# VQRAE: Representation Quantization Autoencoders for Multimodal Understanding, Generation and Reconstruction

Supplementary Material

In the supplementary materials, Section A provides a concise overview of the motivation underlying this study and highlights the distinctions between VQRAE and other related works; Section B illustrates the detailed training configurations of the tokenizer (Section B.1) and autoregressive models for understanding and generation (Section B.2) and their evaluation setups (Section B.3). Section C presents more qualitative results in image reconstruction (Section C.1), visual generation (Section C.2) and failure cases (Section C.3). Finally, we summarize the limitations of our work and future exploration in Section D.

## A. Overview

***Related Work.*** Reviewing the evolution of unified models, pioneering works such as Chameleon [63] and EMU-3 [73] employed VQGAN [16] as the encoder for both understanding and generation tasks, albeit at the expense of semantic understanding. The Janus series [8, 45, 83] adopted a dual-encoder architecture, utilizing a semantic encoder like CLIP [52] for multimodal understanding and a pixel encoder such as VQGAN [16] for image generation. This approach, however, hindered interaction and alignment between the two types of representations. The unified tokenizer aims to employ a single encoder-decoder framework to maintain performance on understanding tasks while enabling generation. Nevertheless, many previous methods [10, 25, 36, 43, 51, 58, 85, 89, 103] were overly complex in design or relied on simple contrastive learning to provide weak semantic supervision on latent tokens, making it difficult to achieve a satisfactory trade-off between comprehension and reconstruction. These tokenizers generally still underperform compared to classic understanding-only baseline models like LLaVA-1.5 [39]. Tar [23] and X-Omni [20] were among the first works to utilize pretrained Visual Foundation Models (VFMs) as encoders while discretizing representations to preserve multimodal understanding capabilities. Although they narrowed the performance gap of earlier discrete tokenizers [63, 73] on understanding tasks, they still suffer from quantization errors and lack inherent autoencoder properties. Recent work, RAE [104], discovered that high-dimensional ViT encoders can be directly applied to reconstruction and have been used to replace original VAEs [30] in diffusion-based generative models [48].

***Contributions.*** The contribution of our paper can be summarized in three folds: (i) We propose a vector quantization version of RAE, namely **VQRAE**, which pioneers the first attempt in unifying understanding, generation and reconstruction. (ii) VQRAE is the first unified tokenizer to produce continuous semantic features for understanding and fine-grained discrete tokens for generation and reconstruction **simultaneously**. (iii) VQRAE features the first **high-dimensional** VQ codebook (comparable to CLIP encoders) with a nearly **100%** utilization ratio, which is contrary to previous explorations in this field.

## B. Implementation Details

### B.1. Tokenizer Training Details

VQRAE is pretrained on BLIP3-o [7] open-sourced data, which consists of 27M samples recaptioned by Qwen2.5-VL-7B [2], 5M samples from CC12M [5], and 4M synthesized images from JourneyDB [59] (Table 1 & 2). We train three variants of encoder: InternViT-300M-448px [106], SigLIP2-so400m-256px [68] and SigLIP2-so400m-512px [68]. The decoder adopts the symmetric design with encoder. Specific training configurations are provided in Table 7. The code implementation is adapted from TiTok [97] repository.

### B.2. Autoregressive Training Details

For image understanding, we follow the setups in LLaVA-1.5 [39], which uses LLaVA-Pretrain-595K and LLaVA-v1.5-mix-665K for pretraining and SFT (Table 1 & 3 & 6). For visual generation, we train it on BLIP3-o [7] data and additional 80M high quality images (Table 4). We conduct ablation studies on VQ codebook on ImageNet-1K with 20 epochs training for efficiency (Table 5 & 6). Note that, the results in Table 6 are attained through training on stage 1. The LLM backbones include: Vicuna-v1.5-7B [11], Vicuna-v1.5-13B [11] and Qwen2.5-7B [65] for understanding; Qwen3-0.6B [95] for generation. **Besides, we did not conduct specific training on understanding tasks for our pretrained tokenizer.** Specific training configurations are provided in Table 8 & 9. The code implementation is adapted from the LLaVA [39] repository.

### B.3. Evaluation Details

For image reconstruction, we evaluate on the 256 x 256 ImageNet 50k validation set to compute the rFID, PSNR and SSIM as in TiTok [97] official codebase in Table 1 & 2. For multimodal understanding, we evaluate the SigLIP2-Vicuna variants using the LMMs-Eval codebase [32, 102]. We directly replace the ViT with our tokenizer without specific

