Investigating Permutation-Invariant Discrete Representation Learning for Spatially Aligned Images

Jamie Stirling[1], Noura Al-Moubayed[1], and Hubert P. H. Shum[1]

Durham University, United Kingdom

Abstract. Vector quantization approaches (VQ-VAE, VQ-GAN) learn discrete neural representations of images, but these representations are inherently position-dependent: codes are spatially arranged and contextually entangled, requiring autoregressive or diffusion-based priors to model their dependencies at sample time. In this work, we ask whether positional information is necessary for discrete representations of spatially aligned data. We propose the permutation-invariant vector-quantized autoencoder (PI-VQ), in which latent codes are constrained to carry no positional information. We find that this constraint encourages codes to capture global, semantic features, and enables direct interpolation between images without a learned prior. To address the reduced information capacity of permutation-invariant representations, we introduce matching quantization, a vector quantization algorithm based on optimal bipartite matching that increases effective bottleneck capacity by 3.5× relative to naive nearest-neighbour quantization. The compositional structure of the learned codes further enables interpolation-based sampling, allowing synthesis of novel images in a single forward pass. We evaluate PI-VQ on CelebA, CelebA-HQ and FFHQ, obtaining competitive precision, density and coverage metrics for images synthesised with our approach. We discuss the trade-offs inherent to position-free representations, including separability and interpretability of the latent codes, pointing to numerous directions for future work.

Keywords: Representation learning · Image synthesis · Vector quantization

1 Introduction

Vector quantization approaches (VQ-VAE, VQ-GAN) have emerged as powerful methods for learning discrete neural representations of images [12, 37, 40, 51]. However, these learned representations are inherently position-dependent: the same visual feature appearing at different spatial locations may map to different codes, and the emergent "visual vocabulary" is highly contextual, requiring autoregressive or diffusion-based priors to model spatial dependencies at sample time [5, 12]. This position- and context-dependence limits the utility of learned codes for interpretability and direct manipulation. It also stands in contrast to

Method	Discrete?	$\mathcal{O}(1)$-time sampling?	Interpolation?
BigGAN [11]	-	✔	-
VDVAE [8]	-	-	-
DDGAN [50]	-	✔	-
StyleGAN-XL [43]	-	✔	✔
LDM-4 [41]	-	-	-
ADM-IP [36]	-	-	-
RDM [46]	-	-	-
WaveDiff [38]	-	-	-
VAE [26]	-	✔	✔
β-VAE [20]	-	✔	✔
VQ-VAE [37]	✔	-	-
VQ-GAN (TT) [12]	✔	-	-
VQ-GAN (UT) [5]	✔	-	-
PI-VQ (Ours)	✔	✔	✔

Table 1. Comparison of the capabilities of our method with the state-of-the-art. "$\mathcal{O}(1)$-time sampling" indicates methods which can sample an image with a single forward-pass of a neural network.

Fig. 1. Visualisation of (approximately) smooth interpolations on CelebA 64x64 (source images to left and right). The discrete and permutation-invariant nature of the latents allows us to generate multiple, equally plausible paths between two images.

the goals of *controllability* and *disentanglement* pursued in the face representation literature [3, 49, 55], which aim to identify direct correspondences between interpretable features (such as pose, lighting [10], skin tone and subject identity [33]) and factors of variation in learned representations. Such properties have been explored extensively for continuous latent variables in VAEs [20] and GANs [15, 24, 32], with recent advances in flow-based approaches [9, 29] demonstrating efficient sampling. Meanwhile, discrete representations remain comparatively under-explored in this regard. In this work, explore the consequences of enforcing permutation invariance on VQ representations, finding that this constraint encourages codes to capture global, semantic features and enables novel capabilities, including direct interpolation, without any need for learned priors over the latent representations.

In this paper, we propose permutation-invariant discrete representation learning (PI-VQ), a technique for learning position-free discrete representations of spatially aligned face data. We impose a permutation-invariance constraint on the discrete latent codes—a principled inductive bias guaranteeing that no positional information is encoded. We argue that discrete, permutation-invariant representations are especially well-suited to aligned images (such as faces) because: (1) such images exhibit many discrete ground-truth attributes (e.g., eye colour, hair colour, etc. for faces), and (2) spatial alignment reduces the need to encode positional information, allowing codes to focus on global features. As a consequence of this design, we obtain memory-efficient representations that support direct interpolation and $\mathcal{O}(1)$-time sampling without any need for training a secondary prior over the latent codes.

