Holistic Tokenizer for Autoregressive Image Generation

Supplementary Material

First, we compare Hita with other vanilla tokenizers and further discuss the token fusion module. Then, we elaborate on more ablations in detail. Later, we present more visualization samples egarding zero-shot style transfer, zero-shot in-painting, and class-conditional image generation.

A. Comparison with other image tokenizers.

In this subsection, we first compare Hita with the other vanilla tokenizers. Next, we depict the differences between VAR [12] in detail. Next, we present further discussion on token fusion modules.

A.1. Comparison with vanilla image tokenizers

In Table. 1, we compare with the prevalent image tokenizer, including VQGAN [3], MaskGiT [1], ViT-VQGAN [13] and TiTok [14]. The image reconstruction quality is measured by rFID [6] and rIS [10] metrics, which are evaluated on 256×256 ImageNet [2] 50k validation benchmark.

Following VQGAN [3], we adopted ℓ_2-normalization into codebook vectors, low codebook vector dimension, and a codebook size of 16,384. Compared with other counterparts like VQGAN [3] and ViT-VQGAN [13], the proposed tokenizer represents an image with fewer tokens (569 *vs.* 1024), while achieving a better reconstruction quality with **100%** utilization for both holistic and patch-level codebooks. Additionally, we observed that our approach achieves a better rIS **198.5** compared with the VQGAN proposed in LlamaGen [11]. rIS quantifies the KL-divergence between the original label distribution and the logit distribution of reconstructed images after softmax normalization [10]. In other words, rIS measures the semantic consistency between the reconstructed images and the original ones. A higher rIS confirms that our holistic tokenizer is more effective at preserving the semantic consistency of the reconstructed images.

A.2. Comparison with VAR

As discussed in Sec.2, VAR's multi-scale tokens can also be considered as a combination of semantic tokens and patch tokens. However, as depicted in Table. 2 we find removing the initial coarse-scale tokens seldom effects its reconstruction. Meanwhile, linear probing conducted on the cumulative coarse-scale tokens reveals poor semantic information.

Approach	f	*setup*			*img recon.*		*usage(%)*↑	
		size	dim	#toks	rFID↓	rIS ↑	$\mathcal{Q}_H$	$\mathcal{Q}_P$
TiTok [14]	–	8,192	64	256	1.05	191.5	–	100.0
VQGAN$^{\text{oim.}}$ [3]	8	256	4	1024	1.44	–	–	–
VQGAN [3]		8192	256		1.49	–	–	–
ViT-VQGAN [13]		8192	32		1.28	192.3	–	95.0
VQGAN$^{\text{oim.}}$ [3]		16384	4		1.19	–	–	–
VQGAN [3]	16	1024	256	256	7.94	–	–	–
MaskGiT [1]					2.28	–	–	–
Var [12]	16	4096	32	680	**0.92**	196.0	–	100.0
RQ-VAE [8]	32	16384	256	1024	1.83	–	–	–
VQGAN [3]	16	16384	256	256	4.98	–	–	–
VQGAN [11]			8	441	1.21	189.1	–	99.2
VQGAN [11]			8	576	0.95	197.3	–	99.7
Hita			8/12	569	1.03	**198.5**	100.0	100.0

Table 1. Comparison with other image tokenizers. $^{\text{oim.}}$ indicates training on OpenImages [7]. $\mathcal{Q}_H/\mathcal{Q}_P$ denote the codebook usage in holistic and patch-level quantizers, respectively.

n	0	1	2	3	4	5	6	7	8
rFID	1.31	1.31	1.32	1.31	1.41	1.78	3.22	10.42	92.3
rIS	198.6	199.3	198.8	198.9	196.4	190.2	171.8	119.8	16.4
Acc	2.2	4.9	7.2	8.3	9.2	9.5	9.8	10.1	10.3

Table 2. Analysis of VAR's multi-scale tokens. Acc indicates top-1 accuracy for linear probing estimation on the ImageNet [2].

