Scalable Image Tokenization with Index Backpropagation Quantization

Supplementary Material

Model	Parameters	Width w	Head h	Depth d	Lr	Batch Size	Epoch
IBQ-B	342M	16	16	1024	3e-4	768	300
IBQ-L	649M	20	20	1280	3e-4	768	350
IBQ-XL	1.1B	24	24	1536	3e-4	768	400
IBQ-XXL	2.1B	30	30	1920	3e-4	768	450

Table 1. **Model sizes and architecture configurations of IBQ.**

Model	Optimization	Training	Inference	rFID↓	Usage↑
Soft VQ	Corrupted	Soft	Soft	16.17	2.5%
Soft VQ	Corrupted	Soft	Hard	233.17	2.5%
IBQ (Ours)*	Stable	Hard	Hard	4.03	99%
IBQ (Ours)	Stable	Hard	Hard	1.37	96%

Table 2. **Comparison with Soft Vector Quantization.** Soft VQ training corrupts after a few epochs. When adopting hard quantization for inference, there is a significant drop in rFID. * denotes IBQ with the same training epochs as Soft VQ.

Model	Codebook Size	Parameters	Memory	Time/epoch	Usage
VQGAN	1,024	89.6M	19.5G	3h15min	44%
	8,192	91.5M	19.7G	3h18min	-
	16,384	93.6M	19.8G	3h21min	5.3%
	262,144	156M	21.2G	4h	∼0%
IBQ	1,024	89.6M	19.5G	3h20min	99%
	8,192	91.5M	19.7G	3h30min	98%
	16,384	93.6M	20G	3h40min	96%
	262,144	156M	30.5G	9h	84%

Table 3. **Training computational costs comparison between VQGAN and IBQ.** (Tested on 8 A6000 gpus)

1. Autoregressive Model Configurations

We show the detailed autoregressive model configurations and training settings in Tab. 1. We scale up the autoregressive models from 300M to 2.1B parameters, following the scaling rules proposed in VAR [10].

2. Comparison with Soft Vector Quantization

To comprehensively illustrate the rationality of our IBQ, we compare it with another global update method, Soft Vector Quantization (Soft VQ). During training, it adopts the weighted average of all code embeddings as the quantized feature v_q and incorporates a cosine decay schedule of the temperature ranging from 0.9 to $1e$-6 for one-hot vector approximation. As for inference, it switches back to the original VQGAN way, which selects the code with the highest probability for hard quantization.

As shown in Tab. 2, Soft VQ is far behind IBQ in both reconstruction quality and codebook usage. In the experiments, we observe that the training process of Soft VQ corrupts within a few epochs (< 10). This may stem from the unstable adversarial training where the adaptive weight of the GAN loss appears enormous and ends up with NAN. In addition, the soft-to-hard manner for one-hot vector approximation brings more difficulty in optimization and incurs inconsistency of quantization between training and inference, as demonstrated by a significant reconstruction quality drop ($16.17\text{rFID} \rightarrow 233.17\text{rFID}$).

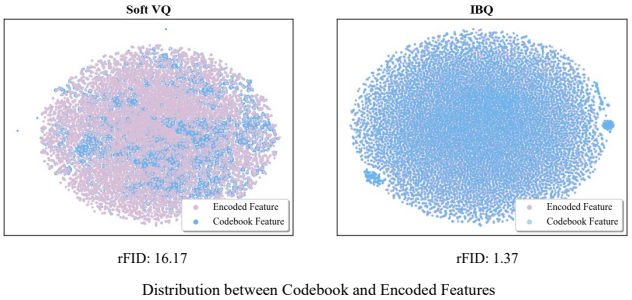
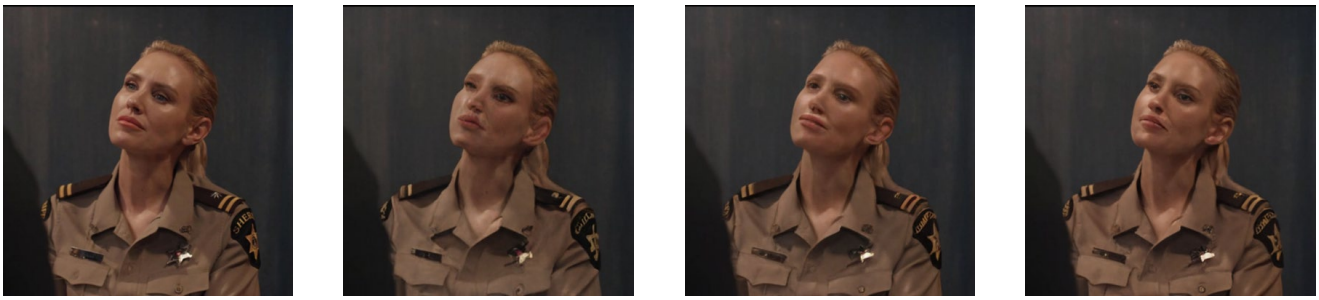


Figure 1. **Distribution Gap.** The T-SNE results of the codebook (16,384 codebook size and 256 dimension) and sampled encoded features.

