
Appendix of Dimensional Collapse in VQVAEs: Evidence and Remedies

Anonymous Author(s)

Affiliation

Address

email

1	Contents	
2	A Related work	1
3	A.1 VQVAE and its variants	1
4	A.2 Multi-codebook methods	2
5	A.3 Codebook collapse and dimensional collapse	2
6	B Case study: visualization of each effective dimension	2
7	B.1 Purpose	2
8	B.2 Methodology	2
9	B.3 Visualization setup	3
10	B.4 Findings	3
11	C Implementation details	3
12	C.1 Codebase	3
13	C.2 Training infrastructure	5
14	C.3 Datasets	5
15	C.4 Hyperparameters for CelebA and CIFAR-10	5
16	C.5 Hyperparameters for ImageNet	5
17	C.6 Architecture Details for CelebA and CIFAR-10	5
18	C.7 Architecture Details for ImageNet	6
19	C.8 GPU usage	6
20	D Additional results for correlations	6
21	A Related work	
22	A.1 VQVAE and its variants	
23	Vector-Quantized Variational Autoencoders (VQVAEs) [11] have emerged as a cornerstone in discrete	
24	representation learning, enabling high-fidelity generative modeling across vision [9], audio [18],	
25	video [13], and structural biology [12, 3]. The central idea is to discretize a continuous latent space	
26	into learnable embeddings, typically via nearest-neighbor assignment to a codebook.	

Successive works have aimed to scale and stabilize this framework. VQVAE-2 [9] introduced hierarchical quantization, while VQVAE [5] proposed recursive residual quantization. Innovations like grouped quantization [15], the rotation trick [2], and joint codebook updates in SimVQ [19] further refine training dynamics and representation efficiency. Lookup-Free Quantization (LFQ) [17] and Finite Scalar Quantization (FSQ) [8] offer scalar-level discretization with extremely low latent dimensions (e.g., 6–8), improving code utilization but at the cost of expressivity.

Despite extensive progress in improving generation quality and mitigating codebook collapse, our work focuses on dimensional collapse, a more subtle and previously overlooked issue.

A.2 Multi-codebook methods

Several recent works have proposed architectural strategies that superficially resemble our approach by partitioning the latent space into independently quantized subspaces. For example, XQ-GAN and IMAGEFOLDER [7, 6] adopt *product quantization* to split the latent space into low-dimensional branches, each quantized separately. However, their primary goal is to align tokens with spatial or semantic features to facilitate autoregressive modeling and reduce sequence length, rather than addressing the underlying structure of the latent space. In contrast, our work is driven by a systematic analysis of *dimensional collapse* in VQVAEs, where high-dimensional embeddings are compressed into narrow subspaces, limiting expressivity.

Other methods, such as RQVAE [5] and grouped quantization [15], improve expressivity by introducing multi-stage or groupwise quantization to better approximate the encoder outputs. Visual AutoRegressive modeling (VAR) [10] further complements this line by redefining generation as a coarse-to-fine, multi-scale prediction task, achieving state-of-the-art results in efficiency and quality. While these approaches adopt multi-codebook designs to enhance generation, our proposed DCVQ leverages them as a means to address latent space underutilization.

A.3 Codebook collapse and dimensional collapse

Codebook collapse—the underutilization of codebook entries—has long been recognized as a key limitation in VQVAEs [11]. Numerous approaches have been proposed to mitigate this issue, including dead-code replacement [18], exponential moving average (EMA) updates [9], and architectural modifications that improve gradient flow and codebook dynamics [19, 2]. Interestingly, several works [8, 16] have shown that reducing the embedding dimension can lead to improved code usage, suggesting a link between codebook utilization and the underlying structure of the latent space.

However, most of these efforts focus on usage statistics such as codebook perplexity or entropy, without examining the geometry of the latent space itself. A more fundamental issue—*dimensional collapse*—has only recently begun to attract attention. This refers to the observation that, despite operating in high-dimensional latent spaces, VQVAEs often encode information in a much lower-dimensional subspace. While the term "dimensional collapse" has previously been used in contrastive self-supervised learning [4] to describe degeneracy in feature representations, its occurrence in VQ-based generative models has not been systematically studied.