B. Explanation on Token Fusion Module

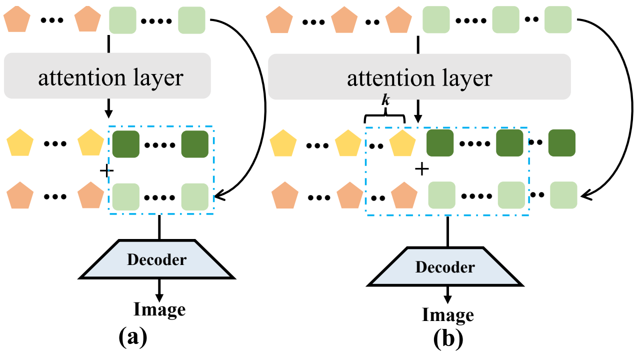

Figure 1. Token fusion module comparison. (a).Token fusion module composed of vanilla attention layers. (b). Our designed token fusion layers.

As shown in Fig. 1, If a vanilla transformer is used to construct the token fusion module, due to the existence of the skip connection and the patch-level tokens contain enough information for image reconstruction, the patch-level tokens can directly flow through the skip connections into the decoder to reconstruct the image. Thus, the token fusion module can learn a trivial solution, which overlooks the holistic tokens and leads to holistic codebook collapse (Fig. 1a). Once the last k holistic tokens participate in the image reconstruction, the information of the patch tokens flowing through the skip connection is incomplete. It needs to interact with the holistic tokens to obtain complete information for image reconstruction(Fig. 1b). This simple operation emphasizes the holistic tokens and avoids codebook collapse. To better align the token sequence with the nature of the AR generation model, here we adopt causal attention to build the token fusion module.

C. More Ablation Studies.

We elaborate on further ablation studies on the design of our approach. Next, we quantitatively analyze Hita's zero-shot inpainting performance.

C.1. AR generation with resolution of 256×256

Given that Hita ensures a fair comparison with other approaches by controlling the number of tokens, here we directly train the image tokenizer and the AR generation model on 256×256 images, enabling them to reconstruct and generate 256×256 images, respectively, which is also in line with common practice. We initialize and train the image tokenizer for 40 epochs, and train both Hita-B and Hita-L for 50 epochs as default. As depicted in Table. 3, Hita can not only achieve a better reconstruction performance, but also improve the generation quantity compare to LlamaGen [11].

Approach	*Image recon.*		*code usage*↑		*AR gen.*	
	rFID↓	rIS↑	$\mathcal{Q}_H$	$\mathcal{Q}_P$	gFID↓	gIS↑
LlamaGen-B	2.22	169.8	–	95.2%	7.22	178.3
LlamaGen-L					4.21	200.0
Hita-B	1.40	186.6	100%	100%	**6.58**	**210.2**
Hita-L					**4.04**	**242.2**

Table 3. Hita performs image reconstruction and AR generation with image resolution of 256×256.

C.2. Other Token Fusion Variants

As depicted in Sec 3.2.3, for simplicity, we choose the last k holistic tokens combined with the patch-level tokens to reconstruct the image. Here, we refer to MAE [5] and Titok [14] and design 2 different variants to generate features for the first k patch tokens: 1). **Partial**: A mask token combined with the positional embedding of the first k patch tokens are used to generate their feature. In this scenario, the holistic tokens can be treated as condition; 2). **Full**: The patch tokens are completely removed, which is consistent with TiTok [14]. Then, a mask token, along with the positional embeddings, generates the features for all patch tokens. To achieve this, we initialize and train 2 different holistic tokenizers for 40 epochs, and then train 2 AR generation models – Hita-B and Hita-L based on those tokenizers with default training settings for 50 epochs. As listed in Table 4, the token fusion strategy proposed in Hita shows a better performance in both reconstruction and generation compared to the other variants. Thus, we choose to use the last k holistic tokens combined with the patch-level tokens to reconstruct the image, by default.

