Scaling the Codebook Size of VQGAN to 100,000 with a Utilization Rate of 99%

Yanik Künzi

March 4th 2025, Zurich

Motivation

- LLMs successful for text generation
- Autoregressive models on tokens
- Can we do the same for images?

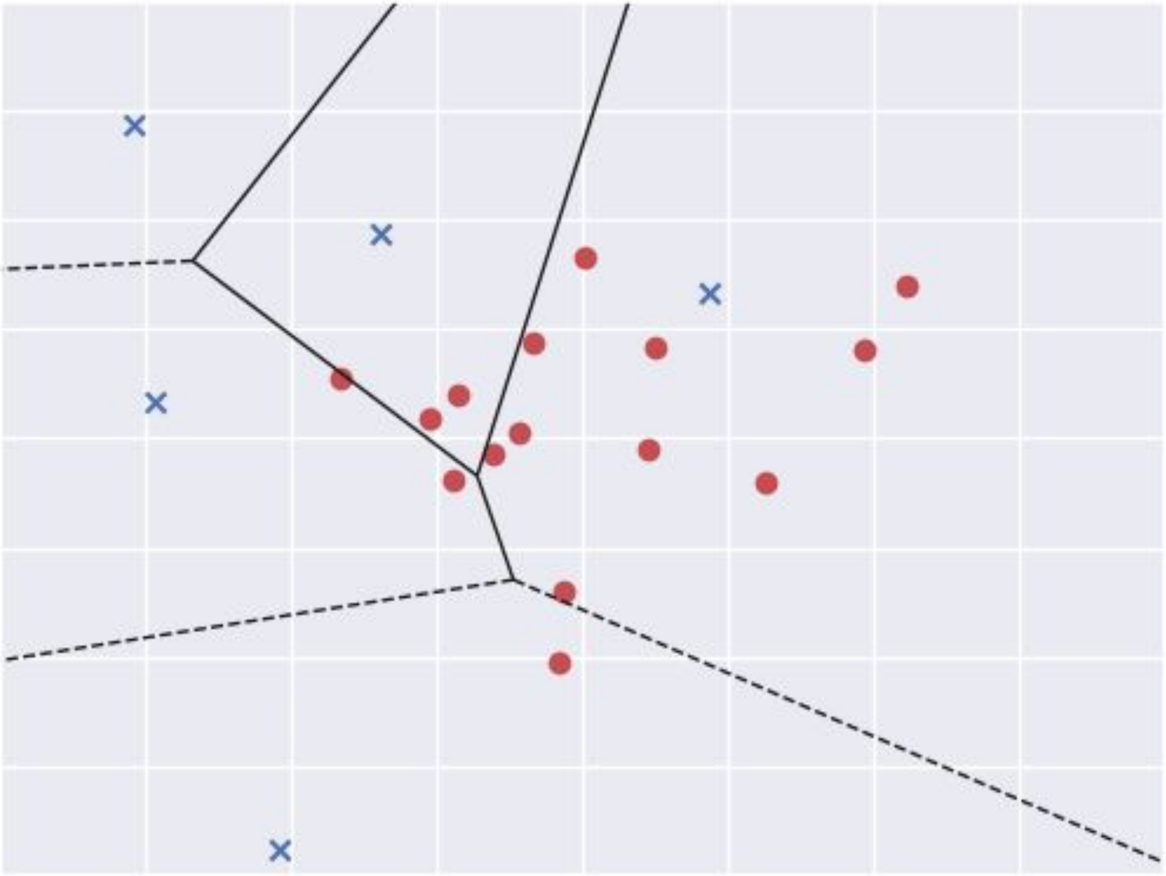
Motivation

- Problem: Images are continuous
- Solution: Discretize![1]
- Tokens used to describe images
- Bigger vocabulary results in more expressive model

[1] van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/1711.00937

Vector Quantization

- Discretization of latent space
- Compression
- Autoregressive models



VQGAN[2]

