Infinity∞: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis

Supplementary Material

1. Predefined Scale Schedules

As listed in Tab.1, for each aspect ratio r, we predefine a specific scale schedule $\{(h_1^r, w_1^r), ..., (h_K^r, w_K^r)\}$. We ensure that the aspect ratio of each tuple (h_k^r, w_k^r) is approximately equal to r, especially in the latter scales. Additionally, for different aspect ratios at the same scale k, we keep the area of $h_k^r \times w_k^r$ to be roughly equal, ensuring that the training sequence lengths are roughly the same. We adopt buckets to support training various aspect ratios at the same time. The consistent sequence lengths of different aspect ratios improves training efficiency. During the inference stage, Infinity could generate photo-realistic images covering common aspect ratios (1:1, 16:9, 4:3, *etc.*) as well as special aspect ratios (1:3, 3:1, *etc.*) following the predefined scale schedules.

Figure 1. Prompt-following qualitative comparison. We highlight text in red that Infinity-2B consistently adheres to while the other four models fail to follow. Zoom in for better comparison.

2. Human Preference Evaluation

In order to measure the overall performance, we have conducted a human preference evaluation. We build a website and recruit volunteers to rank the generated images from different T2I models.

Prompts. We have collected 360 prompts in total, including prompts randomly sampled from Parti [9] and other human-written prompts. As illustrated in Fig.3, these

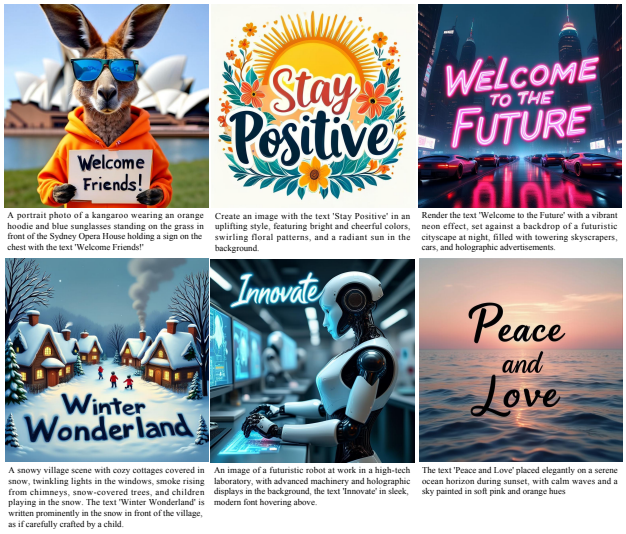

Figure 2. Text rendering results from our Infinity-2B model. Infinity-2B could generate text-consistent images following user prompts across diverse categories.

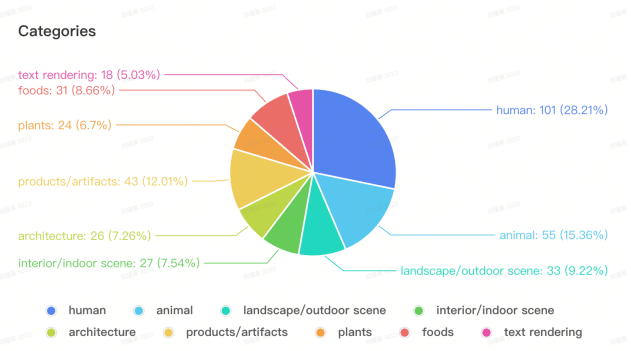

Figure 3. Distribution of Prompt Categories

prompts are divided into nine categories, such as human (28%), animal (15%), products/artifacts (12%), landscape (9%), foods, indoor scene, architecture, plants, and text rendering. It is worth noting that we incorporate a variety of human-related prompts, such as faces, bodies, and movements, in the human category as a supplement to the Parti prompts. In Fig.4, we also list the challenges of these prompts, which includes simple prompts, complex prompts, quantity, positioning & perspective, painting style, detail, semantic understanding, color, and imagination. These statistics demonstrate that the prompts used for evaluation are balanced, covering various categories and challenges

