

---

# FlexVAR: Flexible Visual Autoregressive Modeling without Residual Prediction

---

## A Inference steps

In Tab. 1, we list the scales corresponding to different inference steps. The scales in each step are not fixed and can be flexibly adjusted during inference. Note that during training, we only limit the maximum number of steps to 10 and randomly sample the scale for each step, so the scales during the training process do not follow Tab. 1

Reso	Step	Scale
256px	6	{1, 2, 4, 6, 10, 16}
	7	{1, 2, 3, 5, 8, 11, 16}
	8	{1, 2, 3, 4, 6, 10, 13, 16}
	9	{1, 2, 3, 4, 5, 7, 10, 13, 16}
	10	{1, 2, 3, 4, 5, 6, 8, 10, 13, 16}
	11	{1, 2, 3, 4, 5, 6, 7, 9, 11, 13, 16}
	12	{1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 14, 16}
	13	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16}
384px	11	{1, 2, 3, 4, 5, 6, 8, 10, 13, 16, 24}
512px	12	{1, 2, 3, 4, 5, 6, 8, 10, 13, 16, 23, 32}
	15	{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 23, 32}

Table 1: Scale configurations of various inference steps.

## B Extent visual autoregressive modeling with Mamba

Unlike attention mechanisms that utilize explicit query-key-value (QKV) interactions to integrate context, Mamba faces challenges in handling bi-directional interaction. Therefore, prior Mamba-based visual autoregressive work [2] only used Mamba to model the unidirectional relationship between scales, relying on additional Transformer layers to process tokens within one scale.

In this work, we adopt a composition-recomposition strategy to obtain global information in Mamba network. Specifically, we utilize a Zigzag scanning strategy [1] over the spatial dimension. We alternate between eight distinct scanning paths across different Mamba layers (as shown in Fig. 1), which include:

- (a) top-left to the bottom-right.
- (b) top-left to the bottom-right.
- (c) bottom-left to the top-right.
- (d) bottom-left to the top-right.
- (e) bottom-right to the top-left.
- (f) bottom-right to the top-left.

- (g) top-right to the bottom-left.
- (h) top-right to the bottom-left.

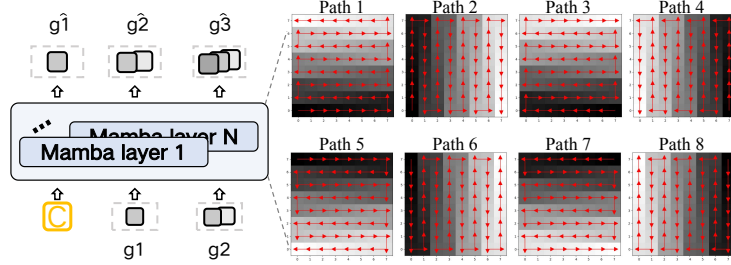


Figure 1: Spatial scan paths for Mamba.

### C Qualitative results with different steps.

In Fig. 2, we show some generated samples with  $\{6, 8, 10, 12\}$  steps. Our FlexVAR uses up to 10 steps for autoregressive modeling during training to avoid OOM (out-of-memory), while it can naturally transfer to any number of steps during inference. The samples generated with different steps are highly similar, differing only in some details. Generally, more steps result in better image details.

### D Qualitative results with various resolutions.

Fig. 3 shows some generated samples with  $\{256, 384, 512\}$  resolutions. FlexVAR uses up to  $256 \times 256$  resolution images for training, it can generate images with higher resolutions such as 384 and 512. The generated images demonstrate strong semantic consistency across multiple scales, and the higher resolutions display more detailed clarity.

### E Qualitative results with different VQVAE tokenizers.

**Image reconstruction.** We compare more image reconstruction results in Fig. 4. First, we encode the image into the latent space and perform multi-scale downsampling, then reconstruct the original image through the VQVAE decoder. It is evident that only our scalable VQVAE can perform image reconstruction at various scales.

**Generate images with GT prediction.** We visualize the generated samples with VQVAE tokenizers from VAR, Llamagen, and ours, corresponding to the 2<sup>nd</sup>, 3<sup>rd</sup> and 5<sup>th</sup> results in Tab. 6 in the main paper. As shown in Fig. 5, the VAR tokenizer, trained with a residual paradigm, fails to generate images under GT prediction; the generation samples of Llamagen’s tokenizer are not up to the mark, due to its discrete tokens at intermediate steps being suboptimal.

### F Additional Visual Results.

We show more generated samples in Fig. 6.