| | Stage1 | | | Stage2 | | |
|---|---|---|---|---|---|---|
| **Model** | **SigLIP2-so400m** | **SigLIP2-so400m** | **InternViT-300M** | **SigLIP2-so400m** | **SigLIP2-so400m** | **InternViT-300M** |
| resolution | 256px | 512px | 448px | 256px | 512px | 448px |
| freeze encoder | true | true | true | false | false | false |
| codebook size | 16384 | 16384 | 16384 | 16384 | 16384 | 16384 |
| codebook dim | 1536 | 1536 | 1536 | 1536 | 1536 | 1536 |
| discriminator start steps | NA | NA | NA | 50000 | 50000 | 30000 |
| discriminator weight | NA | NA | NA | 0.1 | 0.1 | 0.1 |
| distillation weight | NA | NA | NA | 1.0 | 1.0 | 1.0 |
| perceptual loss weight | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 |
| perceptual model | convnext-s | convnext-s | convnext-s | convnext-s | convnext-s | convnext-s |
| augmentation | random crop & random flip | random crop & random flip | random crop & random flip | random crop & random flip | random crop & random flip | random crop & random flip |
| encoder lr | NA | NA | NA | 1e-5 | 1e-5 | 1e-5 |
| decoder lr | 4e-4 | 4e-4 | 4e-4 | 1e-4 | 1e-4 | 1e-4 |
| end lr | 1e-4 | 1e-4 | 1e-4 | 1e-5 | 1e-5 | 1e-5 |
| scheduler | cosine | cosine | cosine | cosine | cosine | cosine |
| weight decay | 1e-4 | 1e-4 | 1e-4 | 1e-4 | 1e-4 | 1e-4 |
| discriminator lr | NA | NA | NA | 4e-5 | 4e-5 | 4e-5 |
| optimizer | AdamW | AdamW | AdamW | AdamW | AdamW | AdamW |
| $(\beta_1, \beta_2)$ | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) | (0.9, 0.999) |
| warmup steps | 2000 | 2000 | 2000 | 2000 | 2000 | 2000 |
| mixed precision | bf16 | bf16 | bf16 | bf16 | bf16 | bf16 |
| max grad norm | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| global batch size | 1024 | 1024 | 768 | 1024 | 1024 | 768 |
| total steps | 100000 | 100000 | 45000 | 70000 | 70000 | 60000 |

Table 7. The detailed training configurations of VQRAE tokenizer.

| Hyperparameters | Pretrain | SFT |
|---|---|---|
| freeze backbone | true | false |
| freeze ViT | true | true |
| freeze connector | false | false |
| mm_projector_type | mlp2x_gelu | mlp2x_gelu |
| mm_vision_select_layer | -1 | -1 |
| deepspeed stage | ZeRO-2 | ZeRO-3 |
| scheduler | cosine | cosine |
| learning rate | 1e-3 | 2e-5 |
| weight decay | 0.0 | 0.0 |
| optimizer | AdamW | AdamW |
| $(\beta_1, \beta_2)$ | (0.9, 0.999) | (0.9, 0.999) |
| warmup ratio | 0.03 | 0.03 |
| mixed precision | bf16 | bf16 |
| max grad norm | 1.0 | 1.0 |
| global batch size | 256 | 128 |
| model max length | 2048 | 2048 |

Table 8. The detailed training configurations of multimodal understanding, SigLIP2-Vicuna-v1.5-7B / 13B.

training. Due to the compatibility with InternVL3 [106], we utilize the OpenCompass VLMEvalKit [15] for evaluation of InternViT-Qwen2.5-7B in Table 3. For visual generation, we use the official evaluation toolkit from GenEval [21] and DPG-Bench [24] in Table 4.

# C. Additional Qualitative Results

We provide more qualitative results on image reconstruction and generation in this section. Also, we present some failure cases in our tokenizer and AR model.

| Hyperparameters | Generation Tuning |
|---|---|
| augmentation | center crop |
| deepspeed stage | ZeRO-1 |
| scheduler | constant |
| learning rate | 1e-4 |
| weight decay | 0.0 |
| optimizer | AdamW |
| $(\beta_1, \beta_2, \epsilon)$ | (0.9, 0.999, 1e-15) |
| warmup steps | 4000 |
| mixed precision | bf16 |
| max grad norm | 1.0 |
| global batch size | 512 |
| model max length | 1536 |

Table 9. The detailed training configurations of visual generation.

## C.1. Reconstruction Results

As shown in Figure 7, our VQRAE can achieve fine-grained reconstruction in human faces, scenes and objects.

## C.2. Generation Results

We present additional visual generation results in Figure 8. Our method can generate images with various styles, subjects, and scenarios.

## C.3. Failure Cases

As shown in Figure 9, our tokenizer remains flawed in text reconstruction and high-density scenarios, which is likely

attributable to the trade-off between semantic representation and reconstruction performance and specific text data tuning. Moreover, in terms of image generation, in Figure 10, certain artifacts persist in fingers and human faces, issues that may primarily necessitate resolution through post-training [7, 20, 70].

## D. Limitation and Future Work

The primary limitation of VQRAE lies in the lack of exploration of alternative and more effective methods to balance understanding and reconstruction performance to minimize the compromise on understanding capability. The potential for reconstruction and generation to enhance understanding remains underexplored. Additionally, the quantization loss inherent in the discrete tokenizer makes it challenging for VQRAE to compete with state-of-the-art continuous VAEs. There is still room for improvement in generation quality, particularly in understanding spatial relationships, texture rendering, and addressing artifacts in faces and fingers.