Variants	*image recon.*		*code usage*↑		*AR gen.*		
	rFID↓	rIS↑	$\mathcal{Q}_H$	$\mathcal{Q}_P$	Model	gFID↓	gIS↑
Partial	1.05	198.2	100.0%	100.0%	B	6.59	209.8
					L	3.96	243.1
Full	2.07	170.3	100.0%	–	B	11.64	172.1
				–	L	6.75	219.7
Hita	1.03	198.5	100.0%	100.0%	B	**5.85**	**212.3**
					L	**3.75**	**262.1**

Table 4. Image reconstruction and AR generation. with different token fusion strategies.

C.3. Attention Modules Study.

As outlined in Sec. 2, the attention modules consists of one standard transformer $\mathcal{E}_{\text{trans}}(\cdot)$, two causal transformers $\mathcal{E}_{\text{causal}}(\cdot)$ and $\hat{\mathcal{E}}_{\text{causal}}(\cdot)$. $\mathcal{E}_{\text{trans}}(\cdot)$ for holistic feature capture, $\mathcal{E}_{\text{causal}}(\cdot)$ is for causal latent space alignment, and $\hat{\mathcal{E}}_{\text{causal}}(\cdot)$ for holistic codebook learning. Here we study their effectiveness as follows: 1) We remove $\mathcal{E}_{\text{trans}}(\cdot)$ and introduce new attention mask into $\mathcal{E}_{\text{causal}}(\cdot)$ to learn its contribution to holistic feature capture; 2). We substitute $\mathcal{E}_{\text{causal}}(\cdot)$ to study its effect on causal latent space alignment. Here, we initialize and train the tokenizers with different attention modules. Then, we train an AR generation model (Hita-B) on those holistic tokenizers to estimate their generation quality.

As depicted in Table. 5, it can be observed when only $\mathcal{E}_{\text{trans}}$ is directly discarded from the tokenizer, the quality of image reconstruction and generation slightly drops. This is because the subsequent causal transformer $\mathcal{E}_{\text{causal}}$ can simultaneously achieve the requirement of holistic feature capture, semantic-aware feature injection, and feature space alignment. Similarly, only removing $\mathcal{E}_{\text{causal}}$ leads to a slight degradation in image reconstruction and generation, indicating that incorporating causal attention $\mathcal{E}_{\text{trans}}$ within the tokenizer helps in learning a latent space that better aligns with the causal nature of AR models. With all the modules integrated, we achieve the best performance in terms of reconstruction and generation.

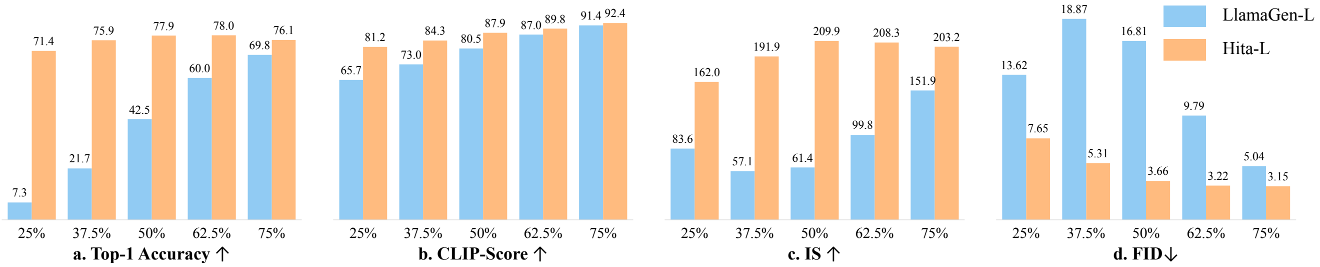

Figure 2. Quantitative zero-shot in-painting analysis with Hita-L conducted on ImageNet [2] evaluation benchmark.