To complement PI-VQ, we propose **matching quantization**, a novel vector-quantization algorithm based on optimal bipartite matching. Unlike standard nearest-neighbour quantization [1,37], matching quantization eliminates repeated codes within each image representation, reducing redundancy and increasing the effective information capacity of the permutation-invariant bottleneck.

Finally, we demonstrate that the compositional structure of the learned representations enables **fast interpolation-based sampling**: novel images can be synthesised by recombining discrete codes from existing images, requiring only a single forward pass. This sampling approach is inherently interpretable—it can be understood as recombining discrete visual *traits*—while approaching competitive scores across multiple quality metrics [18, 35, 42] (Table 1).

We evaluate PI-VQ on CelebA, CelebA-HQ and FFHQ, showing that the learned representations are concise, semantic, and global. Our method achieves competitive precision, density and coverage metrics, while we discuss limitations on FID and recall in Section 5.

Our contributions are as follows:

- We propose **permutation-invariant vector quantization (PI-VQ)**, an architecture for discrete representation learning which enforces that latent codes carry no positional information, and investigate the properties of the resulting representations.
- We introduce **matching quantization**, a vector quantization algorithm based on optimal bipartite matching which eliminates code repetition within each image, multiplying the effective information capacity of the permutation-invariant bottleneck by a factor of 3.5.
- We demonstrate that the learned representations enable **interpolation-based sampling**, allowing us to synthesise novel images in $\mathcal{O}(1)$ forward passes by recombining discrete codes, without requiring a learned prior.
- We further show that logistic regression over the learned features is useful for predicting human-annotated features on FFHQ, indicating that the learned representations capture features which separable and interpretable without any explicit disentangling objective.

2 Related Work

The concept of **disentanglement** is widespread in deep learning literature [19], especially in the context of image synthesis [10] and representation learning [49]. In continuous representations, disentanglement aims to find low-dimensional vector representations of high-dimensional spaces [7] such that the components of the representation correspond to underlying ground-truth factors of variation (such as pose, colour etc.). VAE [26] and β-VAE [20] have been proposed as robust approaches to disentangled representation learning. Disentanglement is closely related to the idea of **compositional generalization** [25], by which a model generalizes to unseen combinations of seen concepts. Recent work has explored disentangling facial motion attributes [44], demonstrating the value of separable representations for controllable synthesis.

Discrete representation learning has also emerged as the discrete counterpart to continuous VAE approaches [37]. Discrete representation learning is based on the concept of vector-quantization (VQ) [1], whereby features from a continuous vector space are mapped to an element of a finite set of codebook vectors. This so-called "VQ-family" of models includes VQ-VAE [37] and VQ-GAN [12]. The naive "nearest-neighbour" approach to vector quantization usually results in some degree of codebook collapse, in part due to straight-through gradient estimation [1]. Recent work has explored improved gradient flow through vector quantization layers [13], addressing limitations of straight-through estimation. Multiple methods have been proposed to mitigate or even eliminate codebook collapse, including EdVAE [4], OptVQ [53], and HyperVQ [17]. Our proposed matching quantization method addresses the problem of codebook collapse in the specific setting of permutation-invariant latent codes.

Permutation-invariant techniques aim to learn operations which are invariant to the ordering of their inputs [28]. Empirically, permutation-invariance has been found to have greater robustness to corruption of the inputs and better generalization to unseen situations [45]. Janossy pooling [34], Set transformer [28] and Memory-based Exchangeable Model [21] have been proposed as specialized architectures for learning operations on sets. Concurrent work has explored permutation-invariant embeddings for automated feature selection in tabular data [30]. While sharing the goal of position-free discrete representations, this approach embeds predefined feature indices for downstream search. To our knowledge, our work is the first to propose a *discrete* analogue to permutation-invariant learning.

Sparse binary autoencoders have been applied in the mechanistic interpretability literature for obtaining sparse, semantic and efficient representations of LLM activations [39]. Our learned discrete set representations of faces can be viewed as a special case of binary autoencoders, with the notable distinction that the fixed code length exactly constrains the sparseness of the representation.