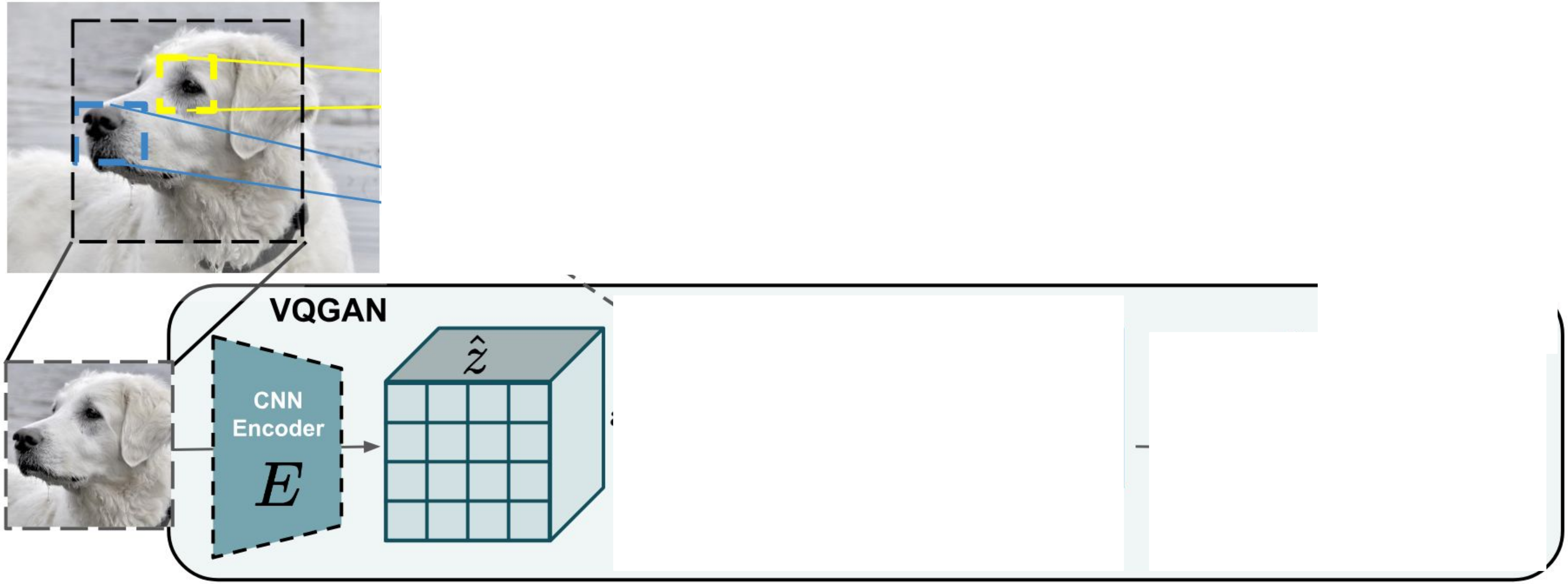


Figure 2 [2]

[2] Esser, P., Rombach, R., & Ommer, B. (2020). Taming Transformers for High-Resolution Image Synthesis. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2012.09841

Quantization Loss in VQGAN[2]

