ViQ: Text-Aligned Visual Quantized Representations at Any Resolution

Xumin Yu[1,*], **Zuyan Liu**[1,2,*], **Zhenyu Yang**[1,4,*], **Yuhao Dong**[3], **Shengsheng Qian**[4], **Jiwen Lu**[2], **Han Hu**[1], **Yongming Rao**[1,†]

[1] Tencent HY Vision Team [2] Tsinghua University

[3] Nanyang Technological University [4] Chinese Academy of Sciences

GitHub: https://github.com/yuxumin/ViQ
HuggingFace: https://huggingface.co/XuminYu/ViQ_weights

Abstract

A unified representation for text and vision is a natural pursuit, as it enables simpler multimodal modeling and more efficient training. However, representing images as discrete signals in the same way as text inevitably introduces severe information loss. Existing work struggles to balance low-level details and high-level semantics in discrete representations: reconstruction-oriented representations often lack semantic information, whereas semantically stronger features typically suffer from severe loss of detail. We present **ViQ**, a **Vi**sual **Q**uantized Representations framework, which is designed to balance semantics and details in discrete representations while supporting inputs at native resolutions, thereby enabling it to serve as a unified and general discrete representation for arbitrary visual inputs. Our approach structures quantization learning into two stages: text-aligned pre-training and feature discretization. With text-aligned pre-training, we enhance the visual encoder semantic-rich supervision from the pretrained language model and enable it to process native-resolution visual inputs. During discretization, we propose a proximal representation learning strategy to progressively compact the feature space, along with a position-aware head-wise quantization mechanism that enables flexible processing of arbitrary resolutions. Extensive experiments on multimodal tasks demonstrate that ViQ achieves competitive performance compared to state-of-the-art multimodal vision encoders with continuous and high-dimensional visual features, while maintaining high precision in low-level reconstruction. We also show that multimodal training with visual quantized representations largely improves efficiency, yielding up to 20%-70% acceleration with different base LLMs and training recipes.

Keywords: Visual Tokenization · Vector Quantization · Multimodal Representation Learning · Native Resolution

1 Introduction

The rapid development of multimodal large language models (MLLMs) QwenTeam (2024); Chen et al. (2024b); Zhu et al. (2025a) has created a growing demand for high-quality, unified visual representations. A key challenge in this field lies in aligning visual signals into a form that can be seamlessly interpreted by language models. Early and dominant approaches have primarily relied on continuous visual encoders Zhai et al. (2023); Radford et al. (2021); Fini et al. (2025), pre-trained through contrastive vision-language learning,

[*] Equal contribution. [†] Corresponding author.

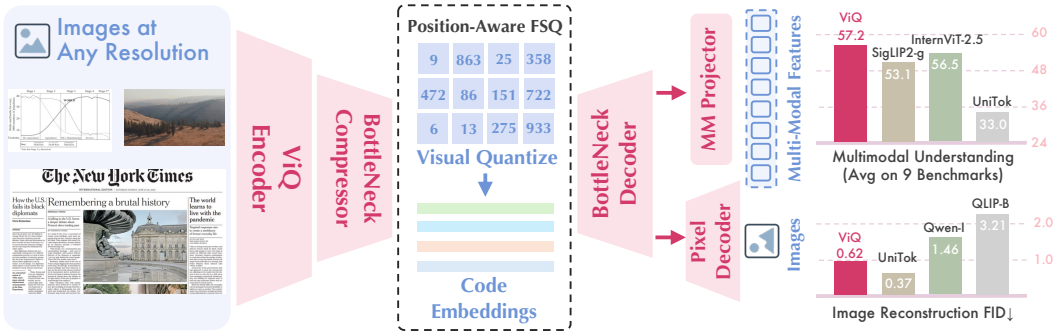

Fig. 1: **ViQ delivers high-quality multimodal quantized representations with both high-level semantics and low-level details.** The quantized visual codes in ViQ support high-level multimodal understanding and low-level image reconstruction, with state-of-the-art performance compared with continuous visual encoders.

or specialized encoders fine-tuned for multimodal tasks Chen et al.; QwenTeam (2024); Liu et al. (2024c). Although this paradigm has achieved considerable success, it introduces a fundamental representational discrepancy: the continuous nature of visual features is intrinsically mismatched with the discrete token-based modeling of text. Furthermore, the computational process of extracting high-dimensional visual features places significant hardware strain during model training.

Inspired by the success of vector quantization Gersho and Gray (2012); Mentzer et al.; Yu et al. in visual representation learning, a growing body of research has begun exploring discrete visual representations in multimodal domains Ma et al. (2025); Zhao et al. (2025). These approaches typically quantize visual inputs into a finite set of discrete tokens, analogous to words in a textual vocabulary. However, a fundamental challenge remains in balancing low-level visual details with high-level semantic information. Reconstruction-focused autoencoders often fail to preserve semantic structure in their latent features, while semantically rich visual encoders tend to incur significant information loss during quantization. This creates a critical gap in multimodal representations, which require both fine-grained visual details and rich semantic content to handle complex tasks effectively. As a result, the application of quantized visual encoders has been largely limited, and they have yet to match the high fidelity and robustness of continuous encoders in demanding real-world scenarios.

To address this issue, we propose ViQ (Visual Quantized Representations), a novel approach that elevates the performance of discrete visual representations in multimodal tasks to a level comparable with widely-used continuous features, while supporting native-resolution visual inputs. Our method structures quantization learning into two phases: text-aligned pre-training and visual quantization learning. In the initial stage, the ViQ model learns semantically rich alignments through supervision from vision-language pairs. To mitigate information loss during quantization, we introduce a proximal representation learning strategy that constrains the latent visual space, followed by a quantization mechanism enhanced with position encoding and expanded visual features. This design supports quantization at arbitrary resolutions and improves the representational capacity of visual codes. Our work not only strongly enhances the efficiency of multimodal encoders in training process but also unifies visual and linguistic representations into a cohesive discrete framework.