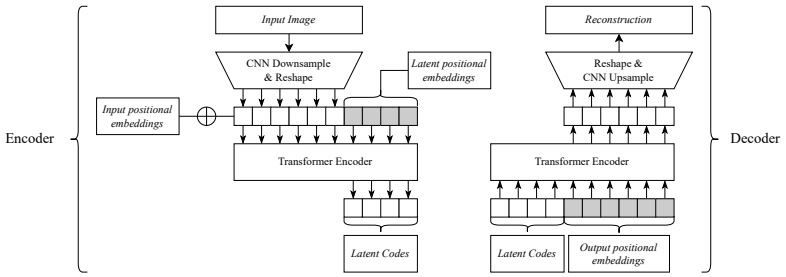

Fig. 2. The encoder and decoder components of the permutation-invariant autoencoder (quantization step omitted for clarity). Positional embeddings are not applied to the latent codes before decoding: as a result, the decoder is invariant to permutation of the latent codes.

3 Methods

3.1 Notation

The following notation is used from this section onward:

- d denotes the dimension of latent embeddings and codebook entries.
- K denotes the total number of vector entries in a learned codebook, i.e. codebook $\in \mathbb{R}^{K \times d}$.
- L denotes the number of codes used to represent each image (fixed for a given training run).
- x and $\hat{x}$ denote an input image and its reconstruction, respectively.
- C_x denotes the set of integers that is the discrete coded representation of x, such that $|C_x| = L$.
- $\binom{n}{r}$ denotes the binomial coefficient used in combinatorics [14], defined as $\frac{n!}{r!(n-r)!}$.

3.2 Permutation-Invariant Autoencoder

The first step in formulating the PI-VQ model is to identify a differentiable sequence of operations which can map the input space to a set of vectors (which lack explicit positional information) and then back to the input space. In addition to the improved generalization and robustness of permutation-invariant representations [30, 45], this architectural constraint forces codes to capture global information about an image. This is particularly appropriate for aligned faces, where many salient features are inherently global (lighting, skin-tone, expression) rather than spatially localized.

To achieve permutation invariance, we exploit the fact that the transformer encoder architecture is invariant to the permutation of its inputs by design [47]. We intentionally omit positional embeddings when supplying latent codes to the decoder (Fig. 2). Enforcing this architectural constraint guarantees that the decoding step is unaffected by the permutation of latent codes.

Below is a description of how the encoder, quantizer and decoder work together to extract discrete permutation-invariant representations of images. Fig. 2 aids in illustrating how information flows through the architecture.

1. **Encoder**: The input image x first goes through a series of convolutional down-sampling and residual layers to produce a $W \times H \times d$ tensor. This is then "flattened" into a sequence of $(W \times H)$ vectors of dimension d. Input positional embeddings are added, and a further sequence of L learned positional embeddings (also of dimension d) are concatenated. The new sequence is then fed to a transformer encoder, of which the final L output elements are the latent codes.
2. **Quantizer**: Each vector in the latent codes is mapped to a similar element of the codebook (Sections 3.3 and 3.4).
3. **Decoder**: The quantized latent vectors are concatenated to a sequence of learned output embeddings. The new sequence is fed to a transformer encoder. Of the resulting output sequence, the elements following the first L are reshaped and fed to a series of convolutional up-sampling and residual layers, resulting in the reconstructed image.

3.3 Vector Quantization

Having identified a suitable architecture for permutation-invariant autoencoder, the next step is to *quantize* the latent representations using a learned mapping $\mathbb{R}^d \to \mathbb{Z}$. The original VQ-VAE [37] uses nearest-neighbour vector quantization, in which encoder outputs are mapped to their nearest neighbour in a learned codebook. Specifically, for each position-free encoder output vector z_e, the corresponding quantized vector is computed as the nearest codebook entry e_c, where:

$$c = \arg\min_j ||z_e - e_j||_2 \tag{1}$$

where $e_0, e_1 ... e_{K-1}$ are entries in a learned vector codebook of length K.