B Case study: visualization of each effective dimension

B.1 Purpose

This section aims to provide an intuitive understanding of what is encoded in each effective dimension of the VQVAE codebook. By visualizing the role of individual principal components (PCs), we explore how semantic information is distributed across the most significant axes of variation in the learned discrete latent space.

B.2 Methodology

We perform principal component analysis (PCA) on the codebook embeddings to identify the most informative directions. Specifically, we construct a series of reduced codebooks that contain only the first k principal components ($k = 1$ to 6). Each level of this PCA-reduced codebook represents an increasingly complete approximation of the full latent space.

75 For a given input image, we extract its discrete token indices using the encoder. We then replace the
76 standard codebook embeddings with their PCA-reduced counterparts before decoding. This allows
77 us to visualize how reconstructions evolve as more effective dimensions are included.

78 We conduct the experiment on VQGAN. Notably, our analysis reveals that over 99% of the variance
79 in the codebook is captured by the first 4 principal components. Thus, we focus on up to the first 6
80 PCs to analyze marginal contributions beyond the main informative axes.

81 B.3 Visualization setup

82 The visualizations are arranged in two rows for each example:

- 83 • **Top row:** reconstructions using progressively richer PCA embeddings—from PC1 only up
84 to PC6. The baseline input and full VQGAN reconstruction are also included for reference.
- 85 • **Bottom row:** the latent activations corresponding to each individual PC, rendered as
86 grayscale images, to show spatial distribution and intensity.

87 B.4 Findings

88 From the visualizations in Figure 1, we observe the following:

- 89 • **PC1** captures the global layout or coarse structure of the image. For example, the silhouette
90 or contour of a person is often visible even with only the first PC.
- 91 • **PC2 and PC3** encode finer details such as texture and local contrast, improving visual
92 fidelity and definition.
- 93 • **PC4 and PC5** introduce color and tone variations, filling in stylistic and chromatic informa-
94 tion.
- 95 • **PC6** contributes marginally, with minimal visible structure in either the reconstruction or
96 the latent visualization, indicating that it carries negligible additional semantic content.

97 This experiment supports the interpretation that VQVAE codebooks suffer from a dimensional bottle-
98 neck, where only a few directions in the latent space carry most of the meaningful information. This
99 visualization method complements quantitative measures by providing direct human-interpretable
100 evidence of dimensional collapse.

101 C Implementation details

102 C.1 Codebase

103 Our implementation is based on several publicly available GitHub repositories:

- 104 • lucidrains/vector-quantize-pytorch [14] (MIT License)
- 105 • kakaobrain/rq-vae-transformer [5] (Apache 2.0 License)
- 106 • cfifty/rotation_trick [2] (No explicit license; used under academic fair use as per
107 official ICLR release)
- 108 • thuanz123/enhancing-transformers (MIT License)
- 109 • CompVis/taming-transformers [1] (MIT License)

110 Specifically, the training script for VQVAE on CelebA and CIFAR-10 was adapted
111 from cfifty/rotation_trick, and the ViT architecture implementation was taken from
112 thuanz123/enhancing-transformers. The CNN-based encoder/decoder backbone was based
113 on CompVis/taming-transformers. The RQ-VAE training procedure and the training script for
114 ImageNet were derived from kakaobrain/rq-vae-transformer. We re-trained RQVAE on our
115 own setup and observed that the performance closely matched the results reported in their paper.



Figure 1: Reconstructions and principal component visualizations for selected examples.

116 C.2 Training infrastructure

117 All experiments were conducted on NVIDIA A100 GPUs with 80GB of memory. Training was
 118 performed on a single GPU per run. Although A100s were used, our models did not fully utilize
 119 the available memory, making it feasible to train them on lower-end GPUs as well. We employed
 120 Weights and Biases (wandb) for experiment tracking and conducted uniform random hyperparameter
 121 sweeps. Reported training times (in GPU-hours) are taken directly from the wandb platform.

122 C.3 Datasets

123 We conduct experiments on three datasets: ImageNet-256, CelebA-64, and CIFAR-10. ImageNet
 124 images are resized to 256×256 resolution. CelebA images are resized to 64×64 , and CIFAR-10
 125 consists of 32×32 images. These datasets cover a range of visual complexity and resolution, allowing
 126 us to evaluate the model’s behavior across diverse settings.