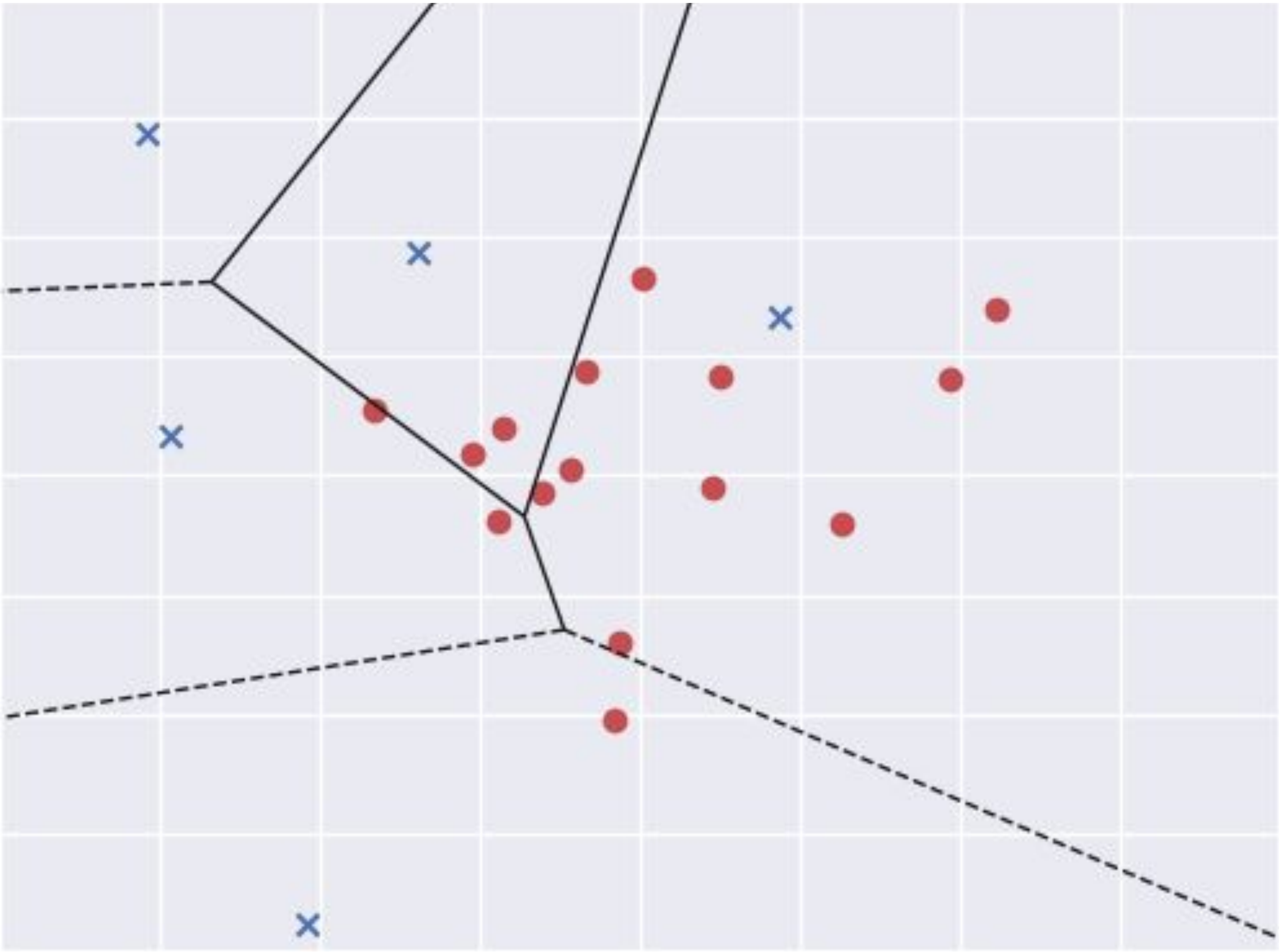
- Straight-through gradient estimation

$$\mathcal{L}_{VQ}(E, G, \mathcal{Z}) = \underbrace{\|x - \hat{x}\|^2}_{\text{Reconstruction loss}} + \underbrace{\|\mathrm{sg}[E(x)] - z_q\|^2}_{\text{Codebook loss}} + \underbrace{\|\mathrm{sg}[z_q] - E(x)\|^2}_{\text{Commitment loss}}$$

[2] Esser, P., Rombach, R., & Ommer, B. (2020). Taming Transformers for High-Resolution Image Synthesis. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2012.09841

Codebook Collapse

- Few updates to codebook per batch
- Many codebook vectors rarely used



Previous Approaches to Counteract Codebook Collapse

- Factorized Codes (FC)[3]
- Exponential Moving Average (EMA)[1],[4]

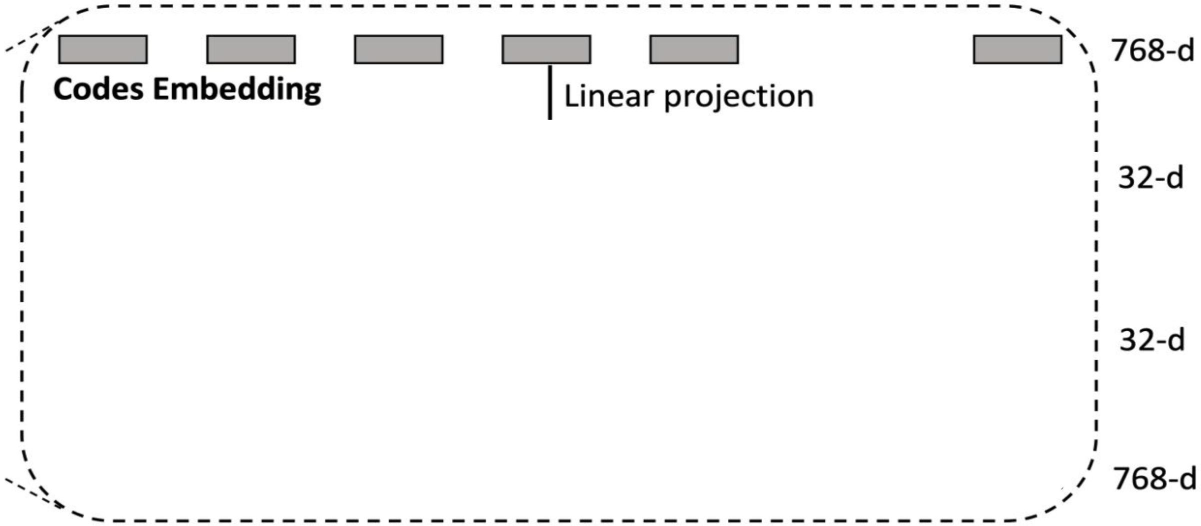


Figure 4 (cropped) [3]

[1] van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/1711.00937
[3] Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., … Wu, Y. (2022). Vector-quantized Image Modeling with Improved VQGAN. *arXiv [Cs.CV]*. Retrieved from http://arxiv.org/abs/2110.04627
[4] Razavi, A., van den Oord, A., & Vinyals, O. (2019). Generating Diverse High-Fidelity Images with VQ-VAE-2. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/1906.00446