Aspect Ratio	Resolution	Scale Schedule											
1.000 (1:1)	1024×1024	(1,1)	(2,2)	(4,4)	(6,6)	(8,8)	(12,12)	(16,16)	(20,20)	(24,24)	(32,32)	(40,40)	(48,48)	(64,64)
0.800 (4:5)	896×1120	(1,1)	(2,2)	(3,3)	(4,5)	(8,10)	(12,15)	(16,20)	(20,25)	(24,30)	(28,35)	(36,45)	(44,55)	(56,70)
1.250 (5:4)	1120×896	(1,1)	(2,2)	(3,3)	(5,4)	(10,8)	(15,12)	(20,16)	(25,20)	(30,24)	(35,28)	(45,36)	(55,44)	(70,56)
0.750 (3:4)	864×1152	(1,1)	(2,2)	(3,4)	(6,8)	(9,12)	(12,16)	(15,20)	(18,24)	(21,28)	(27,36)	(36,48)	(45,60)	(54,72)
1.333 (4:3)	1152×864	(1,1)	(2,2)	(4,3)	(8,6)	(12,9)	(16,12)	(20,15)	(24,18)	(28,21)	(36,27)	(48,36)	(60,45)	(72,54)
0.666 (2:3)	832×1248	(1,1)	(2,2)	(2,3)	(4,6)	(6,9)	(10,15)	(14,21)	(18,27)	(22,33)	(26,39)	(32,48)	(42,63)	(52,78)
1.500 (3:2)	1248×832	(1,1)	(2,2)	(3,2)	(6,4)	(9,6)	(15,10)	(21,14)	(27,18)	(33,22)	(39,26)	(48,32)	(63,42)	(78,52)
0.571 (4:7)	768×1344	(1,1)	(2,2)	(3,3)	(4,7)	(6,11)	(8,14)	(12,21)	(16,28)	(20,35)	(24,42)	(32,56)	(40,70)	(48,84)
1.750 (7:4)	1344×768	(1,1)	(2,2)	(3,3)	(7,4)	(11,6)	(14,8)	(21,12)	(28,16)	(35,20)	(42,24)	(56,32)	(70,40)	(84,48)
0.500 (1:2)	720×1440	(1,1)	(2,2)	(2,4)	(3,6)	(5,10)	(8,16)	(11,22)	(15,30)	(19,38)	(23,46)	(30,60)	(37,74)	(45,90)
2.000 (2:1)	1440×720	(1,1)	(2,2)	(4,2)	(6,3)	(10,5)	(16,8)	(22,11)	(30,15)	(38,19)	(46,23)	(60,30)	(74,37)	(90,45)
0.400 (2:5)	640×1600	(1,1)	(2,2)	(2,5)	(4,10)	(6,15)	(8,20)	(10,25)	(12,30)	(16,40)	(20,50)	(26,65)	(32,80)	(40,100)
2.500 (5:2)	1600×640	(1,1)	(2,2)	(5,2)	(10,4)	(15,6)	(20,8)	(25,10)	(30,12)	(40,16)	(50,20)	(65,26)	(80,32)	(100,40)
0.333 (1:3)	592×1776	(1,1)	(2,2)	(2,6)	(3,9)	(5,15)	(7,21)	(9,27)	(12,36)	(15,45)	(18,54)	(24,72)	(30,90)	(37,111)
3.000 (3:1)	1776×592	(1,1)	(2,2)	(6,2)	(9,3)	(15,5)	(21,7)	(27,9)	(36,12)	(45,15)	(54,18)	(72,24)	(90,30)	(111,37)

Table 1. Predefined scale schedules $\{(h_1^r, w_1^r), ..., (h_K^r, w_K^r)\}$ for different aspect ratios. Following the text guided next-scale prediction scheme, Infinity takes K=13 scales to generate a 1024×1024 (or other aspect ratio) image.

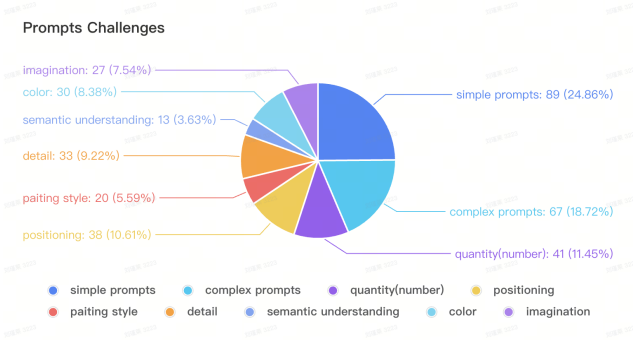

Figure 4. Distribution of Prompts Challenges

well.

Generated Images. We compare Infinity with four open-source models: PixArt-Sigma [1], SD3-Medium [2], SDXL [6], and HART [7]. The images of other models are generated by running their official inference code. No cherry-picking for any models.