127 C.4 Hyperparameters for CelebA and CIFAR-10

128 For the experiments on CelebA and CIFAR-10 presented in Section 3 and Section 5 of the main text,
 129 we conducted hyperparameter searches over selected variables as described in Table 1 (main text).
 130 Other hyperparameters were held fixed across runs. The full configuration used in these experiments
 131 is detailed in Table 1. We use a custom learning rate scheduler that linearly warms up, then decays
 132 via cosine annealing to a fixed minimum learning rate.

Hyperparameter	Value
Batch size	32
Optimizer	Adam
Learning rate	0.0001
Weight decay	0.0001
Epochs	100
Model type	VQVAE with rotation trick
Codebook type	cosine
Warmup iterations	3000
Decay iterations	50000
Stochastic sampling of codes	False
Dropout	0
Seed	0

Table 1: Fixed hyperparameter configuration for the experiments in Section 3.

133 C.5 Hyperparameters for ImageNet

134 For the ImageNet experiments in Section 5, we adopt the default hyperparameters provided by the
 135 RQVAE codebase, which could be found in Table 2.

136 C.6 Architecture Details for CelebA and CIFAR-10

137 We use a standard VQ-VAE framework composed of an encoder, a vector quantizer, and a decoder.
 138 The encoder and decoder architectures differ depending on the model configuration:

- 139 • **CNN-based:** We adopt a convolutional encoder-decoder architecture inspired by VQGAN.
 140 The model downsamples the input by a factor of 4 or 16, using 3 or 5 convolutional blocks
 141 respectively. Each block contains 2 residual layers, and the number of channels is determined
 142 by a base width of 128 multiplied by a channel multiplier. The channel multiplier is set to
 143 $[1, 2, 4]$ for a downsampling factor of 4, and $[1, 1, 2, 2, 4]$ for a factor of 16. The encoder
 144 output is projected to a fixed dimensionality of 256 using a convolutional layer.
- 145 • **ViT-based:** We employ a Vision Transformer (ViT) encoder-decoder architecture for image
 146 tokenization and reconstruction. The input image is partitioned into non-overlapping patches
 147 of size $f \times f$, where f denotes the patch size (e.g., $f = 4$ or 16). Each patch is embedded

Hyperparameter	Value
Batch size	32
Epochs	10
Optimizer	Adam
Learning rate	4.0e-05
Betas	(0.5, 0.9)
Weight decay	0.0
Learning rate scheduler	fixed learning rate
Discriminator loss	Hinge
Discriminator start epoch	0
Discriminator weight	0.75
Generator loss	Vanilla
Perceptual loss weight	1.0

Table 2: Non-architectural hyperparameters used in the experiments in Section 5.

via a convolutional projection into a 768-dimensional vector. Fixed 2D sinusoidal positional embeddings are added to the patch sequence, which is then processed by a Transformer encoder comprising 12 layers, each with 12 self-attention heads and an MLP of width 3072. The encoded patch representations are quantized and then passed to a symmetric Transformer decoder to reconstruct the image. This architecture enables global context modeling and adaptive spatial compression.

In both architectures, a linear projection maps the encoder output to the quantizer input space, which allows tunable codebook dimensionality. After quantization, the embeddings are projected back before being passed to the decoder. The quantizer is based on cosine similarity and is updated using exponential moving average (EMA).

C.7 Architecture Details for ImageNet

For experiments on ImageNet, the settings are similar but we only have CNN-based backbone with $f = 8$. We follow the RQVAE architecture with the following configuration:

- **CNN-based (ImageNet):** We utilize a deep convolutional encoder-decoder architecture designed for high-resolution inputs. The encoder comprises 6 stages, each with 2 residual blocks, totaling 12 residual layers. A base channel size of 128 is used with channel multipliers $[1, 1, 2, 2, 4, 4]$, reaching up to 512 channels in deeper layers. The spatial resolution is reduced by a factor of 8. The encoder output is projected to a 256-dimensional latent space. A commitment loss with weight 0.25 is applied to encourage consistency between the encoder output and the quantized representation.

C.8 GPU usage

Table 3 summarizes the GPU usage for the experiment in Section 3. The GPUs we use were A100 80GB.