Contrastive Language-Image Pre-training (CLIP)[6]

- Contrastive Learning
- Matching image to text description
- Image encoder based on ViT

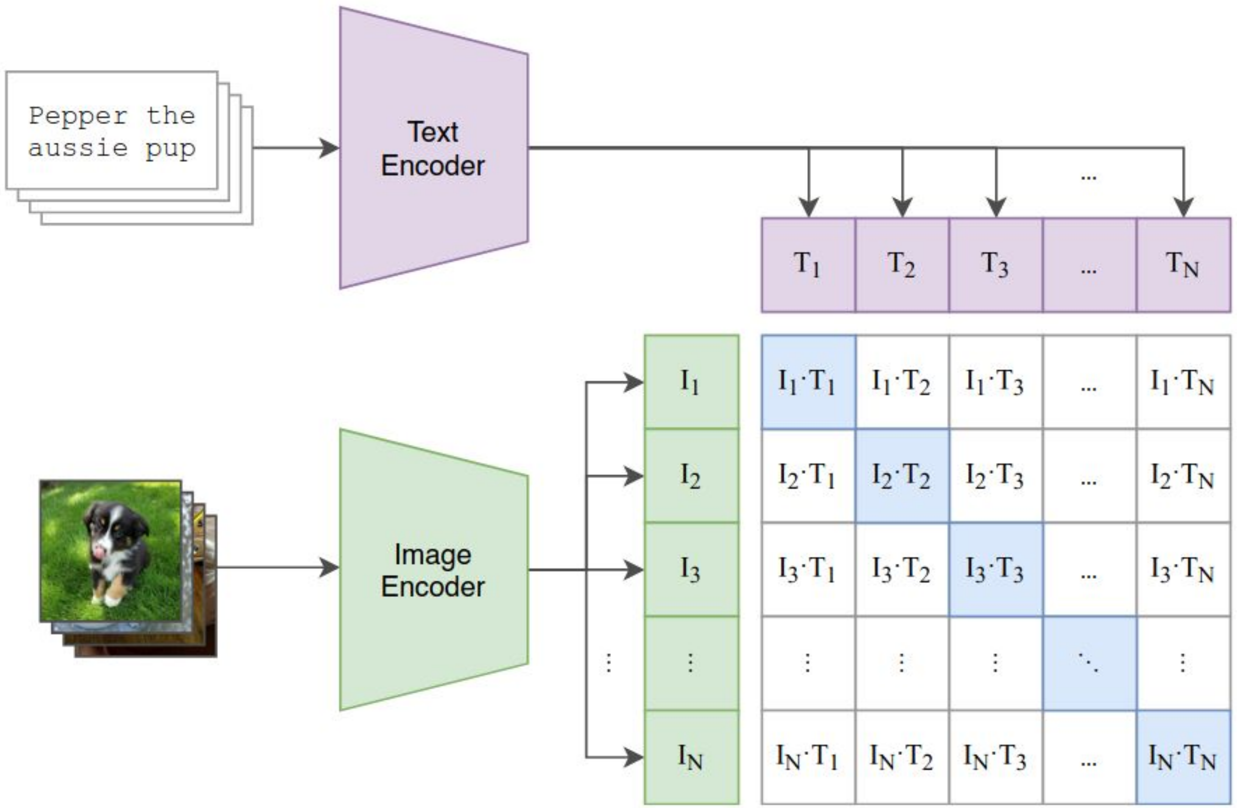
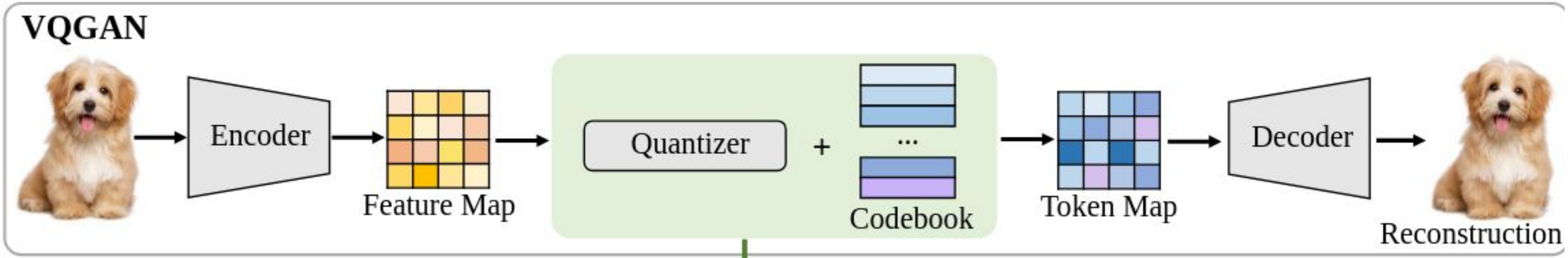

Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
⋮ ⋮ ⋮ ⋱ ⋮
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$

Figure 1 (1) [6]

[6] *Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv [Cs.CV]. Retrieved from* http://arxiv.org/abs/2103.00020

Recap



(a) Architecture of VQGAN.