Human Evaluation. For the human evaluation process, we build a website which presents two images from two anonymous models at the same time. There is one image generated by Infinity while the other is from other four models. Volunteers are required to pick a better one from two images in terms of *overall quality, prompt following, and visual aesthetics*, respectively. Besides the aforementioned criterion, we make sure each side-by-side comparison is evaluated by at least two volunteers to reduce human bias. We filter out pairs with opposite results evaluated by two volunteers. These contradictory pairs are sent to a third volunteer to assess. Then we take the consensus from three as the final results. Note that the whole process of human evaluation is completely double-blind. That is, a volunteer doesn't know which model it is, as well as other volunteers' results when performing a side-by-side comparison.

Results. As in Fig.6 of the submitted manuscript, we observe a remarkable human preference for Infinity over the other four open-source models. Especially for the comparison with HART [7] (another SOTA AR-based model), Infinity earns 90.0%, 83.9%, and 93.2% win rate in terms of overall quality, prompt following, and visual aesthetics, respectively. As for the diffusion family, Infinity earns 76.0%, 79.0%, 66.0% win rate to PixArt-Sigma, SDXL and SD3-Medium, respectively. What's more, Infinity reaches 71.1% win rate towards SD3-Medium regarding visual aesthetics. These results reveal that Infinity is more capable of generating visually appealing images. We attribute these great advantages to the proposed bitwise modeling, which has lifted the upper limits of AR models by large margins.

3. Ablation Studies

Optimal Strength for Bitwise Self-Correction. Algorithm 1 shows the detailed procedure of Bitwise Self-Correction. As illustrated in Tab.6, Bitwise Self-Correction mitigates the train-test discrepancy caused by teacher-forcing training. Here we delve into the optimal strength for applying bitwise self-correction in Tab.2. We empirically find that mistake imitation that is too weak (10% and 20%) fails to fully leverage the potential of Bitwise Self-Correction. Random flipping 30% bits yields the best results.

Positional Embedding. Learnable APE adopted in VAR [8] brings too many parameters and get confused when the sequence length varies. However, the sequence length changes frequently when training with various aspect ratios. Simply applying RoPE2d [3] or normalized RoPE2d [5] can not distinguish features from different resolutions. In this

Method	ImageReward↑	HPSv2.1↑
w/o Bitwise Self-Correction	0.515	29.53
Bitwise Self-Correction ($p = 10\%$)	0.751	30.47
Bitwise Self-Correction ($p = 20\%$)	0.763	30.71
Bitwise Self-Correction ($p = 30\%$)	**0.775**	**31.05**

Table 2. Comparison between different strengths of Bitwise Self-Correction. Experiment with 5M hight quality data and 512×512 resolution.

work, we apply RoPE2d and learnable scale embeddings on features of each scale. RoPE2d preserves the intrinsic 2D structure of images. Learnable scale embeddings avoids confusion between features of different scales. To verify the effectiveness, we compare it with the learnable APE in Fig.5. It's obvious that applying RoPE2d along with learnable scale embeddings on features of each scale converges faster and reaches higher training accuracy.

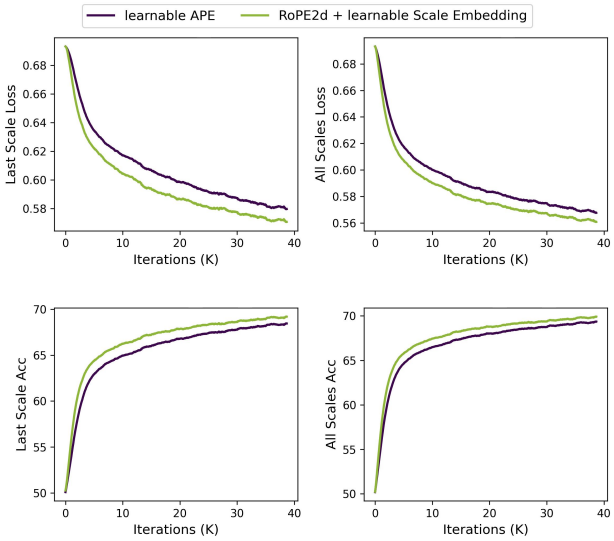

Figure 5. Comparison between learnable APE and our positional embeddings. Our method, *i.e.*, applying RoPE2d along with learnable scale embeddings on features of each scale, converges faster and reaches higher training accuracy.