We evaluate the ViQ model against state-of-the-art continuous multimodal visual encoders and quantized multimodal semantic encoders across a range of experiments. Under consistent fine-tuning data and training protocols, ViQ not only outperforms existing quantized models by a large margin but also achieves competitive results compared to well-established continuous encoders such as InternViT Zhu et al. (2025a), AIMv2 Fini et al. (2025), and SigLIP2 Tschannen et al. (2025). On the aggregated score over nine benchmarks—spanning visual question answering, world knowledge, and document and chart recognition—ViQ

attains an average of 57.2 with Qwen2.5-1.5B as the backbone LLM and 63.9 with Qwen2.5-7B, surpassing the previous state-of-the-art scores of 57.0 and 63.8, respectively, under 6B number of visual encoder parameters. The quantized representation also brings substantial efficiency gains in real-world multimodal training. Compared to conventional training strategies, using the ViQ visual encoder yields speedups of 20% to 70% in training recipes varying in sequence lengths. Furthermore, ViQ preserves rich low-level visual details: when fine-tuned with an image decoder, it achieves a PSNR of 22.73 and an rFID score of 0.62, ranking first among mainstream discrete visual autoencoders. This strong performance in both understanding and reconstruction tasks underscores the effectiveness and unity of our discrete representation.

2 Related Works

Visual Representations for MLLMs. The visual encoder serves as a critical bridge between vision and language modalities in multimodal learning. Conventional multimodal language models typically employ CLIP-style models—such as CLIP Radford et al. (2021), SigLIP Zhai et al. (2023), SigLIP2 Tschannen et al. (2025), AIM Fini et al. (2025), and DFN Fang et al. (2023)—as visual encoders. However, such models often rely on fixed input resolutions, which constrains their flexibility in multimodal tasks, and their contrastively learned representations may not align optimally with downstream multimodal objectives. To address these limitations, several specialized MLLM visual encoders have been developed and trained end-to-end on multimodal tasks. These include open-source models like InternViT Chen et al. (2024b); Zhu et al. (2025a), OryxViT Liu et al. (2024c), and SAILViT Yin et al. (2025), as well as proprietary encoders integrated into large MLLM systems such as Qwen-VL Team (2025), Kimi-VL Team et al. (2025), and Seed-VL Guo et al. (2025). More recently, researchers have begun exploring unified discrete representations for vision and language. Quantized multimodal encoders such as QLIP Zhao et al. (2025) and UniTok Ma et al. (2025) apply quantization to image inputs. Nevertheless, these quantized visual encoders still exhibit a large performance gap compared to continuous models, particularly in tasks requiring textual understanding or fine-grained visual details. This gap can be attributed to the high compression ratio of discrete codes and the high sensitivity of multimodal models to detailed visual information.

Quantized Visual Encoders. Image tokenization is pivotal in bridging raw pixels with compact latent representations for visual generation. Among existing approaches, vector-quantized (VQ) Gersho and Gray (2012) tokenizers have gained prominence due to their discrete latent space, which is well-suited for visual generation. The seminal VQ-VAE Van Den Oord et al. (2017) introduced a learnable codebook to discretize continuous features via nearest-neighbor assignment. Subsequent methods (such as FSQ Mentzer et al., RPQ Chiu et al. (2022), BSQ Zhao et al., and LFQ Yu et al.) explored structurally constrained codebooks with predefined geometries, often decoupled from end-to-end training. Enhancements like VQ-GAN Esser et al. (2021) further improved visual quality by incorporating adversarial and perceptual losses. While these tokenizers excel at learning compact, generative-friendly representations, they often struggle to preserve fine-grained visual details crucial for dense prediction and high-fidelity understanding tasks. To address this, we propose ViQ, a novel framework designed to minimize information loss and align discrete visual tokens with semantic and textual contexts.

3 ViQ: Visual Quantized Representations

The visual quantization training for ViQ follows a two-stage process, as depicted in Fig. 2. In Stage 1, text-aligned pre-training is performed on continuous features to enhance multimodal alignment. Subsequently, Stage 2 applies quantization in a progressive manner.

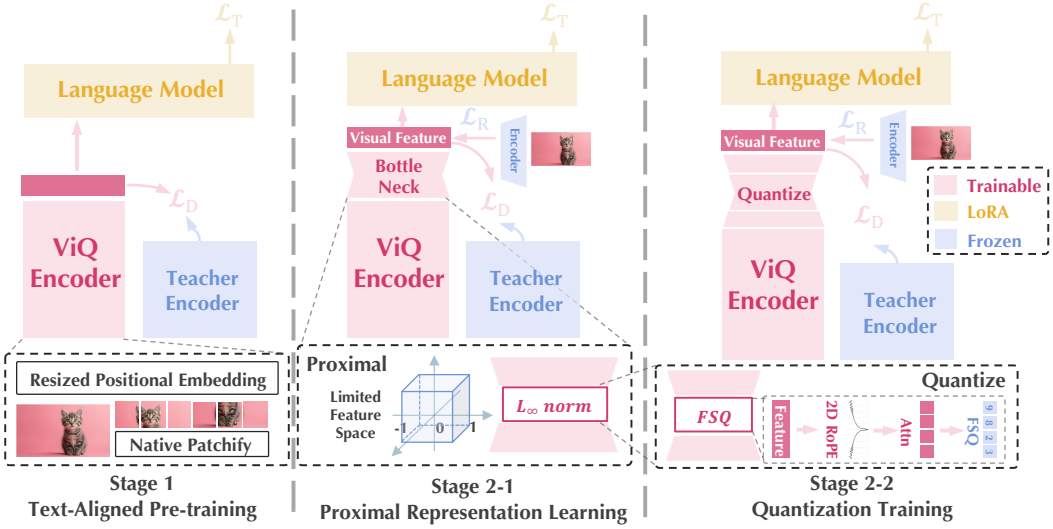

Fig. 2: **Approach of ViQ Representation Learning.** Stage 1 enables multimodal alignment with language supervision, while Stage 2 compresses the high-dimensional visual features into discrete codes in a progressive learning manner.

3.1 Text-Aligned Pre-Training at Any Resolution

The first stage of ViQ training aims to create a visual encoder that functions in a multi-modal manner. This is accomplished by leveraging text-aligned visual pre-training to align the vision and language embeddings within ViQ, thereby following an optimization strategy similar to conventional multi-modal training.

Any Resolution Adaptation. While Language-Image Pre-training models such as CLIP Radford et al. (2021) and SigLIP Zhai et al. (2023) offer a strong initialization for building multi-modal visual encoders, most existing models are constrained by a fixed input size. In multi-modal learning, native image resolution often provides a more natural and efficient form of visual representation. To overcome the limitations of fixed-scale pre-training, we follow the approach of NaViT Dehghani et al. (2024) and replace the original positional embedding layer with one that supports an adequately dimensioned size. This allows the model to resize positional parameters when processing images at arbitrary resolutions dynamically. We adopt the techniques introduced in OryxViT Liu et al. (2024c) to maintain computational efficiency with variable-length visual tokens.