Figure 2 (cropped) [5]

VQGAN-LC[5]

- Static codebook, learned projection
- Patch-wise CLIP[6] embeddings
- K-means clustering

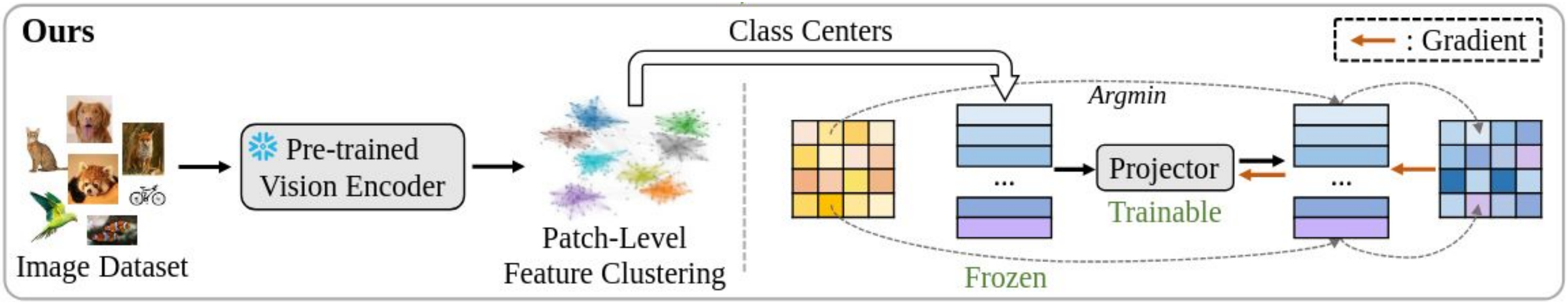
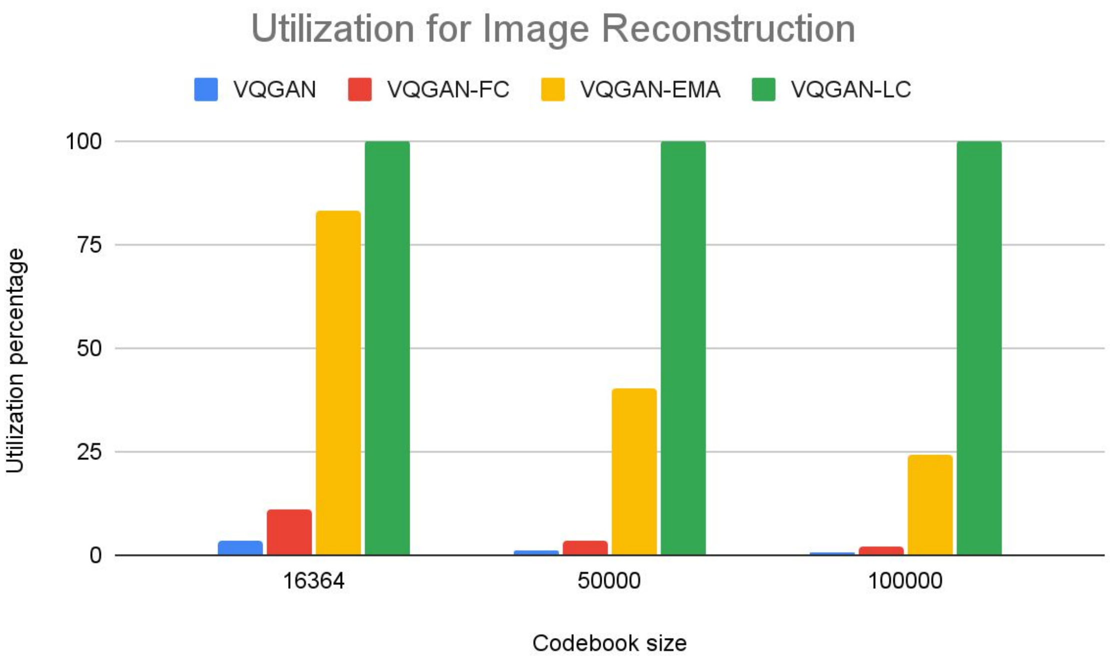
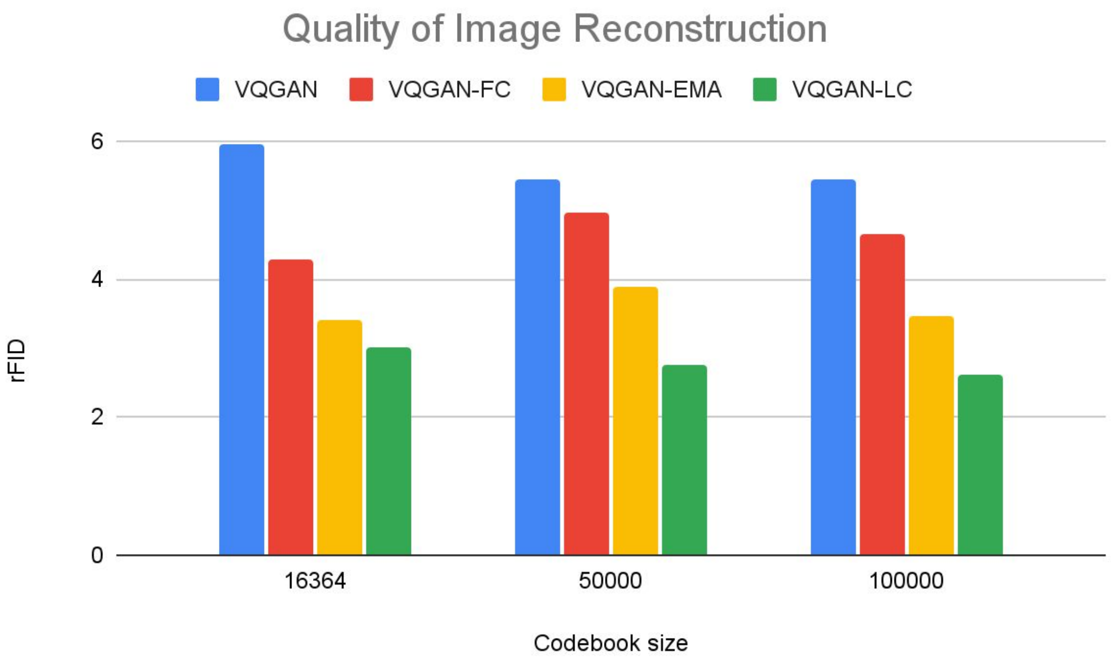