Decoding. Decoding is crucial for improving generation quality. VAR adopts the pyramid Classifer-Free Guidance (CFG) on predicted logits. That is, the strength of CFG increases linearly as the scale goes from 1 to K. Such a pyramid scheme is to tackle the issue that the model collapses frequently when applying large CFG at early scales. We found that Infinity supports large CFG values even in very early scales equipped with Bitwise Self-Correction. Since Infinity is more robust to sampling, we revisit different decoding methods and find the best as illustrated in Tab.3. We visualize the comparison results of different decoding methods in Fig.6. We achieve the best generation results.

Method	Param	FID↓	ImageReward↑	HPSv2.1↑
Greedy Sampling	$\tau = 0.01, cfg = 1$	9.97	0.397	30.98
Normal Sampling	$\tau = 1.00, cfg = 1$	4.84	0.706	31.59
Pyramid CFG	$\tau = 1.00, cfg = 1 \rightarrow 3$	3.48	0.872	**32.48**
Pyramid CFG	$\tau = 1.00, cfg = 1 \rightarrow 5$	2.98	0.929	32.32
CFG on features	$\tau = 1.00, cfg = 3$	3.00	0.953	32.13
CFG on logits	$\tau = 1.00, cfg = 3$	2.91	0.952	32.31
CFG on logits (Ours)	$\tau = 1.00, cfg = 4$	**2.82**	**0.962**	32.25

Table 3. Comparison between different decoding methods.

Figure 6. Comparison of different sampling method. In contrast to Greedy Sample, Normal Sample and Pyramid Sample, our method could generate images with richer details and higher text-image alignments.

Algorithm 1 Bitwise Self Correction

Inputs: raw feature $\mathbf{F}$ processed by VAE Encoder
Hyperparameters: random flip proportion p, scale schedule $\{(h_1^r, w_1^r), ..., (h_K^r, w_K^r)\}$, $\mathbf{R}_{queue} = []$, $\widetilde{\mathbf{F}}_{queue} = []$
for $k = 1, \cdots, K$ **do**
  $\mathbf{R}_k = \text{quant}(\text{down}(\mathbf{F} - \mathbf{F}_{k-1}^{flip}, (h_k, w_k)))$
  queue_push($\mathbf{R}_{queue}, \mathbf{R}_k$)
  $\mathbf{R}_k^{flip} = \text{Random\_Flip}(\mathbf{R}_k, p)$
  $\mathbf{F}_k^{flip} = \sum_{i=1}^{k} \text{up}(\mathbf{R}_i^{flip}, (h, w))$
  $\widetilde{\mathbf{F}}_k = \text{down}(\mathbf{F}_k^{flip}, (h_{k+1}, w_{k+1}))$
  queue_push($\widetilde{\mathbf{F}}_{queue}, \widetilde{\mathbf{F}}_k$)
end for
Return: $\mathbf{R}_{queue} = \{\mathbf{R}_1, ..., \mathbf{R}_K\}, \widetilde{\mathbf{F}}_{queue} = \{\widetilde{\mathbf{F}}_1, ..., \widetilde{\mathbf{F}}_K\}$

4. More Qualitative Results

Fig.7 shows the qualitative comparison results among Infinity and other top-tier models. The images of other models are obtained either by querying their open-source demo