Text-Guided Multi-Modal Pre-training. We optimize the continuous feature space of ViQ to be more compatible with multi-modal learning during our text-aligned pre-training stage. Specifically, given a data triplet $[I, T, A]$ consisting of an image I, a text query T, and an answer A, we employ a temporary language model to compute the text supervision loss $\mathcal{L}_{\text{text}}$, which is defined as:

$$\mathcal{L}_{\text{text}} = \text{Cross Entropy}[\,\text{LLM}(\text{ViQ}(I),\ T),\ A] \tag{1}$$

Self Distillation. During the multi-modal pre-training stage, we apply self-distillation to regularize the ViQ model, preventing it from overfitting to the multi-modal data and disregarding the knowledge acquired from large-scale language-image pre-training, which is crucial for maintaining the generalizability and foundational capabilities. We use the original fixed-resolution model as the teacher to supervise the cosine similarity of the semantic token (i.e., the class token in mainstream architectures), ensuring that the representations produced at any resolution preserve the original high-level semantic information:

$$\mathcal{L}_{\text{distill}} = 1 - \cos\left(\mathbf{z}_s^{\text{student}}, \mathbf{z}_s^{\text{teacher}}\right) \tag{2}$$

where $\mathbf{z}_s^{\text{student}}$ and $\mathbf{z}_s^{\text{teacher}}$ denote the semantic tokens from the student (any-resolution) and teacher (fixed-resolution) models, respectively.

Training Recipe for Native Resolution. We employ a progressive training strategy that transitions from fixed-resolution to any-resolution inputs. Specifically, the training begins with low-resolution images at their native aspect ratios, and the total number of pixels is gradually increased throughout the training process until native resolution is attained. During training, multi-modal data are packed to fit the maximum token length for higher pre-training efficiency. The image complexity, question-answering difficulty, and overall sequence length are progressively raised to steadily adapt the model to multi-modal tasks. Further implementation details are provided in the appendix. We combine the text loss $\mathcal{L}_{\text{text}}$ and the distillation loss $\mathcal{L}_{\text{distill}}$ for joint optimization.

3.2 Visual Quantized Representation Learning

The second stage of ViQ training involves quantizing the continuous features into discrete codes. A core challenge in this process is to fully optimize the discrete representation space so as to preserve fine-grained visual information. Through our approach, we demonstrate that the resulting ViQ representations are effective in supporting high-quality visual understanding in multimodal contexts.

Proximal Representation. The primary objective of quantization is to map continuous visual features from a high-dimensional space $\mathbb{R}^C$ into a constrained discrete space defined by a codebook $\mathcal{Z}$. However, we observe that directly quantizing high-dimensional visual features results in significant precision loss. To mitigate this, our proposed ViQ method progressively reduces the complexity of the latent space using a proximal representation. Given a high-dimensional visual feature $f \in \mathbb{R}^C$, we first apply a bottleneck layer to compress it into an intermediate-dimensional feature $f_1 \in \mathbb{R}^D$. We then further constrain the feature space complexity via a regularization function, referred to as a proximal representation. In our implementation, we apply the L_∞-norm to the compressed latent feature, projecting all features onto a hypercube surface such that $\|f_1\|_\infty = 1$. This step progressively reduces the feature space complexity prior to discrete quantization. The proximal representation helps regulate the distance between features and quantization anchors, thereby reducing information loss during quantization. The bottleneck compression operations can be formulated as:

$$f_1 = L_\infty(\text{BN}(f)), \hat{f} = \text{BN}'(f_1) \tag{3}$$

where BN denotes the bottleneck fully connected layer for dimension compression and BN$'$ denotes the inverted bottleneck layer respectively.

Multi-Head Finite Scalar Quantization. After learning an initial continuous proximal representation within a constrained latent space, we replace the normalization function with a bottleneck downsampling layer that projects the latent features into a lower-dimensional space $f_2 \in \mathbb{R}^d$, where $d \ll D \ll C$. The final dimension d is kept sufficiently small to facilitate effective quantization. We employ Finite Scalar Quantization (FSQ) Mentzer et al. to discretize the compact continuous features f_2, as this method offers stable training behavior without requiring additional optimization during quantization. The FSQ process can be formulated as:

$$z = \text{round}(\mathcal{Q}(f_2)) \tag{4}$$

where $\mathcal{Q}$ indicates the Finite Scalar Quantization function.

To enhance the representational capacity of the quantized codes, we utilize a multi-head attention mechanism that expands each visual patch into 2×2 visual codes. Specifically, the latent feature $f_2 \in \mathbb{R}^{B\times N\times d}$ is up-projected to $\mathbb{R}^{B\times 4N\times d}$, where N denotes the number of original image patches. The expanded patches are then processed via multi-head self-attention across each patch. After self-attention, the tokens are concatenated along the token dimension prior to quantization. Following quantization, the sequence length is restored via a projection layer to maintain the original visual downsampling rate. It is important to note that the expanded image patches are processed independently; as a result, the quantized

Table 1: **Comparison results on multi-modal understanding tasks.** We conduct experiments with various visual encoder counterparts on general multi-modal benchmarks. All the methods are trained with the same data collection.

Base LLM													
Visual Encoder	Size	AnyRes	Discrete	MMStar	MMMU	SimpleVQA	InfoVQA	TextVQA	DocVQA	OCRBench	AI2D	ChartQA	Avg.
Qwen2.5 - 1.5B													
OAI-CLIP-L Radford et al. (2021)	0.3B	✗	✗	44.9	40.7	21.3	24.2	58.9	52.9	460.0	68.1	55.0	45.8
SigLIP2-g Tschannen et al. (2025)	1.1B	✗	✗	48.1	42.4	25.6	28.2	73.1	67.8	590.0	71.5	62.0	53.1
DinoV2-g Oquab et al. (2023)	1.1B	✗	✗	47.1	41.8	24.0	31.7	72.1	76.9	619.0	68.8	61.8	54.0
OryxViT Liu et al. (2024c)	0.4B	✓	✗	46.4	42.1	23.2	31.8	71.8	73.5	622.0	68.2	62.1	53.4
AIMv2-H Fini et al. (2025)	0.7B	✗	✗	48.5	41.8	23.5	31.9	71.6	73.7	623.0	69.8	62.5	53.9
InternViT-2.5 Chen et al.	0.3B	✗	✗	47.9	40.3	23.6	35.5	73.0	81.7	681.0	69.6	**69.2**	56.5
InternViT-2.5-6B Chen et al.	6.0B	✗	✗	**48.5**	42.1	23.7	35.2	**75.5**	80.1	**690.0**	**70.7**	67.8	57.0
QLIP Zhao et al. (2025)	0.3B	✗	✓	39.9	36.9	13.7	14.8	45.1	12.2	290.0	61.9	14.1	29.7
UniTok Ma et al. (2025)	0.3B	✗	✓	41.0	36.1	15.5	15.9	39.7	11.6	323.0	61.2	43.8	33.0
ViQ	1.3B	✓	✓	47.8	**42.6**	**26.0**	**41.6**	74.3	**84.2**	636.0	69.7	65.2	**57.2**
Qwen2.5 - 7B													
OAI-CLIP-L Radford et al. (2021)	0.3B	✗	✗	53.9	47.1	25.4	33.9	66.4	61.4	544.0	76.6	65.1	53.8
SigLIP2-g Tschannen et al. (2025)	1.1B	✗	✗	**57.2**	48.3	28.5	37.3	78.7	75.0	671.0	**79.5**	72.8	60.5
OryxViT Liu et al. (2024c)	0.4B	✓	✗	56.4	48.1	26.5	39.9	78.5	79.8	660.0	78.2	72.1	60.6
AIMv2-H Fini et al. (2025)	0.7B	✗	✗	55.2	48.2	26.8	41.8	79.1	80.1	687.0	77.8	72.5	61.1
InternViT-2.5-6B Chen et al.	6.0B	✗	✗	55.3	48.1	28.4	44.9	**79.9**	85.7	**757.0**	78.7	**77.4**	63.8
ViQ	1.3B	✓	✓	54.2	**49.1**	**28.5**	**55.3**	78.5	**88.9**	711.0	76.7	72.8	**63.9**