In this work, our main objective is to develop a unified tokenizer that provides more effective representations for understanding, generation, and reconstruction tasks. However, leveraging such representations to integrate various tasks into a single model requires further investigation. Issues such as conflicts and synergies among different tasks, as well as efficient model scaling, are left for future work.

## E. Acknowledgements

This work was supported by Kuaishou Technology. We extend our sincere gratitude to all collaborators involved in this project. Moreover, this work is also supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008).

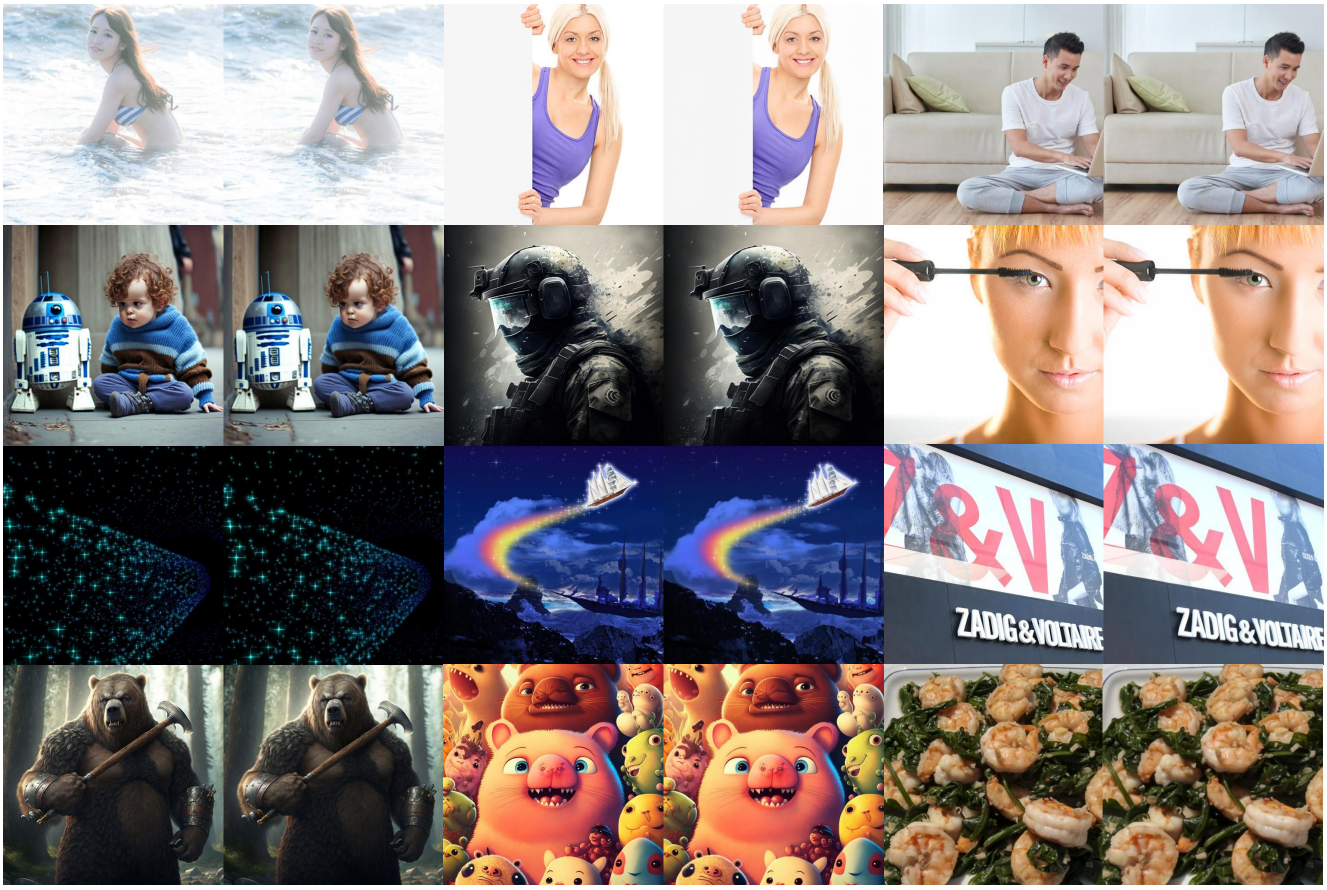


Figure 7. Additional visualization of reconstruction results from VQRAE-InternViT. Left: input image; Right: output image.

Figure 8. Additional visualization of generation results at 512 x 512 px.

Figure 9. Failure reconstruction cases from VQRAE-InternViT. Left: input image; Right: output image.


Figure 10. Failure generation cases. Our model still has certain artifacts in generating fine-grained text, small human faces and fingers, which can be addressed with extensive training data and reinforcement learning as explored in [12, 20, 70].