Since the quantization step is non-differentiable, it is necessary to estimate the gradients during backpropagation. For this purpose, we use straight-through gradient estimation [37], whereby the gradients are copied directly from the decoder input z_c to the encoder output z_e.

3.4 Matching Quantization for Improved Information Bottleneck

Naive formulations of VQ-family approaches [37] use the "nearest-neighbour" quantization method [1] described above. Applying the nearest-neighbour approach alongside PI-VQ, we find in practice that the maximum per-sample codebook usage K_{img} is much smaller than the number of codes per sample L by the end of training (Table 2). This happens because the quantization approach often assigns many embedding vectors to the same codebook entry (Figure 3), resulting

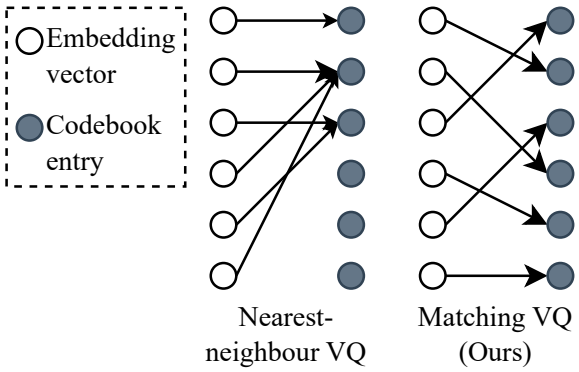

Fig. 3. Matching quantization: Our novel approach to vector quantization (right) ensures that no two embedding vectors are mapped to the same codebook element in a given image, effectively minimising redundancy in the discrete representation.

in redundancy due to repetition. This upper-bounds the per-image information capacity of the latent bottleneck at:

$$B = \log_2 \left[\binom{K_{\text{data}}}{K_{\text{img}}} \times \binom{L + K_{\text{img}} - 1}{K_{\text{img}} - 1} \right] \text{bits}, \tag{2}$$

where K_{data} is the actual codebook usage over the whole dataset (see supplemental for proof). In practice, nearest-neighbour quantization results in a very low value of K_{img} (Section 4), so the value of (2) is only a fraction of the theoretical maximum (see supplemental for proof) of around 2313 bits ($K = 4096$ and $L = 512$).

To remedy this practical limitation, we propose *matching quantization* as a robust means of improving per-sample codebook usage and the corresponding information capacity. Instead of mapping each latent embedding to its nearest neighbour in the learned codebook, we instead find the minimal-cost *one-to-one matching* between the set of latent embeddings and the elements of the codebook (where cost is the total euclidean distance between latent codes and their matching codebook vector). Formally, this can be expressed as finding the optimal permutation $P \in \{0, 1\}^{K \times K}$ of the distance matrix $D \in \mathbb{R}^{K \times L}$ where D_{ij} is the euclidean distance between the codebook entry i and latent embedding j [6]:

$$P^* = \arg\min_P \text{Tr}(PD). \tag{3}$$

The row index i of the 1 in the j^{th} column of P is then the index of the codebook entry to which latent embedding j is mapped. In practice, we use the Hungarian algorithm [6] to compute the optimal matching, which is an $\mathcal{O}(n^3)$ algorithm where $n = \max(L, K)$ (compare with the nearest-neighbour quantization which is $\mathcal{O}(n^2)$). We observe that the added complexity of matching

quantization (during both training and inference) is a necessary trade-off for the increased effective information capacity.

This approach to vector quantization ensures that the per-image codebook usage is always exactly the size of the representation L for as long as $K \geq L$, i.e., each distinct code is used no more than once per image representation. The corresponding theoretical information capacity is then:

$$B = \log_2 \left[\binom{K_{\text{data}}}{L} \right] \text{ bits}, \tag{4}$$

where K_{data} is the actual codebook usage over the dataset observed at the end of training (typically smaller than the hyperparameter K [27]). This theoretical upper bound on information capacity is derived from the fact that there are K_{data} codebook entries from which to choose when representing an image, of which L are chosen to represent a single image, resulting in a theoretical representation space of cardinality $\binom{K_{\text{data}}}{L}$.