representations in ViQ remain independent, making them more suitable for representation learning and downstream tasks.

Rotary Position Embedding. For quantized representations at arbitrary resolutions, we incorporate a 2D rotary position embedding (RoPE) Su et al. (2024) to explicitly encode spatial resolution information. The RoPE layer is inserted prior to the quantization step. Specifically, 2D RoPE extends the 1D RoPE formulation used in large language models by embedding both height and width information of the visual content. Given a token feature $f_m \in \mathbb{R}^d$ at position (h, w) in the 2D feature map, the rotary encoding is applied as:

$$\tilde{f}_m = f_m \odot e^{i(h\theta_h + w\theta_w)} \tag{5}$$

where θ_h and θ_w are frequency parameters for the height and width dimensions, respectively, and $\odot$ denotes element-wise multiplication with complex exponentials. This formulation ensures that relative spatial relationships are preserved across varying resolutions.

Multi-Stage Training with Low-Level Supervision. Building upon the progressive quantization framework, we learn the compressed feature representation in a multi-stage manner. First, starting from the proximal representation in the constrained space, we introduce a bottleneck layer with a regularization function between the continuous input features and the final output. We retain both the text loss $\mathcal{L}_{\text{text}}$ and the self-distillation loss $\mathcal{L}_{\text{distill}}$ during this phase, keeping all other settings consistent with Stage 1 training. To enhance low-level detail preservation and improve latent feature quality, we incorporate a pre-trained visual autoencoder that encodes the original input images into the low-level latent features. This encoding is supervised using a VAE-style latent loss along with a negative log-likelihood (NLL) term, expressed as:

$$\mathcal{L}_{\text{recon}} = \text{NLL}(\hat{f}, \text{Encoder}(x)) \tag{6}$$

where x denotes the input image and $\hat{f}$ denotes the recovered features after compression or quantization. Here the NLL term is computed under a Gaussian likelihood with fixed (unit) variance over the target VAE latent $\text{Encoder}(x)$, so that the loss reduces, up to a constant, to a mean-squared error between $\hat{f}$ and $\text{Encoder}(x)$, i.e., $\mathcal{L}_{\text{recon}} = \frac{1}{2}\|\hat{f} - \text{Encoder}(x)\|_2^2 + \text{const}$. This makes the objective a simple and stable regression on the pre-trained VAE latent space rather than a pixel-level reconstruction.

Once the bottleneck compression is adequately learned, we replace the proximal regularization function with a quantization module in the subsequent quantization training stage. Since our vector quantization (VQ) implementation is optimization-free, no additional

Table 2: **Comparison results on multi-modal understanding tasks.** We conduct experiments with various visual encoder counterparts on general multi-modal benchmarks. All the methods are trained with the same data collection.

Base LLM													
Visual Encoder	Size	AnyRes	Discrete	MMStar	MMMU	SimpleVQA	InfoVQA	TextVQA	DocVQA	OCRBench	AI2D	ChartQA	Avg.
Qwen2.5 - 1.5B													
OAI-CLIP-L Radford et al. (2021)	0.3B	✗	✗	44.9	40.7	21.3	24.2	58.9	52.9	460.0	68.1	55.0	45.8
SigLIP2-g Tschannen et al. (2025)	1.1B	✗	✗	48.1	42.4	25.6	28.2	73.1	67.8	590.0	71.5	62.0	53.1
DinoV2-g Oquab et al. (2023)	1.1B	✗	✗	47.1	41.8	24.0	31.7	72.1	76.9	619.0	68.8	61.8	54.0
OryxViT Liu et al. (2024c)	0.4B	✓	✗	46.4	42.1	23.2	31.8	71.8	73.5	622.0	68.2	62.1	53.4
AIMv2-H Fini et al. (2025)	0.7B	✗	✗	48.5	41.8	23.5	31.9	71.6	73.7	623.0	69.8	62.5	53.9
InternViT-2.5 Chen et al.	0.3B	✗	✗	47.9	40.3	23.6	35.5	73.0	81.7	681.0	69.6	**69.2**	56.5
InternViT-2.5-6B Chen et al.	6.0B	✗	✗	**48.5**	42.1	23.7	35.2	**75.5**	80.1	**690.0**	**70.7**	67.8	57.0
QLIP Zhao et al. (2025)	0.3B	✗	✓	39.9	36.9	13.7	14.8	45.1	12.2	290.0	61.9	14.1	29.7
UniTok Ma et al. (2025)	0.3B	✗	✓	41.0	36.1	15.5	15.9	39.7	11.6	323.0	61.2	43.8	33.0
ViQ	1.3B	✓	✓	47.8	**42.6**	**26.0**	**41.6**	74.3	**84.2**	636.0	69.7	65.2	**57.2**
Qwen2.5 - 7B													
OAI-CLIP-L Radford et al. (2021)	0.3B	✗	✗	53.9	47.1	25.4	33.9	66.4	61.4	544.0	76.6	65.1	53.8
SigLIP2-g Tschannen et al. (2025)	1.1B	✗	✗	**57.2**	48.3	28.5	37.3	78.7	75.0	671.0	**79.5**	72.8	60.5
OryxViT Liu et al. (2024c)	0.4B	✓	✗	56.4	48.1	26.5	39.9	78.5	79.8	660.0	78.2	72.1	60.6
AIMv2-H Fini et al. (2025)	0.7B	✗	✗	55.2	48.2	26.8	41.8	79.1	80.1	687.0	77.8	72.5	61.1
InternViT-2.5-6B Chen et al.	6.0B	✗	✗	55.3	48.1	28.4	44.9	**79.9**	85.7	**757.0**	78.7	**77.4**	63.8
ViQ	1.3B	✓	✓	54.2	**49.1**	**28.5**	**55.3**	78.5	**88.9**	711.0	76.7	72.8	**63.9**

codebook supervision is required at this stage. The overall training objective combines the three losses as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} \tag{7}$$