Figure 2 (cropped) [5]

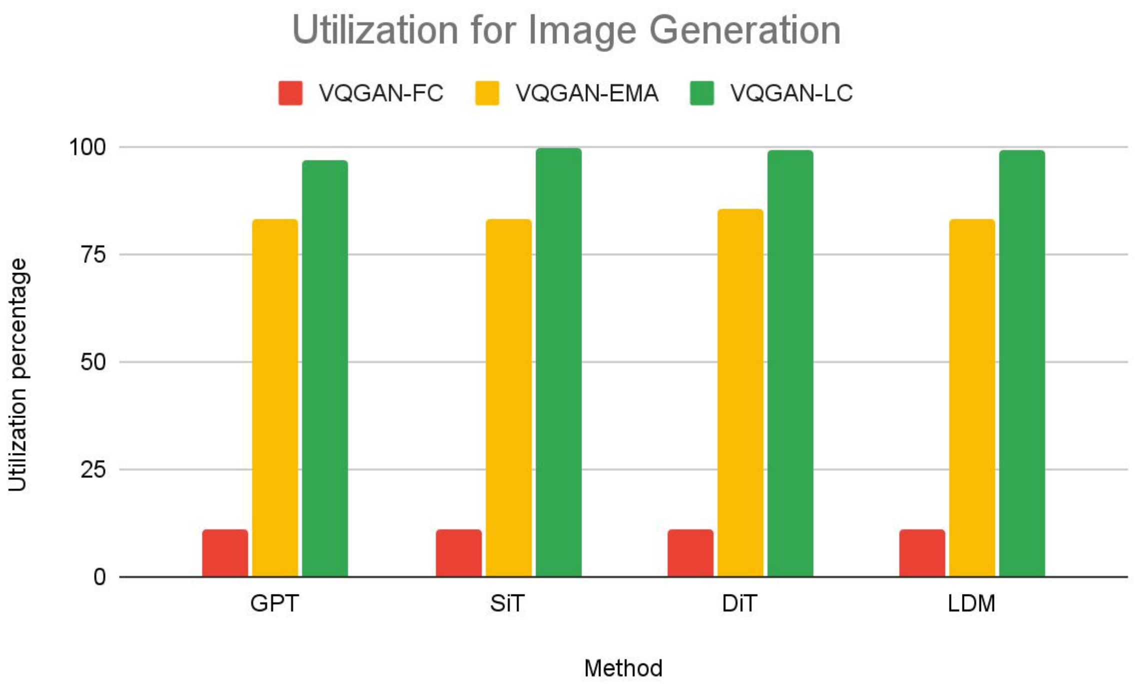
[5] Zhu, L., Wei, F., Lu, Y., & Chen, D. (2024). Scaling the Codebook Size of VQGAN to 100,000 with a Utilization Rate of 99%. *arXiv [Cs.CV]*. Retrieved from http://arxiv.org/abs/2406.11837
[6] *Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv [Cs.CV]. Retrieved from* http://arxiv.org/abs/2103.00020