component			*image recon.*		code usage↑		*AR gen.*	
$\mathcal{E}_{\text{trans}}$	$\mathcal{E}_{\text{causal}}$	$\hat{\mathcal{E}}_{\text{causal}}$	rFID↓	rIS↑	$\mathcal{Q}_H$	$\mathcal{Q}_P$	gFID↓	gIS↑
–	–	–	1.21	189.1	–	99.7%	7.95	166.9
–	✓	✓	1.15	196.0	100.0%	100.0%	6.01	196.3
✓	–	✓	1.20	195.3	100.0%	100.0%	6.12	188.9
✓	✓	✓	**1.03**	**198.5**	100.0%	100.0%	**5.85**	**212.3**

Table 5. Attention modules analysis in Hita. '-' indicates the module was removed from the tokenizer's architecture.

C.4. Quantitative Study of In-painting Quality.

Beyond the demonstration that the holistic tokenizer can help AR generation model maintain a better semantic consistency (see Fig. 4), we conduct a quantitative analysis of this. Similar to the experimental setup described in the manuscript, we only retain a certain fraction of upper part image, *e.g.* 25%, 50%, *etc.* then utilize a tokenizer to discretize it into visual tokens. These tokens are fed as a prefix sequence prompt to a pre-trained AR generation model, which is required to complete the lower part of the image. Here, we take the AR generation models – Hita-L and LlamaGen-L [11] for a fair comparison.

Evaluation metrics. To quantitatively estimate the semantic consistency of the generated images, we adopt top-1 accuracy and CLIP-score as our metrics, along with the generation metrics FID and IS. The top-1 accuracy is derived from image classification tasks, where we utilize a pre-trained ResNet-101 [4] on ImageNet [2] to classify the completed images. CLIP-score measures the similarity between the original and the completed images. Specifically, we feed both original and in-painted images into the CLIP [9] model to extract image features and compute their cosine distance averaged across all samples. Higher top-1 accuracy and CLIP-scores indicate a better semantic consistency is maintained in an AR generation model.

Observation and discussion. As shown in Fig. 2 and Fig. 4, the AR generation model trained with vanilla VQGAN [3] encounters difficulties in producing an image that maintains semantic coherence. In contrast, our approach effectively produces visually consistent content to complete the given image part, maintaining a better overall coherence even with a significantly truncated prefix. Specifically, as observed in Fig. 2.a and Fig. 2.b, our method achieves higher classification accuracy and CLIP-score under various settings, with relatively stable fluctuations in all metrics across different configurations. In contrast, models trained with vanilla VQGAN exhibit more obvious performance variations, especially when only a small portion of the image is provided.

Beyond measuring semantic coherence, we also estimate the completed image quality using the generation evaluation metrics, FID and IS. As shown in Fig. 2.c and Fig. 2.d, compared to the model trained with vanilla VQGAN[3], our approach achieves a better FID and IS, which also illustrates the holistic tokenizer can help the AR generation model generate the lower half of images robustly, maintaining strong semantic consistency.

D. Limitations

Currently, Hita is trained on a limited dataset using basic optimization techniques. Better performance could be achieved with more data and advanced learning objectives. Additionally, Hita can be seamlessly extended to perform text-conditional image generation, which is an ongoing direction of our research.

E. More Visual Examples

In this section, we present more visualization samples including zero-shot style transfer (see Fig. 3), image in-painting (see Fig. 4), and class-conditional image generation (see Fig. 5). For optimal clarity, please zoom in.