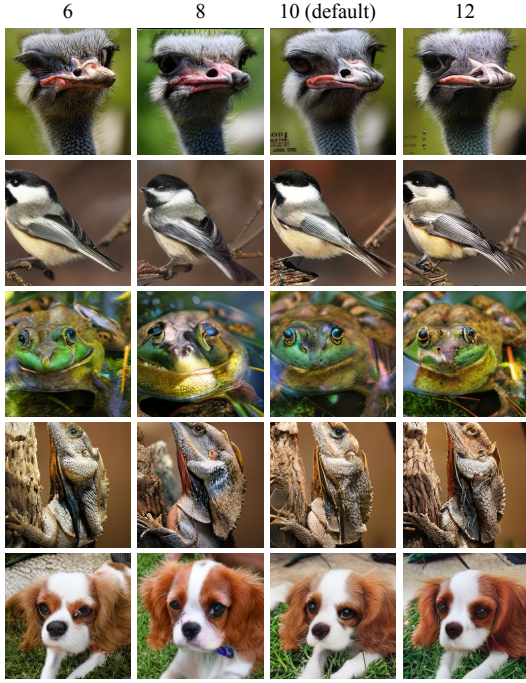


Figure 2: Some generated samples with {6, 8, 10, 12} steps. Note the model is trained with steps  $\leq 10$ . More steps typically result in better image details. Zoom in for a better view.

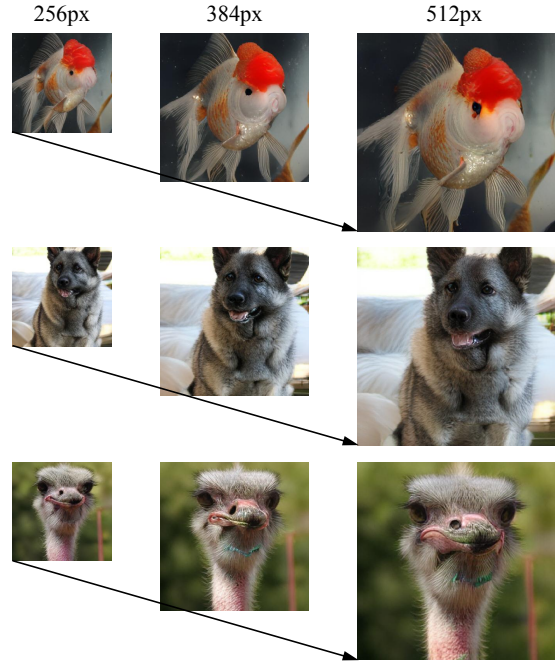


Figure 3: Some generated samples with {256, 384, 512} resolutions. Note the model is trained with a resolution of  $\leq 256 \times 256$ . Zoom in for a better view.

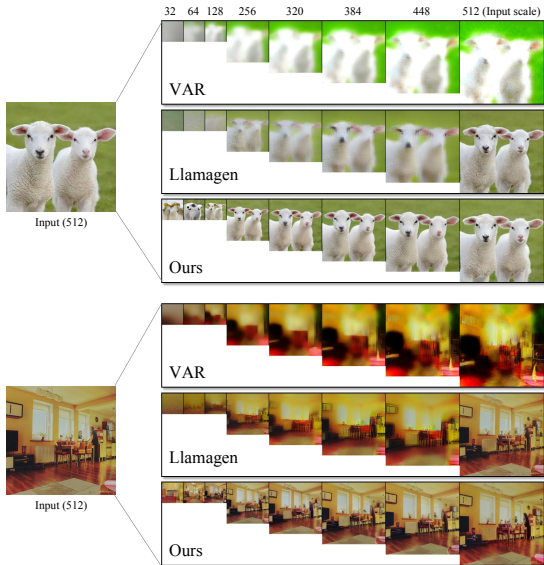


Figure 4: Compared with different VQVAE tokenizers [4, 3] for multi-scale reconstructing images, we downsample the latent features in VQVAE to multiple scales and then use the VQVAE Decoder to reconstruct images. We upsample images  $< 100$  pixels using bilinear interpolation for a better view.

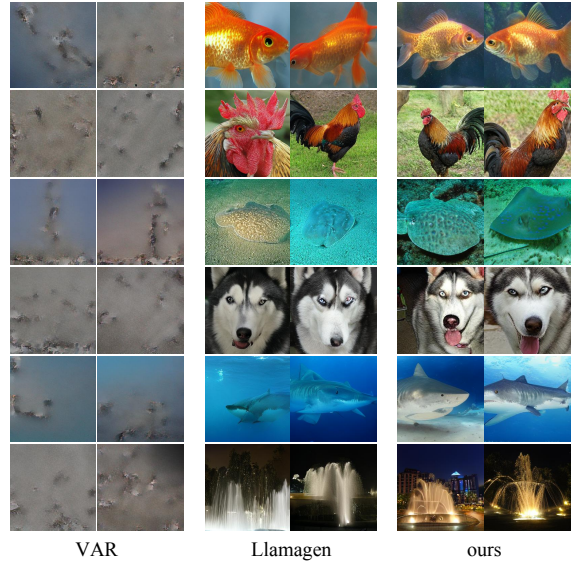


Figure 5: Some generated samples with different VQVAE tokenizers (Llamagen & VAR), corresponding to the  $2^{nd}$  and  $3^{rd}$  results in Tab. 6 in the main paper. We report the results with model scale -d20 trained 40 epochs ( $\sim 70K$  iterations) on ImageNet-1K. Zoom in for a better view.