We compare information capacities from different settings by plotting the upper bound for both quantization methods on the CelebA-HQ dataset, with the standard VQ-VAE capacity included for reference 4. Figure 4 also indicates that PI-VQ information capacity is strictly smaller than standard VQ-family approaches. We address this practically by setting a higher L and K in our experiments (Section 4) than is used in the literature [12] for a comparable VQ-VAE or VQ-GAN training run, resulting in an overall similar information capacity.

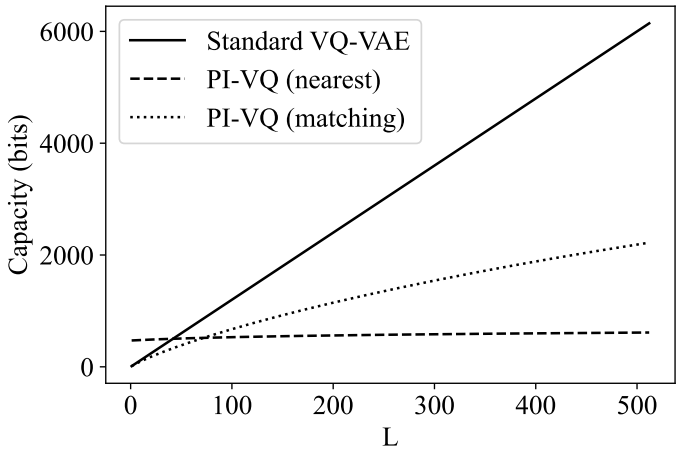

Fig. 4. Information capacity (in bits) in terms of representation length L of three approaches with codebook size $K = 4096$: standard VQ-VAE with no permutation invariance; PI-VQ with nearest neighbour quantization ($K_{\text{img}} = 49$); PI-VQ with proposed matching quantization.

Algorithm 1 Fast interpolation-based sampling

```
procedure INTERPOLATE(Ca[1..L], Cb[1..L])
    Cc = set intersect of Ca and Cb
    Cd = symmetric set difference of Ca and Cb
    Cinterp = Cc
    for j = 1 to L-length(Cc) do
       c = uniform random element from Cd
       add c to Cinterp
       remove c from Cd
    end for
    return Cinterp
end procedure
```

3.5 Fast Interpolation-Based Sampling

Besides increasing the latent bottleneck's information capacity, matching quantization facilitates a simple yet effective approach to sampling from the space of learned representations. Naively, we assume that the set of latent codes extracted from a given face image is representative of the set of visual *traits* present in the image. Under this assumption, we identify the *common traits* between two coded representations C_a and C_b (for images a and b respectively as simply the intersection: $C_c = C_a \cap C_b$, and the *exclusive traits* as the codes in either set but not in both: $C_d = (C_a \setminus C_b) \cup (C_b \setminus C_a)$.

Defining C_c and C_d in this way, it is possible to sample from large (but finite) space of equally valid interpolations between a and b by first initialising $C_{interp} := C_c$, then iteratively sampling elements from C_d (without replacement) and adding them to C_{interp} until $|C_{interp}| = L$. At this point, C_{interp} forms a new representation comprising all the features shared by C_a and C_b as well as a mixture of features which are not shared. C_{interp} is finally passed to the permutation-invariant decoder to produce an interpolated image x_{inter} .

It is also possible to interpolate (approximately) smoothly from one image to another using a similar process: Compute C_c as before. Construct two arrays (*not* sets) R_a and R_b of equal length $|R|$ containing the elements in C_a but not in C_b and vice versa, respectively. Choose a random permutation of each array R'_a and R'_b respectively. Construct interpolated representation C^t_{smooth} for parameter $0 < t < |R|$ as the union of (C_c, the first t elements of R_a, the final $|R| - t$ elements of $R - b$). This way, there exists a unique interpolated path through the latent space for each unique pair of permutations R'_a and R'_b.

3.6 Training

We follow largely the same training procedure as [5]. The model is trained to jointly minimise the L1 reconstruction loss, LPIPS perceptual loss [54] (a learned perceptual distance function based on a pre-trained CNN classifier), and a learned discriminator loss [16]. The discriminator is trained alongside the

decoder to distinguish reconstructed images from real images. We employ differentiable augmentations (DiffAug) [56] and adaptive weight limiting [52], since they have been shown to further boost reconstruction and sample quality in generative adversarial models.