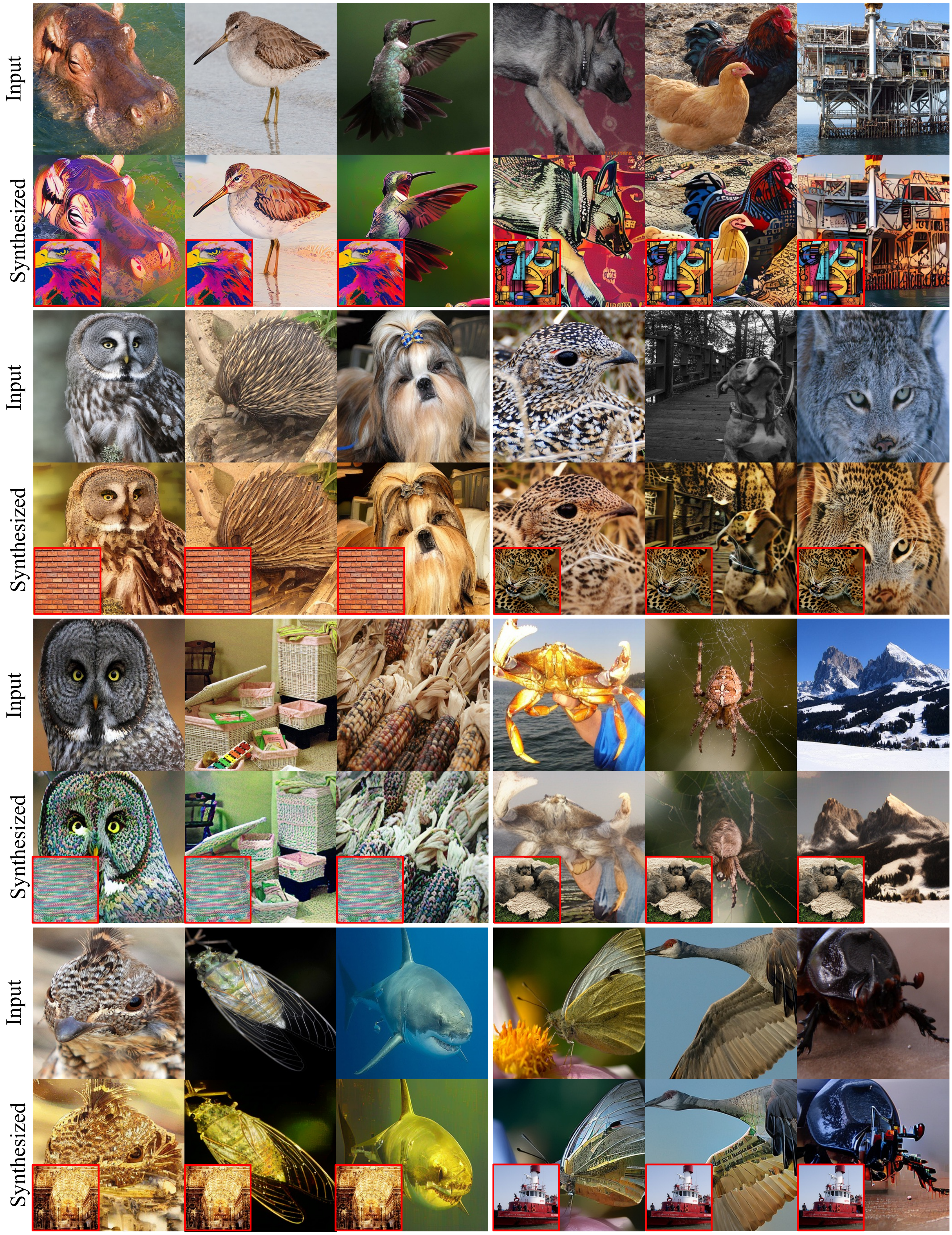

Figure 3. Some zero-shot style-transfer samples by Hita's holistic tokenizer. Best viewed with zoom-in.