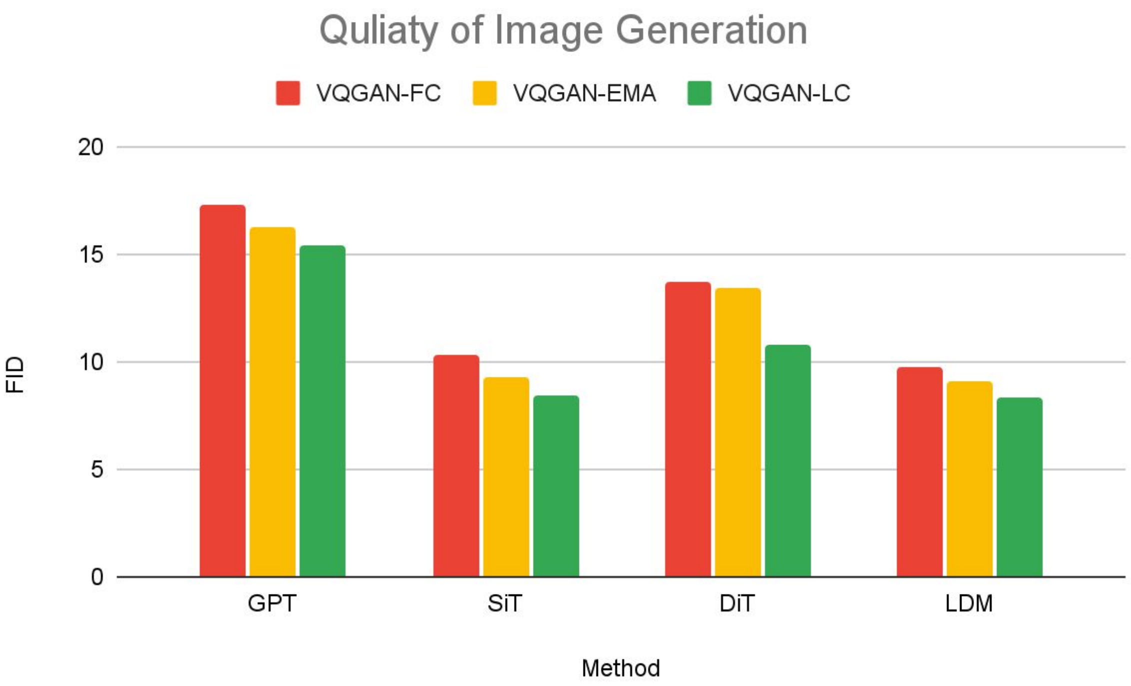
Results


Utilization for Image Reconstruction
VQGAN
VQGAN-FC
VQGAN-EMA
VQGAN-LC
Utilization percentage
100
75
50
25
0
16364
50000
100000
Codebook size


Quality of Image Reconstruction
VQGAN
VQGAN-FC
VQGAN-EMA
VQGAN-LC
rFID
6
4
2
0
16364
50000
100000
Codebook size

Results


Utilization for Image Generation
VQGAN-FC
VQGAN-EMA
VQGAN-LC
100
75
50
25
0
Utilization percentage
GPT
SiT
DiT
LDM
Method