3.7 Delayed Codebook Initialization

In practice, applying vector-quantization from the beginning of the training run leads to a catastrophic failure in learning, with the reconstructions remaining unrecognisable even at the end of training. To solve this problem, we delay the application of the vector quantization step until training is already partially underway. This follows earlier recommendations for the robust training of VQ-VAE models [27].

At a pre-determined training iteration T_q, we apply data-dependent codebook initialization using KMeans++ [2]. The cluster centroids are fit to a history of encoder outputs over the most recent W training iterations, where W is a hyper-parameter chosen to trade-off time complexity against coverage of the latent space. In practice (Section 4), we re-initialize the codebook more than once over successive windows of length W, finding that repeating codebook initialization in this way leads to better codebook usage.

3.8 Probing the Learned Representations

After training a PI-VQ model to auto-encode images, we wish to probe the learned representations for interpretable or semantic content. To achieve this, we train a binary logistic regression model for each of 15 ground-truth binary attributes (Fig. 5 of FFHQ. Where the raw annotations are not binary (e.g. age), binary attributes are derived from raw annotations, e.g. "age_senior". We treat the presence or absence of each latent code as a separate input variable (taking on the value 0 or 1 respectively), totalling one input variable for each code in the codebook. As a baseline predictor, we predict the most frequent label for each attribute (e.g. “not has_beard”). Section 4 discusses the findings of our probing experiment.

4 Results and Discussion

4.1 Setup

Datasets: We repeat all our training runs for three face datasets : CelebA 64x64 [31], CelebaA-HQ 256x256 [24] and FFHQ 256x256 [22]. We evaluate sample quality on the three datasets in terms of FID [18], precision, recall [42], density and coverage [35]. We compare against results reported in the literature where available. We note that some scores for density and coverage are missing in our comparisons with the state-of-the-art in cases where results were not reported, and we are unable to reproduce the results ourselves with our available computational resources.

Quantization method	K_{img} (max) ↑	Capacity (bits) ↑
Nearest	49	614
Matching (ours)	**512**	**2221**

Table 2. Upper bound of latent bottleneck information capacity: comparison between nearest-neighbour and matching quantization for models trained on CelebA-HQ. K_{img} is included as the maximum per-image codebook usage observed over the entire dataset. Capacity is rounded to the nearest bit.

Architecture: In the autoencoder, we use the same convolutional down-sampling and up-sampling architecture as [5] in the encoder and decoder respectively. We additionally include a 4-layer transformer encoder following the down-sampling and preceding the up-sampling modules, as illustrated in Fig. 2. The transformer embedding dimension is 256 and the FFN hidden dimension is 1024 (128 and 512 respectively for CelebA 64x64). We use the same discriminator and differentiable augmentations as [5], since these are shown to improve visual quality.

Hyperparameters: The transformer layers have a feedforward dimension equal to 4 times the input dimension. We use a relatively small batch size of 4 (increasing to 16 for CelebA 64x64) with exponential model averaging with $beta = 0.995$ following [5]. We choose code length $L = 512$ and codebook size $K = 4096$ such that the information bottleneck is comparable to a conventional VQ-GAN with similar architecture (such as the one used in [5]): a conventional discrete bottleneck with $L = 256$ latent codes and codebook size $K = 1024$ corresponds to $256 \log_2 1024 = 2560$ bits of information. A permutation-invariant discrete bottleneck with $L = 512$ and $K = 4096$ achieves a comparable bottleneck of approximately 2221 bits, computed from our derived upper bound (Equation 4, Figure 4). We reduce this to $L = 32$ and $K = 512$ for CelebA 64x64 due to the much smaller image size. Further training details are in supplemental.

4.2 Codebook Usage and Bottleneck Capacity

In order to verify empirically that our proposed matching quantization technique increases effective information capacity, we train a model on CelebA with identical settings, with the exception that the "nearest neighbour" vector quantization method is used in place of matching quantization. We find that the maximum per-image codebook usage K_{img} over the entire dataset is only 49 out of the maximum 512 when the "nearest" approach is used, indicating a large number of repeated codes per image. In comparison, the proposed "matching" technique achieves maximal per-image codebook usage by design (equal to L, which is 512 for our experiments), resulting in over 3.5× the effective upper limit on information capacity (Table 2).