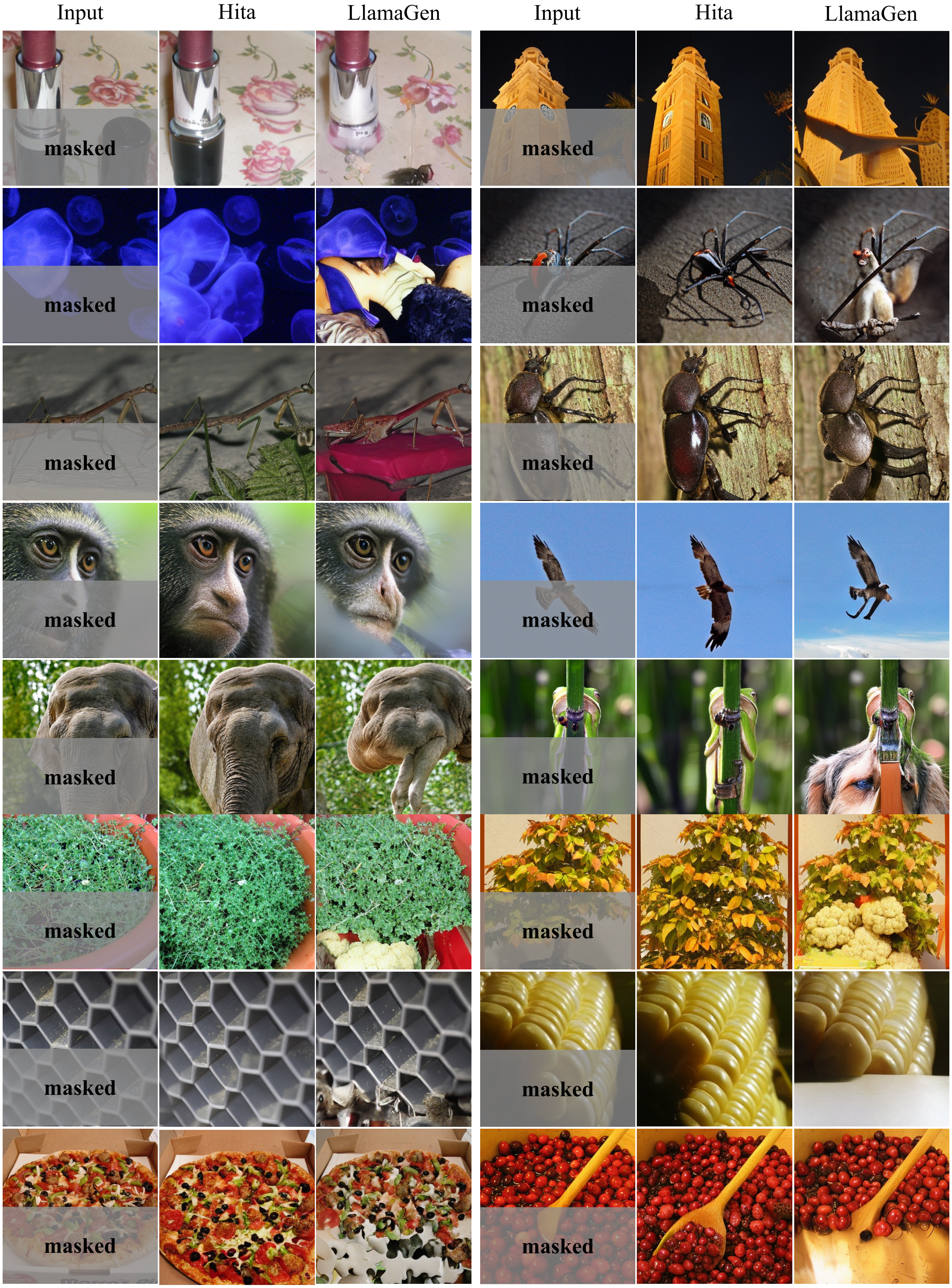

Figure 4. Zero-shot in-painting examples by Hita's AR generation model. Compared with the baseline. Best viewed with zoom-in.