4 Experiments

In our experiments, we comprehensively evaluate the effectiveness of the ViQ model. We first describe the implementation details of ViQ's key components. We then perform fair comparisons on multimodal tasks under consistent fine-tuning data and training protocols, benchmarking against leading multimodal and quantized encoders. To highlight practical benefits, we assess efficiency in real-world training and visual storage scenarios. We further examine the reconstruction quality to analyze low-level feature preservation, and provide ablation studies to validate design choices.

4.1 Implementation Details

We instantiate our ViQ model based on the pretrained SigLIP2-g Tschannen et al. (2025) visual encoder, whose output feature dimension is $C = 1536$. To meet quantization requirements, we reduce the dimensionality through bottleneck layers to $D = 128$ and subsequently to $d = 6$. For quantization, we employ the FSQ method Mentzer et al. with levels $\mathcal{L} = [8, 8, 8, 5, 5, 5]$, yielding a codebook size of 64,000. Multimodal features from ViQ are projected via a simple MLP layer. During training, we use Qwen2.5-VL-0.5B as the language model for text supervision and the pretrained Qwen-Image Wu et al. (2025) encoder for low-level visual supervision, keeping its parameters fixed. Stage 1 pretraining is conducted using 128 NVIDIA A100 GPUs, while Stage 2 quantization training uses 256 NVIDIA A100 GPUs. Further details regarding training stages and datasets are provided in the appendix.

4.2 Multi-Modal Understanding

In this subsection, we present experiments on multi-modal understanding to evaluate the effectiveness of our ViQ models in learning superior visual representations. By comparing ViQ with various baseline models, we show that our approach achieves compact visual representations without compromising their quality.

Setups. We integrate ViQ with large language models of varying scales to evaluate the generalization capability of our vision encoders. To compare with other visual encoders, as ViQ supports native-resolution perception, we adapt the LLaVA-NeXT Liu et al. (2024a)'s "any resolution" training pipeline for other visual encoders that operate at fixed resolutions to ensure a fair comparison. For evaluation, we adopt representative visual understanding benchmarks, including general multi-modal benchmarks MMStar Chen et al. (2024a), MMMU Yue et al. (2024), world knowledge-relevant benchmarks SimpleVQA Cheng et al. (2025), InfoVQA Mathew et al. (2022), text and doc recognition benchmarks including TextVQA Singh et al. (2019), DocVQA Mathew et al. (2021), OCRBench Liu et al. (2023), chart and scientific recognition benchmarks including AI2D Kembhavi et al. (2016) and ChartQA Masry et al. (2022). The comprehensive benchmarks cover a broad range of visual skills, including basic perception, diagram interpretation, OCR, and visual reasoning. Each experiment is conducted on a fixed dataset of 2000K samples drawn from LLaVA-OneVision Li et al. (2024b) to maintain consistency and fairness. We employ LMMs-Eval Li et al. (2024a) as our evaluation toolkit to ensure reproducible results.

Visual Encoder Baselines. To enable a comprehensive comparison with existing visual encoders, we carefully select a diverse set of models across different categories. For general multi-modal visual encoders, we include widely used models such as OpenAI CLIP Radford et al. (2021), SigLIP2 Tschannen et al. (2025), and DINOv2 Oquab et al. (2023). For visual encoders specialized optimized for mutli-modal data and tasks, we incorporate AIMv2 Fini et al. (2025), OryxViT Liu et al. (2024c), and InternViT Zhu et al. (2025a) (including 300M and 6B variants). In addition, we evaluate quantized visual encoders, including QLIP Zhao et al. (2025) and UniTok Ma et al. (2025), to ensure a thorough and balanced comparison.

Results. As shown in Table 2, ViQ achieves competitive performance compared to other visual encoders. Despite being quantized, ViQ matches or surpasses representative general-purpose encoders on the aggregated score, primarily due to its native-resolution perception capability and quantization-aware training strategy, which together preserve strong visual perception throughout quantization. Compared to multimodal encoders that are specifically trained for visual understanding, ViQ remains highly competitive overall and is particularly strong on text- and document-centric tasks. We note, however, that the advantage is not uniform: on certain detail-intensive benchmarks such as OCRBench, ViQ still trails some continuous encoders with fewer parameters, which we attribute to the inherent loss of high-frequency details when the continuous feature space is aggressively compressed into discrete codes. Moreover, ViQ consistently outperforms previous quantized encoders by a large margin, which typically suffer from severe degradation after quantization, highlighting the effectiveness of our proximal representation learning and quantization training strategy in preserving visual quality.

We further observe that ViQ's gains are most pronounced on OCR, document, and infographic understanding tasks. We believe this is because general benchmarks depend more heavily on the backbone LLM's knowledge and reasoning ability, whereas OCR-, document-, and chart-oriented tasks require precise low-level visual details and thus more directly reflect the capability of the visual encoder. From this perspective, the most meaningful contribution of ViQ is that it preserves such fine-grained understanding even after aggressively compressing continuous images into discrete tokens. The residual gap on the most detail-intensive tasks is a systemic property of discrete tokenization rather than a flaw specific to ViQ, and could be further narrowed by orthogonal directions such as multi-scale or residual quantization and incorporating specialized document data. Overall, ViQ serves as an efficient visual encoder that combines high compression with strong perceptual capability, making it well-suited for multimodal understanding tasks.

4.3 Efficiency

In this section, we demonstrate ViQ's capabilities beyond multi-modal understanding through some experiments.