D Additional results for correlations

In Table 2 of the main text, we report the average Pearson correlation between effective dimensionality and various hyperparameters, aggregated over different background dimensionalities. Figure 2 provides a more detailed breakdown of these correlations across individual datasets, architecture, scale factors, and background dimensions.

Sweep Name	Dataset & Model	GPU Days
sweep_celeba_cnn_f=4	CelebA + CNN	47
sweep_celeba_vit_f=4	CelebA + ViT	87
sweep_cifar_cnn_f=4	CIFAR-10 + CNN	15
sweep_cifar_vit_f=4	CIFAR-10 + ViT	19
sweep_celeba_cnn_f=16	CelebA + CNN	34
sweep_celeba_vit_f=16	CelebA + ViT	32
sweep_cifar_cnn_f=16	CIFAR-10 + CNN	15
sweep_cifar_vit_f=16	CIFAR-10 + ViT	17

Table 3: GPU time (in GPU-days) consumed by each sweep. Each sweep includes 64 individual runs, totaling 512 runs across all configurations. The sweeps vary by dataset (CelebA or CIFAR-10), model type (CNN or ViT), and downsampling factor ($f = 4$ or $f = 16$).

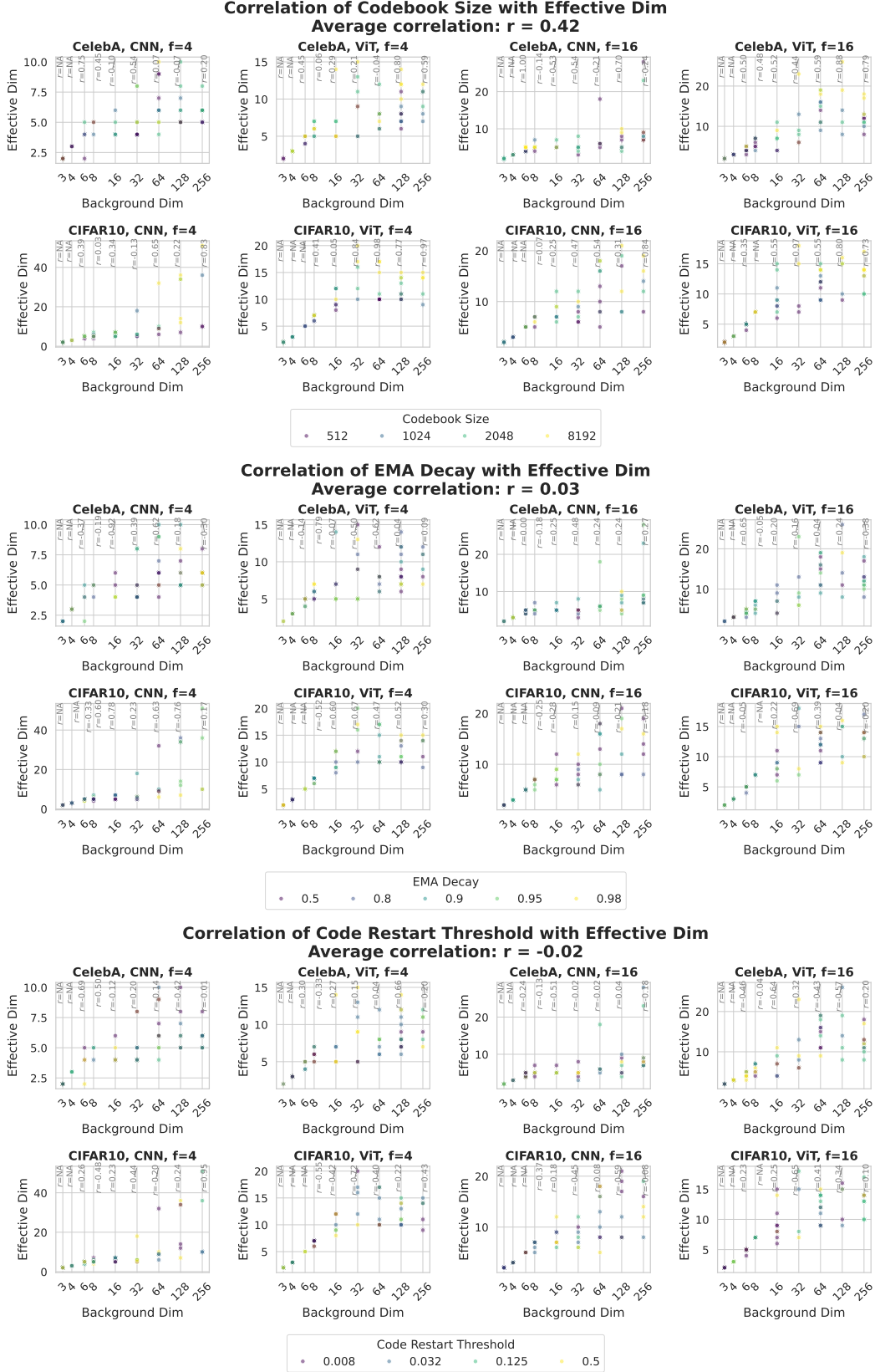


Figure 2: Correlation between effective dimensionality and various hyperparameters (part 1).

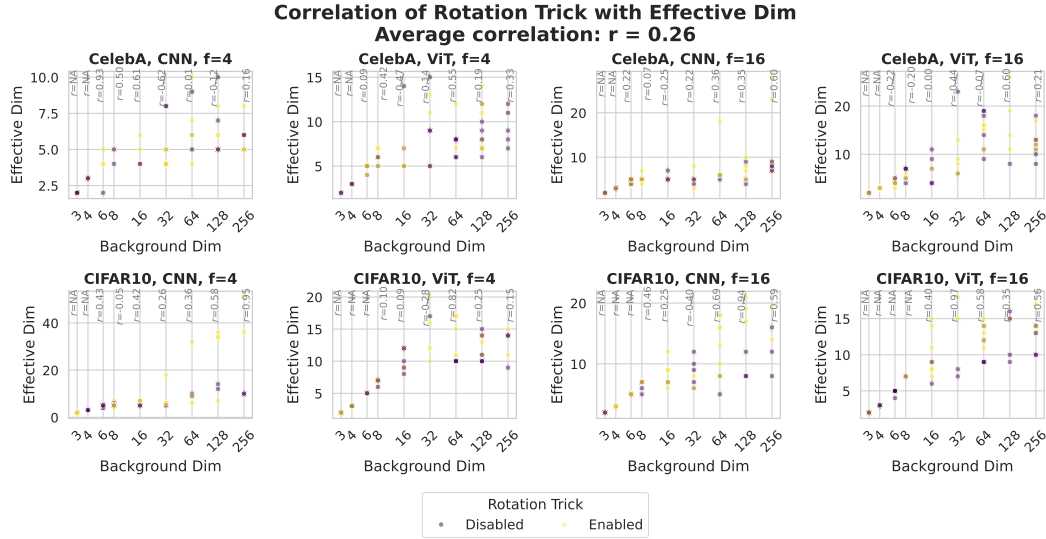
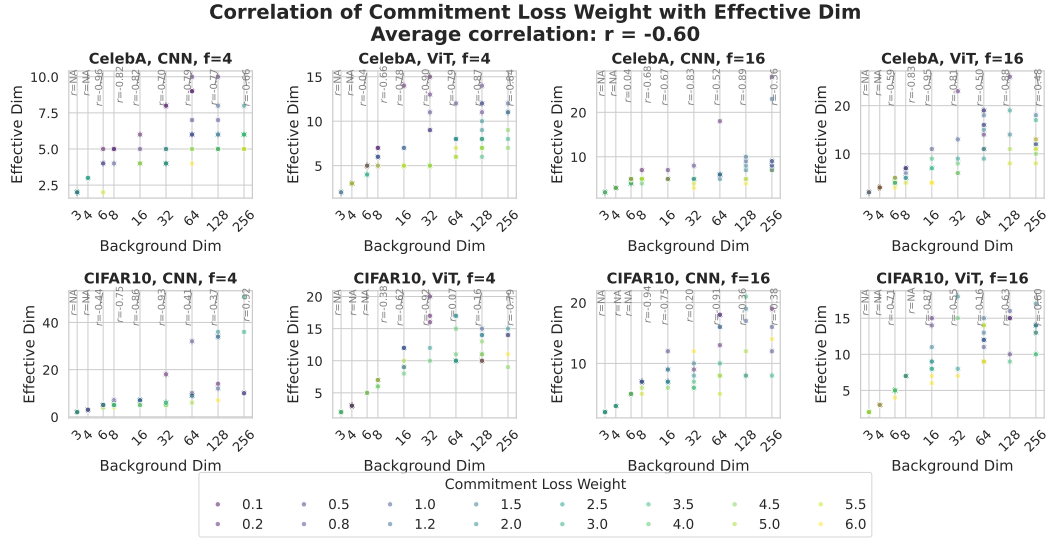