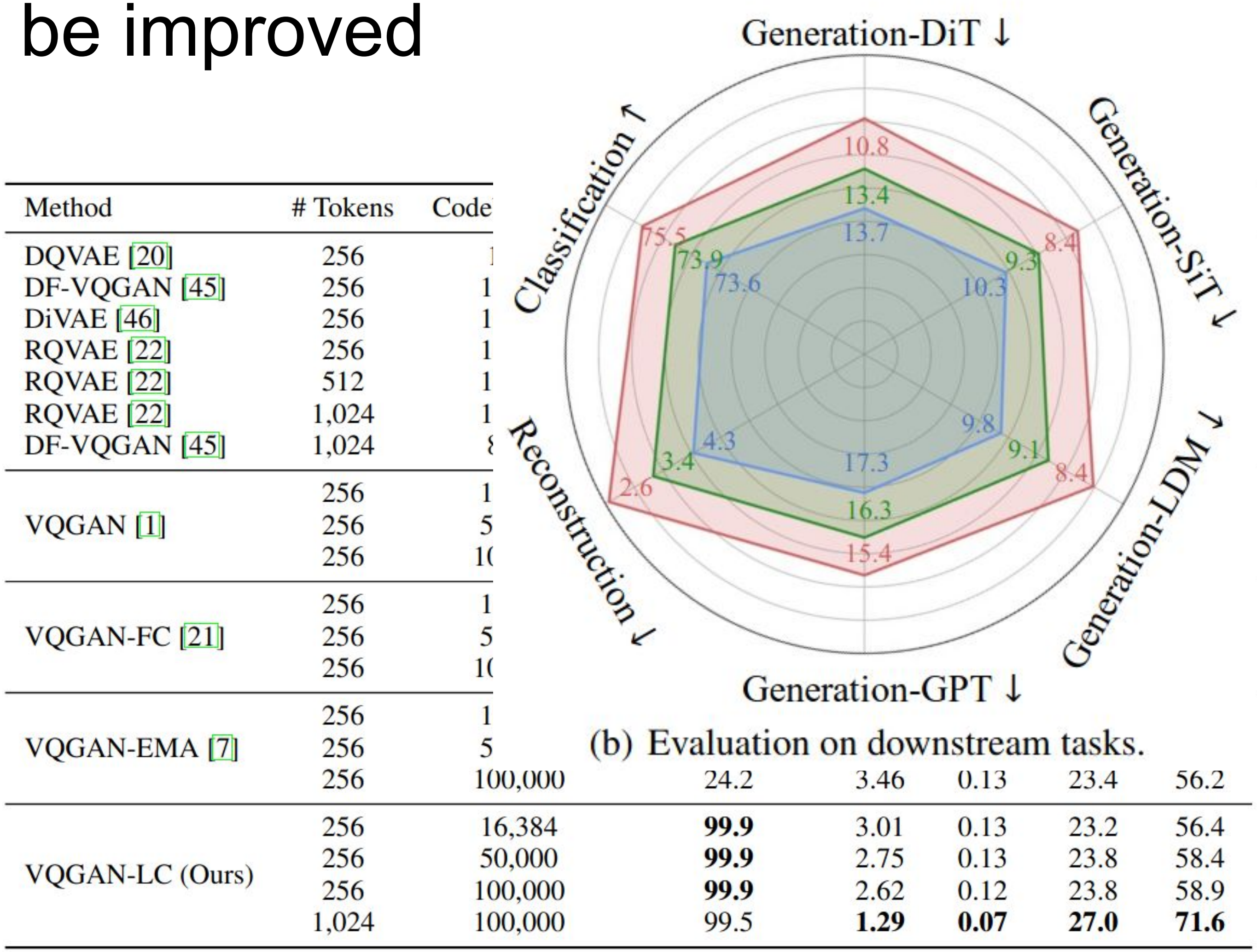

Quliaty of Image Generation
VQGAN-FC
VQGAN-EMA
VQGAN-LC
20
15
10
5
0
FID
GPT
SiT
DiT
LDM
Method

Conclusion

- Interesting and insightful paper
- Easily usable by other methods using VQ
- Effectiveness empirically demonstrated

Conclusion

- Result reporting could be improved
- Experiment setup
- Why does it work?



(b) Evaluation on downstream tasks.

Method	# Tokens	Code					
DQVAE [20]	256	1					
DF-VQGAN [45]	256	1					
DiVAE [46]	256	1					
RQVAE [22]	256	1					
RQVAE [22]	512	1					
RQVAE [22]	1,024	1					
DF-VQGAN [45]	1,024	8					
VQGAN [1]	256	1					
	256	5					
	256	10					
VQGAN-FC [21]	256	1					
	256	5					
	256	10					
VQGAN-EMA [7]	256	1					
	256	5					
	256	100,000	24.2	3.46	0.13	23.4	56.2
VQGAN-LC (Ours)	256	16,384	**99.9**	3.01	0.13	23.2	56.4
	256	50,000	**99.9**	2.75	0.13	23.8	58.4
	256	100,000	**99.9**	2.62	0.12	23.8	58.9
	1,024	100,000	99.5	**1.29**	**0.07**	**27.0**	**71.6**

Discussion

Discussion Icon Vectors by Vecteezy

References

[1] van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/1711.00937

[2] Esser, P., Rombach, R., & Ommer, B. (2020). Taming Transformers for High-Resolution Image Synthesis. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2012.09841

[3] Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., … Wu, Y. (2022). Vector-quantized Image Modeling with Improved VQGAN. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2110.04627

[4] Razavi, A., van den Oord, A., & Vinyals, O. (2019). Generating Diverse High-Fidelity Images with VQ-VAE-2. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/1906.00446

[5] Zhu, L., Wei, F., Lu, Y., & Chen, D. (2024). Scaling the Codebook Size of VQGAN to 100,000 with a Utilization Rate of 99%. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2406.11837

[6] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv [Cs.CV]. Retrieved from http://arxiv.org/abs/2103.00020