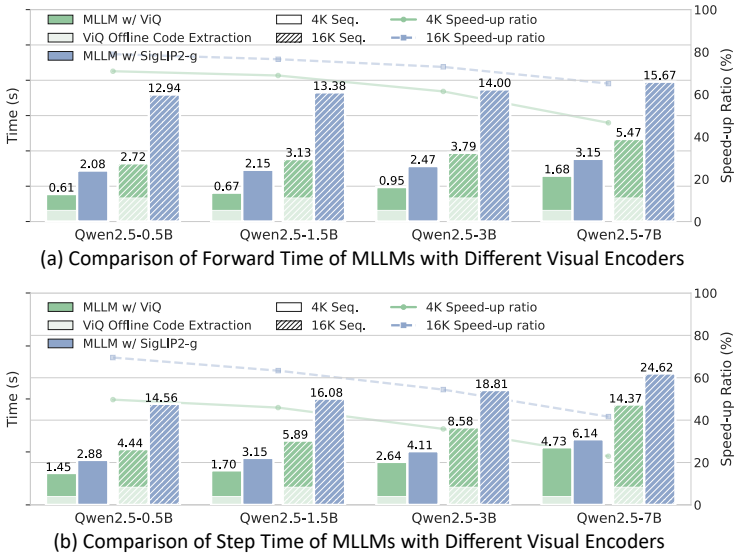

Fig. 3: **Comparisons on Training Efficiency Across Different Visual Encoders.** We conduct the experiments to compare the efficiency of ViQ and SigLIP2-g on the 4k and 16k training.

4.3.1 Training Speed-Up for VLMs

Setup. We integrate ViQ and the popular SigLIP2-g encoder with a series of Qwen2.5 models for VLM SFT as conducted in LLaVA Liu et al. (2024b). All experiments are conducted on a single node to eliminate noise introduced by inter-node communication. We fix the maximum image area to 768^2, while preserving the original aspect ratio. We cover the 4k and 16k training cases in experiments, where image-text pair are packed into a fixed sequence length for stable training.

For SigLIP2-g, following common VLM training practice, we extract features from a list of images and feed them into the LLM together with processed text token embeddings for the forward pass. For ViQ, we first extract the discrete codes of each image offline. During the training, instead of loading raw images, we load the precomputed ViQ codes and project them into the LLM latent space. We discard the first 100 warm-up steps and report the average time required for a single forward pass and for a full training iteration.

Results. As shown in Figure 3, we compare the forward time and step time of SigLIP2-g and ViQ across four different Qwen2.5 model sizes, ranging from 0.5B to 7B. For a fair comparison, we also consider the ViQ offline code extraction time. We can see ViQ provides substantial speed-ups across all model sizes. The gains are especially pronounced for smaller LLMs: for the 0.5B model, ViQ accelerates forward time by 70% and 78% under the 4k and 16k training settings, respectively. And for larger model, like Qwen2.5 7B, ViQ can achieve a 46% and 65% speed-up in forward time under the 4k and 16k settings. Considering a whole iteration step, ViQ offers consistent improvements, exceeding 20% and 40% speed-ups in the 4k and 16k settings.

4.3.2 Storing Any Image as Discrete Codes

As ViQ is a quantized visual encoder, it can convert an image at its native resolution into a series of discrete codes, which can then be reconstructed by a decoder. For an image of shape (H, W) with three channels, storing the raw image consumes $\frac{H\times W\times 3\times 8\text{bits}}{8} = 3HW$ bytes on disk. In contrast, ViQ can translate an image with (H, W) into $\frac{H\times W}{64}$ codes, each ranging from 0 to 64,000, which can be represented using an unsigned 16-bit integer. As a result, storing the ViQ codes requires $\frac{\frac{H\times W}{64}\times 16\text{bits}}{8} = \frac{HW}{32}$ bytes, which is only $\frac{1}{96}$ of the raw

Fig. 4: **Representing images with ViQ.** We show the image compression and visual reconstruction capability of ViQ. ViQ achieves a high-compression-ratio in image storage with high-quality reconstructed images while supporting native-resolution inputs.

Table 3: **Comparison of reconstruction quality on the 256 × 256 ImageNet-1K validation set.** "Und." indicates whether the tokenizer is optimized for understanding tasks. "*" means reproduced results with the official weights and the same evaluation code as our model.

Method	Und.	#Token	PSNR↑	SSIM↑	rFID↓
Continuous Tokenizer					
SD-VAE 3	✗	32×32	31.29	0.87	0.20
FLUX-VAE Laurençon et al. (2024)	✗	32×32	32.74	0.92	0.18
Qwen-Image Wu et al. (2025)	✗	32×32	32.18	0.90	1.46
Cosmos-CI Agarwal et al. (2025)	✗	16×16	25.07	0.70	0.96
Wan2.2 Wan et al. (2025)	✗	16×16	31.25	0.88	0.75
Discrete Tokenizer					
Cosmos-DI Agarwal et al. (2025)	✗	16×16	19.98	0.54	4.40
Show-o Xie et al. (2024)	✗	16×16	21.34	0.59	3.50
LlamaGen Sun et al. (2024)	✗	16×16	20.65	0.54	2.47
MUSE-VL Xie et al. (2025)	✗	16×16	20.14	0.65	2.26
Open-MAGVIT2 Luo et al. (2024)	✗	16×16	22.70	0.64	1.67
QLIP-B Zhao et al. (2025)	✓	16×16	<u>23.16</u>	0.63	3.21
UniTok* Ma et al. (2025)	✓	16×16	**25.32**	**0.77**	**0.37**
ViQ	✓	16×16	22.73	<u>0.66</u>	<u>0.62</u>

image size. This corresponds to a very low target bitrate: matching the same compression ratio with JPEG would require an aggressive quality setting (e.g., $Q \approx 0.08$) that noticeably degrades image quality, whereas ViQ preserves substantially better reconstruction at an equivalent ratio. We therefore compare ViQ against JPEG at a comparable bitrate, and report the compression ratios together with reconstructed visual samples in Figure 4.