Figure 2 (continued): Additional correlation plots.

References

- [1] 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.
- [2] Christopher Fifty, Ronald G Junkins, Dennis Duan, Aniketh Iyengar, Jerry W Liu, Ehsan Amid, Sebastian Thrun, and Christopher Ré. Restructuring vector quantization with the rotation trick. *ArXiv*, abs/2410.06424, 2024.
- [3] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, page eads0018, 2025.
- [4] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.
- [5] 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.
- [6] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Bhiksha Raj, and Zhe Lin. Imagefolder: Autoregressive image generation with folded tokens. *arXiv preprint arXiv:2410.01756*, 2024.
- [7] Xiang Li, Kai Qiu, Hao Chen, Jason Kuen, Jiuxiang Gu, Jindong Wang, Zhe Lin, and Bhiksha Raj. Xq-gan: An open-source image tokenization framework for autoregressive generation. *arXiv preprint arXiv:2412.01762*, 2024.
- [8] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple, 2023.
- [9] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019.
- [10] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. 2024.
- [11] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [12] Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Foldseek: fast and accurate protein structure search. *Biorxiv*, pages 2022–02, 2022.
- [13] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vqvae. *arXiv preprint arXiv:2103.01950*, 2021.
- [14] Phil Wang. vector-quantize-pytorch. <https://github.com/lucidrains/vector-quantize-pytorch>, 2021. Accessed: 2025-04-28.
- [15] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. 2023.
- [16] 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.
- [17] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10459–10469, 2023.

- 223 [18] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi.
224 Soundstream: An end-to-end neural audio codec, 2021.
- 225 [19] Yongxin Zhu, Bocheng Li, Yifei Xin, and Linli Xu. Addressing representation collapse in
226 vector quantized models with one linear layer. 2024.

227 NeurIPS Paper Checklist

228 The checklist is designed to encourage best practices for responsible machine learning research,
229 addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove
230 the checklist: **The papers not including the checklist will be desk rejected.** The checklist should
231 follow the references and follow the (optional) supplemental material. The checklist does NOT count
232 towards the page limit.

233 Please read the checklist guidelines carefully for information on how to answer these questions. For
234 each question in the checklist:

- 235 • You should answer [Yes], [No], or [NA].
- 236 • [NA] means either that the question is Not Applicable for that particular paper or the
237 relevant information is Not Available.
- 238 • Please provide a short (1–2 sentence) justification right after your answer (even for NA).

239 **The checklist answers are an integral part of your paper submission.** They are visible to the
240 reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it
241 (after eventual revisions) with the final version of your paper, and its final version will be published
242 with the paper.

243 The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation.
244 While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a
245 proper justification is given (e.g., "error bars are not reported because it would be too computationally
246 expensive" or "we were unable to find the license for the dataset we used"). In general, answering
247 "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we
248 acknowledge that the true answer is often more nuanced, so please just use your best judgment and
249 write a justification to elaborate. All supporting evidence can appear either in the main paper or the
250 supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification
251 please point to the section(s) where related material for the question can be found.

252 IMPORTANT, please:

- 253 • **Delete this instruction block, but keep the section heading “NeurIPS Paper Checklist”,**
- 254 • **Keep the checklist subsection headings, questions/answers and guidelines below.**
- 255 • **Do not modify the questions and only use the provided macros for your answers.**

256 1. Claims

257 Question: Do the main claims made in the abstract and introduction accurately reflect the
258 paper’s contributions and scope?

