5

[63] Chunqi Wang, Ji Zhang, and Haiqing Chen. Semi-autoregressive neural machine translation. *arXiv preprint arXiv:1808.08583*, 2018. 3

[64] Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. *arXiv preprint arXiv:2406.09399*, 2024. 7

[65] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 1

[66] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024. 3

[67] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 3

[68] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 6, 7

[69] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 1, 3

[70] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023. 7

[71] Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. In *ICLR*, 2024. 1, 3, 7