4.4 Image Reconstruction Experiments

Setups. For image reconstruction tasks, we train the visual decoder with the fixed pretrained ViQ model. The overall regularization loss for VAE is defined as:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{LPIPS}} + \mathcal{L}_{\text{GAN}} + \lambda_{\text{REPA}} \cdot \mathcal{L}_{\text{REPA}} \tag{8}$$

where $\mathcal{L}_{\text{KL}}$ denotes the Kullback-Leibler divergence, $\mathcal{L}_{\text{MSE}}$ represents the Mean Squared Error, $\mathcal{L}_{\text{LPIPS}}$ corresponds to the Learned Perceptual Image Patch Similarity for pixel reconstruction, and $\mathcal{L}_{\text{GAN}}$ is the adversarial loss from the Generative Adversarial Network Goodfellow et al. (2020). To enhance semantic alignment, we leverage DINOv2 Oquab et al. (2023) as an external feature extractor and apply an alignment loss $\mathcal{L}_{\text{REPA}}$ to the ViQ preprocess layer preceding the decoder head within our ViQ architecture. The alignment loss coefficient is empirically set to $\lambda_{\text{REPA}} = 1.5$ for the DINOv2 head. For model optimization, we employ the AdamW optimizer Loshchilov and Hutter (2017); Kingma and Ba (2017) with a constant learning rate of 6×10^{-4} and a global batch size of 4096. Training stability is ensured through gradient clipping and the application of an Exponential Moving Average (EMA) to the generative model parameters. All models are trained for 50,000 steps with a linear learning rate warmup over the initial 5,000 steps. We utilize mixed-precision training with the BF16 format and unfreeze the VAE preprocess layer during training.

Process	avg.(2-2)
Continuous → SimVQ	60.9
Continuous → BN + l_2 → SimVQ	66.6
Continuous → BN + l_2 → FSQ	67.9
Continuous → BN + l_∞ → FSQ	68.7

(a) **Proximal Representations**. Gradually regularizing the latent space from continuous to quantized helps.

Width	avg.(2-1)
32	68.4
128	69.1
512	68.8
1536	69.3

(b) **Bottleneck size**. The bottleneck width can be made significantly narrower than the 1536-dimensional one in SigLIP2-g.

Size	avg.(2-2)
FSQ + 64,000	68.7
FSQ + 128,000	68.3
SimVQ + 2^{15}	66.5
SimVQ + 2^{17}	65.6

(c) **VQ and Codebook size**. The VQ codebook saturates at a size of 2^{17}.

Design	avg.(2-2)
no Position information injected	65.3
RoPE with Attention	68.7
Learnable Pos Emb	65.7

(d) **Position encoding for Quantization**. position information is important.

Case	Self-Distill	Text	Recon	avg.(2-2)
A	✗	✓	✗	61.3
B	✓	✓	✗	66.8
C	✓	✓	✓	68.7

(e) **Loss Combination**. three type of loss.

Loss Type	Time Cost	avg.(2-2)
none	1x	66.8
MSE + LPIPS	2.3x	67.0
DiT	4x	65.8
Vae Latent Loss	1.3x	68.7

(f) **Reconstruction Loss**. Vae Latent loss is efficient.

Table 4: **ViQ ablation experiments**. We conduct a thorough ablation study of ViQ, examining proximal representations, architecture design, and loss combinations. We report the average metrics on MMStar, MMMU, OCRBench, among a total of eight benchmarks, using a fixed SFT training setup on Qwen2.5-3B. (2-1) and (2-2) refer to the training stage of the test model. Default settings are marked in red.

Results. We comprehensively evaluate the reconstruction quality of various tokenizers on the 256×256 ImageNet-1K validation set, as summarized in Table 3. Among discrete tokenizers, ViQ delivers highly competitive reconstruction quality, attaining an SSIM of 0.66 and an rFID of 0.62 (second only to UniTok), while remaining comparable to QLIP-B in PSNR (22.73 vs. 23.16). We note that UniTok reports stronger raw reconstruction metrics (e.g., PSNR 25.32), which we attribute to its direct reconstruction objective combined with contrastive supervision. However, as shown in Table 2, this type of supervision substantially degrades vision-language alignment and thus leads to sub-optimal multi-modal understanding. By contrast, our core objective with ViQ is to deliver a more balanced tokenizer that maintains highly competitive low-level reconstruction while excelling in high-level semantic understanding. This balance indicates that a well-optimized pixel-based decoder remains a powerful and efficient choice for building a discrete visual tokenizer that serves both generative and discriminative purposes. We also acknowledge that, as a discrete tokenizer, ViQ inevitably compromises some fine-grained reconstruction detail relative to continuous tokenizers, as the continuous visual feature space is drastically compressed; nevertheless, it considerably narrows this gap and offers a favorable trade-off between reconstruction fidelity and semantic understanding.

4.5 Ablation Studies

Proximal Representations In Tables 4a and 4b, we show that initializing the model with intermediate proximal representations can substantially improve the final performance. In Table 4a, we can see that optimizing a continuous feature space into a discrete one leads to a severe performance drop. We further investigate the effect of introducing a bottleneck, a bottleneck with L2 normalization, and a bottleneck with L_∞ normalization. The results highlight two key observations: first, a gradually regularized latent space is more suitable for quantization; second, a more appropriate regularization method for quantization (i.e., L_∞) provides additional benefits. Furthermore, Table 4b shows that, when optimized properly, the bottleneck does not necessarily degrade the final performance.

Quantization Designs We compared the performance of commonly used methods such as SimVQ Zhu et al. (2025b) and FSQ Mentzer et al. under different codebook sizes, as shown in Table 4c. We conduct experiments on different position encoding methods for

ViQ, summarized in Table 4d. We observe that FSQ Mentzer et al., a non-optimized vector quantization method, outperforms SimVQ Zhu et al. (2025b), which requires learning a codebook. We also experimented with LFQ Yu et al., vanilla VQ Gersho and Gray (2012), and IBQ Shi et al. (2025), and found consistent conclusions: in our setting, VQ methods without a learnable vocabulary tend to perform better. In addition, we find that a vocabulary size around 60,000 is good. Increasing the codebook size will reduce the utilization rate of learnable codebooks, which in turn degrades performance, whereas this issue is much less pronounced for non-learnable VQ methods. Table 4d further shows that injecting positional encoding significantly enhances the representational ability of VQ. We conduct experiments on both learnable positional embeddings and RoPE-2D, and observe that RoPE leads to a substantial improvement, whereas learnable positional embeddings offer limited gains. This is because learnable positional embeddings increase the optimization difficulty of the VQ module.

Loss Combination. We further study the loss composition used in our training pipeline. As shown in Table 4e. When keep only the text loss, it leads to a substantial performance drop (Case A). In Case B, add self distillation loss with significant enhance the performace of model. While we can see some unsatisfactory performance on tasks such as OCR and Chart. By introducing a reconstruction loss (Case C), the performance notably improves, with most gains coming from detail-intensive tasks such as OCR and Chart, while the gains for captioning and VQA are relatively limited. Table 4f studies different formulations of the reconstruction loss. We adopt a more efficient and simple approach: a VAE latent reconstruction loss, where the model predicts the latent representation of a pretrained VAE network. This achieves effective optimization while maintaining training efficiency.