Method	FID↓	P↑	R↑	D↑	C ↑
FFHQ 256x256					
StyleGAN-XL [43]	**2.2**	0.80	0.39	0.86	0.73
StyleGAN2 [23]	3.8	0.69	0.40	1.12	**0.80**
VQ-GAN (UT) [5]	7.1	0.69	**0.48**	1.06	0.77
VQ-GAN (TT) [12]	9.6	0.64	0.29	0.89	0.59
VDVAE [8]	28.5	0.59	0.20	0 .80	0.50
PR-BigGAN (R) [48]	35.2	0.78	0.10	0.89	0.60
PR-BigGAN (P) [48]	38.2	**0.84**	0.08	**1.15**	0.63
BigGAN [11]	41.4	0.66	0.10	0.52	0.47
Ours	23.4	0.69	0.19	1.02	0.55
CelebA 64x64					
ADM-IP [36]	**1.5**	0.23	**0.65**	0.88	0.24
PR-BigGAN (R) [48]	6.0	0.78	0.56	0.88	**0.50**
BigGAN [11]	9.2	0.78	0.51	0.89	0.48
PR-BigGAN (P) [48]	22.5	**0.84**	0.26	**1.21**	0.43
Ours	73.2	0.58	0.23	0.60	0.44
CelebA-HQ 256x256					
RDM [46]	**3.2**	0.77	**0.55**	-	-
LDM-4 [41]	5.1	0.72	0.49	-	-
WaveDiff [38]	5.9	-	0.37	-	-
DDGAN [50]	7.6	-	0.36	-	-
Ours	22.8	**0.85**	0.10	1.88	0.69
Average (SOTA)	2.3	0.82	0.56	1.18*	0.65*
Average (Ours)	39.8	0.71	0.17	0.81*	0.50*

Table 3. Comparison with state-of-the-art in FID, precision (P), recall (R), density (D), and coverage (C) scores for image quality. Best result for each dataset and metric is in bold. *Averages are taken over only the results where both SOTA and our results are available for comparison.

4.3 Sample Quality and Diversity

To investigate whether the learned permutation-invariant representations support coherent image synthesis, we evaluate interpolation-based sampling using 5 metrics across 50,000 sampled images [1] synthesised from each of our 3 models trained with matching quantization: FID [18], precision, recall [42], density, and coverage [35]. Where available, we compare these values against state-of-the-art results reported in the literature. We choose to evaluate along a broad range of metrics in order to facilitate better comparisons deeper discussion.

We observe that PI-VQ approaches competitive scores in precision, density and coverage, achieving neither the best nor the worst result for each dataset across the majority of metrics. A more in-depth interpretation of these scores is given below:

1. The high density scores on CelebA-HQ and FFHQ indicate that the model is good at sampling from regions of the dataset which are densely packed [35]. It

[1] In the case of CelebA-HQ, we only sample 30,000 images since the reference dataset has only 30,000 ground-truth images with which to compare.

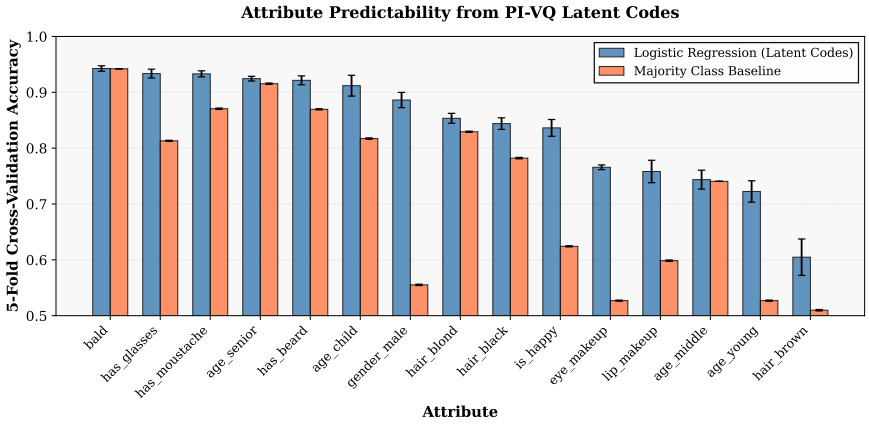

Fig. 5. Logistic regression accuracy with error bars (5-way cross-validated) for predicting ground-truth FFHQ annotations based on learned permutation-invariant representations. For each attribute, we compare against the baseline accuracy (always predicting the majority class).

should be noted that $\mathbb{E}(density) = 1$ if the synthetic distribution is identical to the real distribution [35], so our model achieves a *higher sampler density than the underlying data* on the 256x256 resolution datasets.