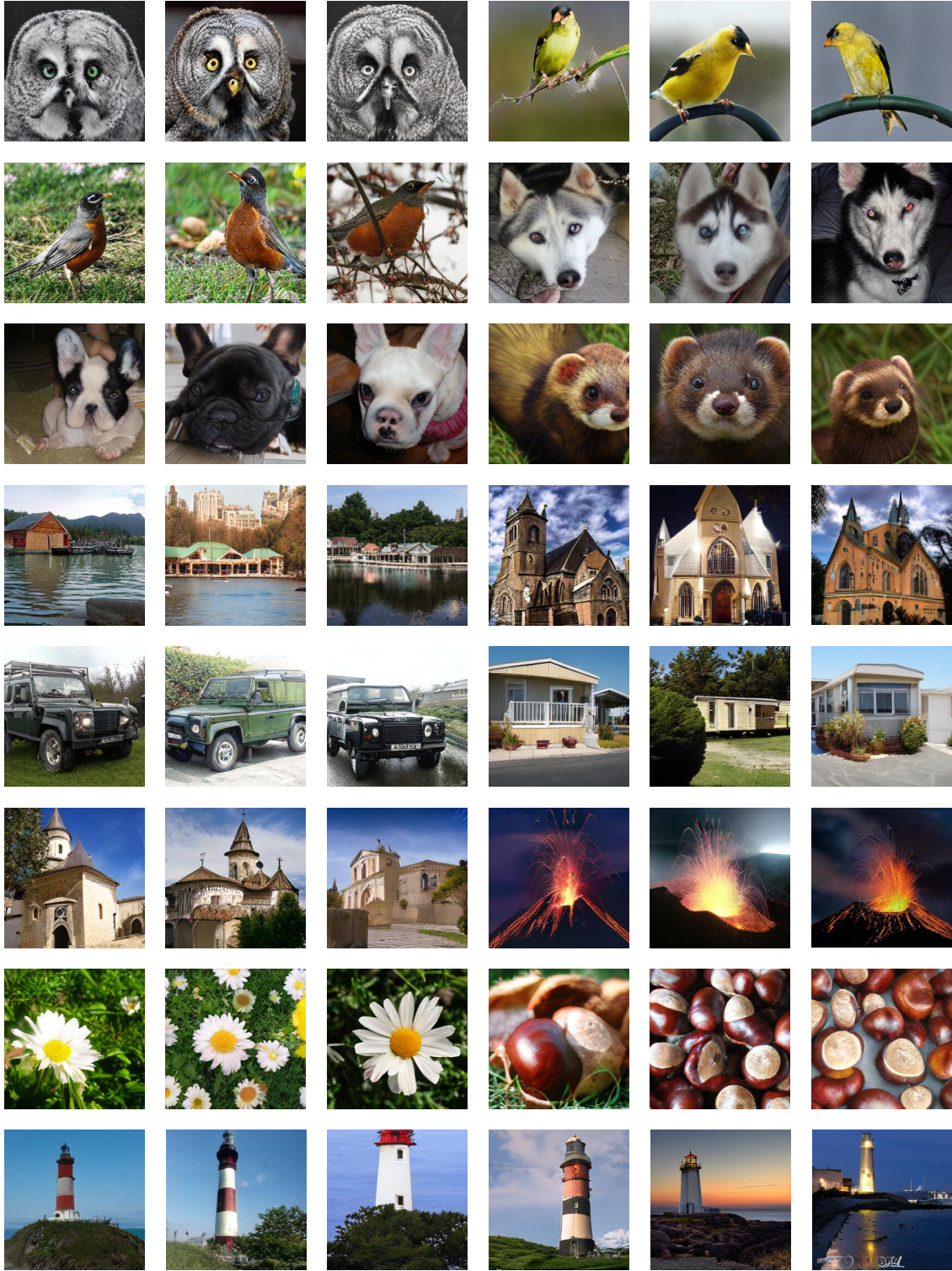


Figure 6: Some generated  $256 \times 256$  samples.

## References

- [1] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Björn Ommer. Zigma: A dit-style zigzag mamba diffusion model. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- [2] Sucheng Ren, Yaodong Yu, Nataniel Ruiz, Feng Wang, Alan Yuille, and Cihang Xie. M-var: Decoupled scale-wise autoregressive modeling for high-quality image generation. *arXiv preprint arXiv:2411.10433*, 2024.
- [3] 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.
- [4] 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.