4.6 Limitations

Although ViQ demonstrates highly competitive performance in unifying visual and language representations into a cohesive discrete framework, there are a few general limitations to consider. While the study thoroughly validates ViQ's efficiency and effectiveness across large language models ranging from 0.5B to 7B parameters, its integration and synergistic effects with massively larger foundation models (e.g., 70B parameters and beyond) remain an area for future empirical exploration. Moreover, like most advanced multimodal models, the robustness of the learned proximal representations relies on the quality and diversity of the large-scale language-image pre-training data, meaning that inherent data biases could marginally affect the model's zero-shot generalization in highly specialized or extreme out-of-domain scenarios.

5 Conclusion

In this work, we introduced ViQ, a quantized multimodal encoder that unifies visual and language representations at native resolution. We designed a two-stage training pipeline to reduce the information losses of learning discrete visual representations with text-aligned pretraining and feature discretization. Through carefully designed architectures and progressive training techniques that minimize information loss, ViQ achieves competitive performance in multimodal tasks compared to both existing continuous and discrete visual encoders. We believe our work not only enhances the quality and efficiency of visual representations in multimodal learning, but also offers a viable pathway toward unifying vision and language within a shared discrete framework.

References

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.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. *URL https://arxiv. org/abs/2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024b.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646, 2025.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2022.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *NeurIPS*, 36, 2024.

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.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor G Turrisi da Costa, Louis Béthune, Zhe Gan, et al. Multimodal autoregressive pre-training of large vision encoders. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9641–9654, 2025.

Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, pages 235–251. Springer, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL `https://arxiv.org/abs/1412.6980`.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, et al. Lmms-eval: Accelerating the development of large multimoal models, 2024a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024b.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024c.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Zhuoyan Luo, Fengyuan Shi, Yixiao Ge, Yujiu Yang, Limin Wang, and Ying Shan. Open-magvit2: An open-source project toward democratizing auto-regressive visual generation. *arXiv preprint arXiv:2409.04410*, 2024.

Chuofan Ma, Yi Jiang, Junfeng Wu, Jihan Yang, Xin Yu, Zehuan Yuan, Bingyue Peng, and Xiaojuan Qi. Unitok: A unified tokenizer for visual generation and understanding. *arXiv preprint arXiv:2502.20321*, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *The Twelfth International Conference on Learning Representations*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

QwenTeam. Qwen2-vl: To see the world more clearly. *Wwen Blog*, 2024. URL `https://qwenlm.github.io/blog/qwen2-vl/`.

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 *ICML*, pages 8748–8763. PMLR, 2021.

Fengyuan Shi, Zhuoyan Luo, Yixiao Ge, Yujiu Yang, Ying Shan, and Limin Wang. Scalable image tokenization with index backpropagation quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16037–16046, 2025.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

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.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Qwen Team. Qwen2.5-vl, January 2025. URL `https://qwenlm.github.io/blog/qwen2.5-vl/`.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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.

Rongchang Xie, Chen Du, Ping Song, and Chang Liu. Muse-vl: Modeling unified vlm through semantic discrete encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24135–24146, 2025.

Weijie Yin, Yongjie Ye, Fangxun Shu, Yue Liao, Zijian Kang, Hongyuan Dong, Haiyang Yu, Dingkang Yang, Jiacong Wang, Han Wang, et al. Sail-vl2 technical report. *arXiv preprint arXiv:2509.14033*, 2025.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

Yue Zhao, Yuanjun Xiong, and Philipp Kraehenbuehl. Image and video tokenization with binary spherical quantization. In *The Thirteenth International Conference on Learning Representations*.

Yue Zhao, Fuzhao Xue, Scott Reed, Linxi Fan, Yuke Zhu, Jan Kautz, Zhiding Yu, Philipp Krähenbühl, and De-An Huang. Qlip: Text-aligned visual tokenization unifies auto-regressive multimodal understanding and generation. *CoRR*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.

Yongxin Zhu, Bocheng Li, Yifei Xin, Zhihua Xia, and Linli Xu. Addressing representation collapse in vector quantized models with one linear layer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22968–22977, 2025b.

Appendix

A More Details

A.1 Training Details

From Fix Resolution to Native Resolution. We first train our ViT with a light-weight LLM on approximately 3B vision–language tokens, covering tasks such as captioning, chart understanding, OCR, and general diagram comprehension. In this stage, we constrained the image size to within 384^2 while preserving each image's native aspect ratio. For training efficiency, we downsampled the vision tokens by a factor of 16 before feeding them into the LLM. After this stage, we increased the image size to 768^2 and introduced another $\sim$3B training tokens. During this phase, the vision tokens were downsampled by a factor of 4 before being passed to the LLM for text loss supervision. The learning rate was gradually decayed from 2e-5 to 5e-5 for above two stage, respectively. We used Qwen2.5-0.5B as the LLM backbone and optimized only the LoRA parameters. For the self distillation loss, we use the cosine similarity loss for the global features between original SigLIP2-g and trained ViT wit a Multi-head Attention Pooling Layer. The images for SigLIP2-g are resized to 384×384.

Fig. 5: **More Reconstructed Visualization Samples of ViQ at Any Resolution.** Left or above is the original image.

From Continuous to Quantized. In Stage 2-1, we optimized a bottleneck with a dimension of 128, where newly added parameters were initialized with a learning rate of 1×10^{-4}, while all other parameters used a learning rate of 5×10^{-5}, half of the initial value. The learning rate was gradually decayed using a cosine scheduler, ultimately reaching 1×10^{-5} and 5×10^{-6} for different parameter groups. The data source used in Stage 1 was identical to the one mentioned earlier, and we trained on 1B vision-language tokens with a resolution of 768 pixels. The bottleneck was optimized during this phase. For the reconstruction branch, we passed the final vision feature through a head designed for VAE latent feature prediction. This head was structured as a 3-layer MHSA and culminated in a convolutional layer for upsampling. Additionally, vision features underwent a 2x2 pooling operation before being fed into the Large Language Model. During this process, we introduced constraints on the bottleneck's 128-dimensional features, such as an L_∞ norm regularization. In Stage 2-2, we further refined the bottleneck from Stage 1. The previously applied non-parameteric constraints like the L_∞ regularization were replaced with the FSQ module. This module incorporates a quantization mechanism along the 6-dimensional feature space, followed by fully connected layers before and after quantization, along with an attention layer that adds Rotary Positional Embedding Su et al. (2024) information. In this stage, we utilized learning

rate of 5×10^{-5} across all components and trained on approximately 30B vision-language tokens.

B More Visualizations

We selected images with different resolutions and themes, containing a lot of details, to showcase the reconstruction effects with our ViQ.