259 Answer: [Yes]

260 Justification: The abstract and introduction accurately summarize the paper’s key contri-
261 butions: (1) the discovery and analysis of dimensional collapse in VQVAEs, and (2) the
262 proposal of Divide-and-Conquer VQ (DCVQ) to overcome this limitation. These contribu-
263 tions are supported by empirical analysis in Section 3 and the proposed method and results
264 in Section 4 and 5.

265 Guidelines:

- 266 • The answer NA means that the abstract and introduction do not include the claims
267 made in the paper.
- 268 • The abstract and/or introduction should clearly state the claims made, including the
269 contributions made in the paper and important assumptions and limitations. A No or
270 NA answer to this question will not be perceived well by the reviewers.

- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: A dedicated "Limitations" paragraph is included at the end of Section 6. We acknowledge that our study is primarily empirical and does not provide a theoretical explanation for codebook collapse. We also note that the applicability of DCVQ to modalities beyond vision remains unexplored and is left for future work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper is primarily empirical and does not include formal theoretical results, assumptions, or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.

- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed descriptions of model architecture, training procedures, datasets, and evaluation metrics in Section 3 and 4 and Appendix. These include hyperparameter settings, codebook configuration, and data preprocessing steps, sufficient to reproduce the main results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will include anonymized code and detailed instructions to reproduce our main experimental results in the supplemental material before the final appendix submission deadline. The code will cover both our proposed method (DCVQ) and baseline comparisons.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide the key experimental settings in the main paper (Section 3 and 5), including datasets used, model configurations, and evaluation metrics. Full training and testing details, including data splits, hyperparameters, and optimizer configurations, will be included in the submitted anonymized code and supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: While we do not explicitly report error bars or confidence intervals, we conduct a large-scale controlled experiment involving 512 independently trained VQVAEs across a diverse set of conditions (Section 3.2). This extensive coverage enables us to systematically analyze the statistical trends and variability of dimensional collapse. The number of runs and control over variables provide strong empirical grounding for the observed U-shaped performance patterns.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We will provide full details on compute resources, training time per experiment, and total compute cost in the Appendix of the supplementary material. All experiments were conducted on NVIDIA A100 GPUs with 80GB memory, though the experiments can be reproduced on GPUs with 40GB memory with mild modifications.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our research fully complies. Our experiments use only publicly available datasets (CIFAR-10, CelebA, and ImageNet) under standard preprocessing. No human subjects, sensitive data, or personally identifiable information are involved.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: Our work focuses on understanding and improving the latent representation properties of VQVAEs, a foundational generative modeling technique. We do not target any specific downstream application or deployment. While improved generative models could, in theory, contribute to misuse, our contribution is abstract and technical in nature. Therefore, we do not explicitly discuss broader societal impacts in the paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper does not release any models or datasets that pose a high risk of misuse. We work with standard, publicly available datasets (CIFAR-10, CelebA, ImageNet) and do not release any pretrained generative models or scraped data. Therefore, no special safeguards are necessary.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We use only publicly available datasets and open-source code, and all assets are properly credited and cited in the paper. **Datasets:** We use CIFAR-10 and CelebA, both publicly available for academic use, and ImageNet (under its standard academic research terms). **Codebases:** (1) lucidrains/vector-quantize-pytorch: MIT License (2) kakaobrain/rq-vae-transformer: Apache 2.0 License (3) cfifty/rotation_trick:

No explicit license, but official code release accompanying an ICLR paper; used under academic fair use (4) thuanz123/enhancing-transformers: MIT License (5) CompVis/taming-transformers: MIT License. We cite all these assets in the main paper and respect their licenses and terms of use. No proprietary or restricted resources are used.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce new assets such as datasets or pretrained models. We propose a new method (DCVQ), and the accompanying code will be released in anonymized form as supplemental material for reproducibility purposes, but it is not considered a structured new asset release.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve any crowdsourcing experiments or research with human subjects. All experiments are conducted on publicly available datasets, and no new data was collected from human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve research with human subjects or crowdsourced data collection. All experiments are conducted on publicly available datasets (CIFAR-10 and CelebA) under standard academic usage terms.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We did not use any large language model (LLM) as a component of the core methods, experiments, or contributions in this research. Any language assistance was limited to standard editing and had no impact on the scientific content or originality.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.