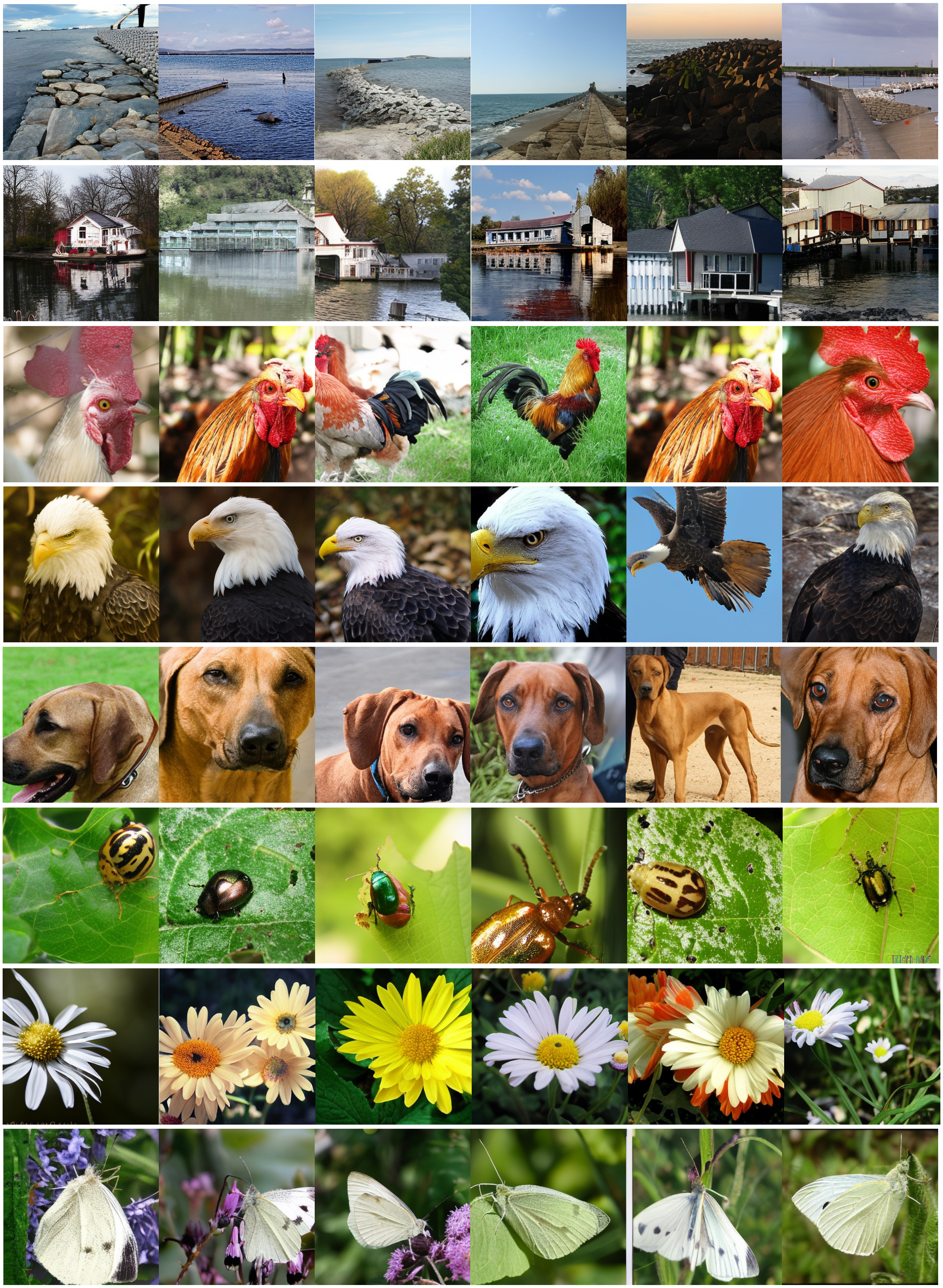

Figure 5. Visualization of class-conditional samples generated by Hita. Best viewed with zoom-in.

References

[1] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3

[3] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 1, 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[6] 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. *Advances in neural information processing systems*, 30, 2017. 1

[7] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 1

[8] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[10] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 1

[11] 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, 3

[12] 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

[13] 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. 1

[14] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *arXiv preprint arXiv:2406.07550*, 2024. 1, 2