Moreover, we provide an in-depth investigation by visualizing the distribution between the codebook and encoded features of Soft VQ. As shown in Fig. 1, although all-code updating strategy is enabled, the inappropriate quantization process tends to cluster codes mistakenly, resulting in low codebook usage (2.5%). We speculate that the force of the weighted average of code embeddings toward the encoded feature will smooth the codebook representation and result in similar and less informative code embeddings. In contrast, IBQ adopts hard quantization with index backpropagation. The hard quantization only involves the selected codes toward the encoded features for discriminative representation, thus ensuring precise quantization, while index backpropagation performs joint optimization of the entire codebook and visual encoder to achieve consistent distribution. Considering the factors above, our proposed IBQ shows dominance in both reconstruction quality and codebook utilization.

3. Training Costs

We evaluate the training costs of VQGAN and IBQ under varying codebook sizes using 8 A6000 GPUs. As shown in Tab. 3, the all-codes updating mechanism of IBQ incurs only a marginal increase in training costs compared to VQGAN when the codebook size is up to 16,384, yet it significantly improves codebook utilization. Specifically, IBQ introduces an additional 0.2 GB of memory usage and extends training time by 19 minutes, but increases codebook utilization from 5.3% to 96%. Furthermore, VQGAN fails to train with an extremely large codebook (i.e., 262,144 entries), whereas IBQ successfully achieves 84% utilization.



Original 16384 262144 262144 + Finetune

Figure 2. **Face reconstruction comparison.** Scaling up tokenizers and finetuning tokenizers on face data can effectively improve facial reconstruction performance.

Method	Ratio	Codebook Size	MS-COCO 2017			Imagenet-1k		
			rFID↓	PSNR↑	SSIM↑	rFID↓	PSNR↑	SSIM↑
LlamaGen† [9]	16	16384	8.40	20.28	0.55	2.47	20.65	0.54
Show-o [11]	16	8192	9.26	20.90	0.59	3.50	21.34	0.59
Cosmos [1]	16	64000	11.97	19.22	0.48	4.57	19.93	0.49
IBQ (Ours)	16	16384	**7.67**	**21.58**	**0.62**	**2.06**	**22.01**	**0.61**
IBQ (Ours)	16	262144	**6.79**	**22.28**	**0.65**	**1.53**	**22.69**	**0.64**

Table 4. **Zero-shot reconstruction performance on ImageNet 50k validation set and MS-COCO val2017.** The tokenizers are trained with large-scale general-domain datasets and aim to serve text-conditional image generation. The results are reported under the same setup for fair comparison. † indicates that LlamaGen loads the model initially trained on Imagenet while the others are training from scratch (*i.e.*, MS-COCO and Imagenet-1k are excluded from training data).

4. Pretraining Tokenizer

We further unveil the representation capacity of our tokenizer by pretraining IBQ on large-scale domain datasets, i.e., 1) General: CapFusion [12], LAION-COCO [4], CC12M [2] and CC3M [8]. 2) High-quality: LAION-aesthetics-12M[1], LAION-aesthetics [7], JourneyDB [5] and LAION-HD[2]. We follow the same training settings stated in the manuscript while the training steps are $\sim$ 800,000. It can be seen in the Tab. 4 that IBQ achieves state-of-the-art performance compared to concurrent methods such as Cosmos [1], Show-o [11]. Although some recent efforts in residual tokenization [3, 6] can achieve better results, they are not listed here because residual techniques are orthogonal and compatible with IBQ. It is anticipated that our improvement on the naive quantization method better benefits the unified visual understanding and generation models compared to the residual one.

5. Improving Face Reconstrution

Visual tokenizers trained on ImageNet may not perform as expected for face reconstruction. Increasing the codebook size can effectively mitigate this limitation. As shown in Fig. 2, increasing the codebook size from 16,384 to 262,144 leads to improved face reconstruction quality. Additionally, incorporating face data into the training set or fine-tuning on face-specific datasets are effective strategies for further enhancement. In particular, fine-tuning IBQ on the FFHQ dataset further enhances reconstruction performance.