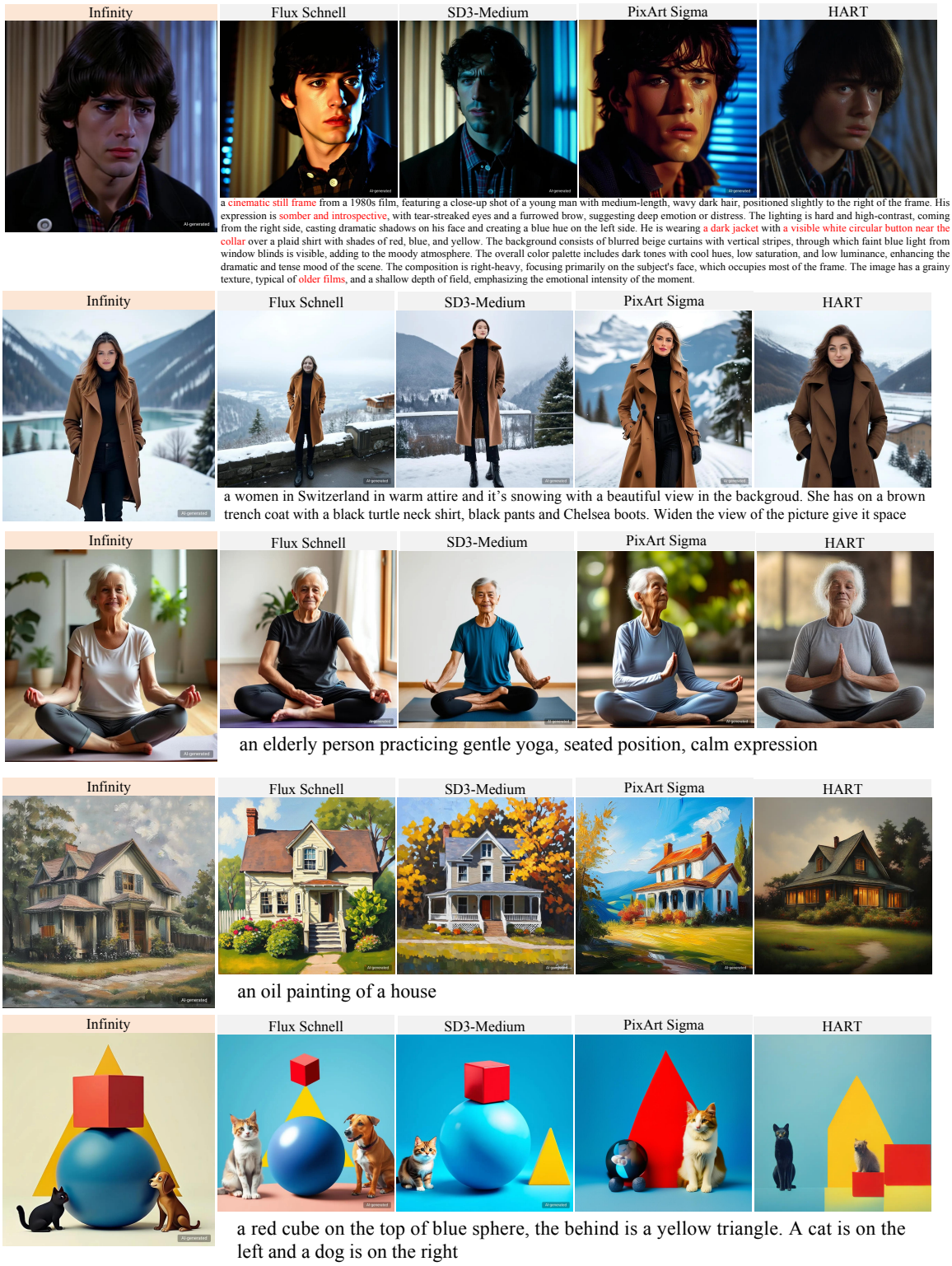

Figure 7. T2I qualitative comparison among our Infinity-2B model and the other four open-source models. Here we select three diffusion models (Flux Schnell, SD3-Medium and PixArt Sigma), one AR model (HART) for comparison. Zoom in for better comparsion.

website (HART [7]) or running their official inference code locally (Flux-Schnell [4], SD3-Medium [2], and PixArt Sigma [1]). Whether a thumbnail or a zoom-in image, we observe significant differences among the generated images from different models. In particular, the AR model like HART generates images with fewer details, blurred small human faces and texture-less background compared to diffusion models. In contrast, Infinity overcomes those shortcomings of AR models and generates better images compared to diffusion models like Flux-Schnell, SD3-Medium, and PixArt Sigma. Especially the third example of Fig.7, the other three models generate distorted human bodies while Infinity generates correct human hands and legs.

References

[1] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024. 2, 5

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 5

[3] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2025. 2

[4] Black Forest Labs. Flux. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024. 5

[5] Xiaoxiao Ma, Mohan Zhou, Tao Liang, Yalong Bai, Tiejun Zhao, Huaian Chen, and Yi Jin. Star: Scale-wise text-to-image generation via auto-regressive representations. *arXiv preprint arXiv:2406.10797*, 2024. 2

[6] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[7] Haotian Tang, Yecheng Wu, Shang Yang, Enze Xie, Junsong Chen, Junyu Chen, Zhuoyang Zhang, Han Cai, Yao Lu, and Song Han. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*, 2024. 2, 5

[8] 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. 2

[9] 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*, 2(3):5, 2022. 1