6. Additional Visualizations

We provide more qualitative reconstruction and generation samples in Fig. 3 and Fig. 4, respectively.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 2

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021. 2

[3] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bit-

[1] https://huggingface.co/datasets/dclure/laion-aesthetics-12m-umap
[2] https://huggingface.co/datasets/yuvalkirstain/laion-hd-subset

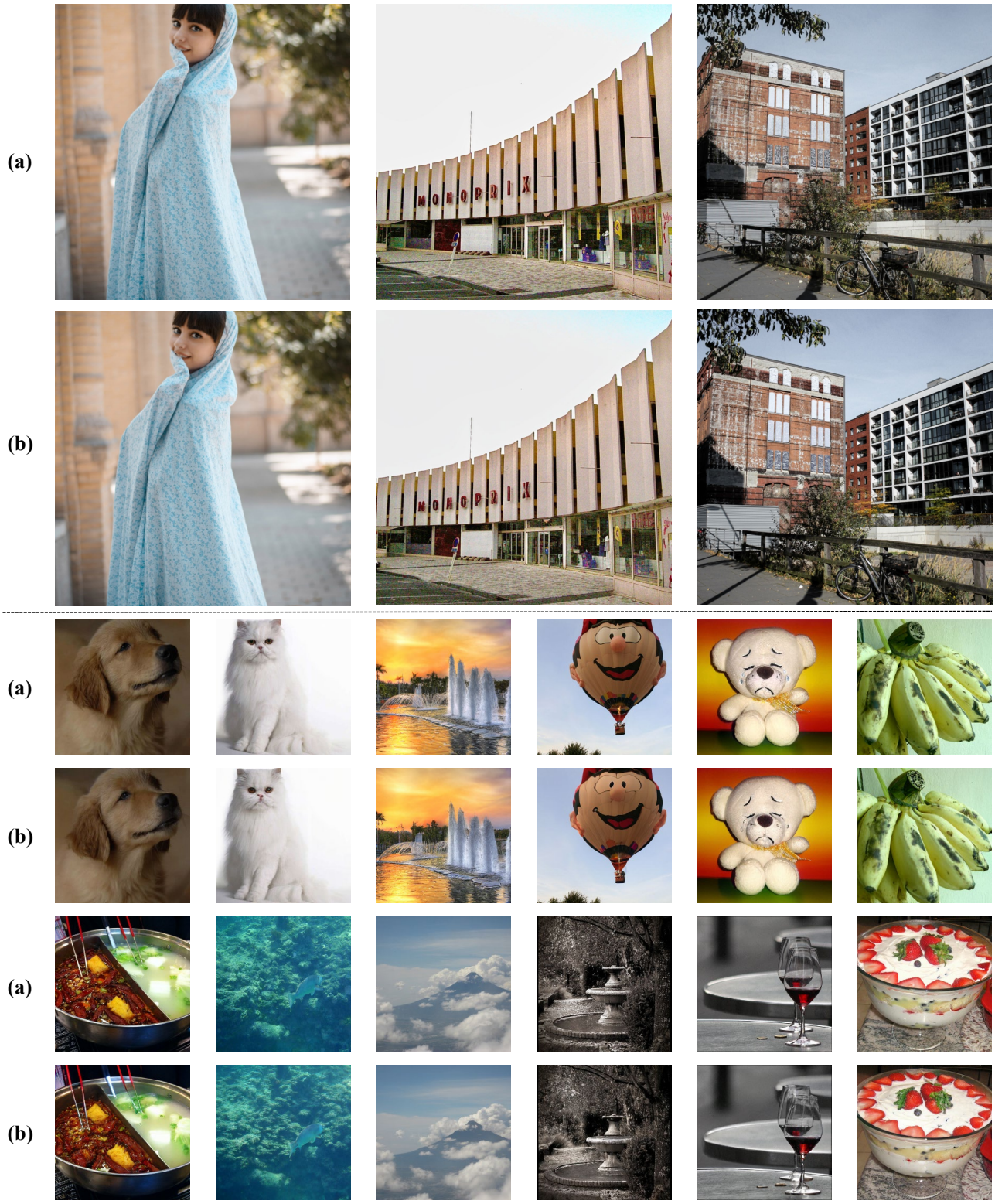


Figure 3. **Reconstruction samples.** The upper part illustrates the IBQ tokenizer tested at 1024 $\times$ 1024 Unsplash. While the second part showcases the IBQ tokenizer tested at 256 $\times$ 256 Imagenet. (a) indicates the original images and (b) signifies the reconstructions.

Figure 4. **Generation samples.** We showcase the 256 × 256 class conditional generation samples on Imagenet.

wise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024. 2

[4] LAION. Laion-coco 600m. https://laion.ai/blog/laion-coco, 2022. 2

[5] Junting Pan, Keqiang Sun, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *arXiv preprint arXiv:2307.00716*, 2023. 2

[6] Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024. 2

[7] Christoph Schuhmann and Romain Beaumont. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. 2

[8] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu

Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. pages 2556–2565, 2018. 2

[9] 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. 2

[10] 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

[11] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2

[12] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14022–14032, 2024. 2