2. Coverage measures the fraction of real samples whose neighbourhoods contain at least one synthetic sample [35]. The high coverage in the case of our method indicates that the model has good coverage of the data.
3. The lower FID and recall scores suggest that the permutation-invariant bottleneck may limit the representation's capacity to capture the full diversity of the data distribution. We discuss this trade-off in Section 5.

4.4 Qualitative Results

In this subsection, we examine the structure of the learned representation space by visualising interpolations between encoded images. We also demonstrate that our approach, while being fundamentally discrete in nature, can approximate a smooth interpolation between two images.

Figure 1 shows 2 valid interpolations between two pairs of images (for a total of 4 interpolations) sampled using our smooth interpolation approach. The model successfully captures multiple plausible paths through the representation space, demonstrating that while individual codes are discrete, the combinatorial structure of the latent space permits approximate smoothness. This suggests the learned codes capture separable, compositional visual features. The number of possible interpolation paths for a given image pair is equal to $(|R|!)^2$, which likely contributes to the ability to sample densely from the 256x256 datasets.

To further probe the semantic content of the learned codebook entries, we fit logistic regression models to predict the ground-truth FFHQ annotations,

treating the presence or absence of each unique latent code as a dummy binary variable (for a total number of variables equal to the codebook size). Figure 5 indicates the 5-fold cross-validation accuracy of each regression model for the top 15 most "predictable" attributes, side-by-side with the majority class predictor (e.g. always predicting "not bald" as the most frequent annotation for hair presentation in FFHQ).

We find that, for certain attributes such as "gender_male", "is_happy" and "eye_makeup", the logistic regression is significantly more accurate than the baseline. This indicates that the semantic codes capture separable, interpretable features which are predictive of human-annotated attributes, both global (e.g. gender) and local (e.g. eye_makeup). Meanwhile, for some attributes, the logistic regression model is not much better than predicting the majority class (e.g. "bald", "age_middle", "age_senior"), indicating that human annotations aren't universally captured by the learned codes in a linearly separable way.

5 Limitations

The desirable properties demonstrated in Section 4 (direct interpolation, $\mathcal{O}(1)$-time sampling, global codes) follow directly from the intended constraints of the proposed architecture. However, the design of our permutation-invariant method comes with a number of limitations, and our qualitative results reflect this trade-off.

We observe that while precision, density and coverage scores are comparable to the state-of-the-art, our method falls behind on FID and recall. The low recall (indicating mode-dropping) is likely a consequence of two factors. The permutation-invariant bottleneck has fundamentally lower information capacity than position-dependent alternatives, limiting the representation's ability to capture rare or fine-grained features. While the proposed matching quantization technique is effective in improving the effective information content of latent codes, it does not entirely prevent codebook collapse. Future work could combine our approach with techniques targeting codebook collapse directly.

Finally, the permutation-invariance constraint is inherently suited to global features and may struggle with fine-grained spatial details, which could contribute to perceptual differences captured by FID. Our evaluation focuses on aligned face datasets, where spatial structure is fixed. The approach is best understood as complementary to standard VQ methods—appropriate when global, interpretable discrete features are prioritised over spatial fidelity.

6 Conclusion

We have investigated the consequences of removing positional information from discrete image representations. The proposed PI-VQ architecture enforces permutation-invariance on learned codes, and this constraint encourages codes to capture

global, semantic features rather than position-dependent patterns. As a consequence, the learned representations support direct interpolation between images and $\mathcal{O}(1)$ sampling without requiring a learned prior.

We introduced matching quantization as an alternative to nearest-neighbour vector quantization suited to the permutation-invariant setting, increasing the effective information capacity of the bottleneck by a factor of 3.5×. Our experiments on aligned face datasets demonstrate competitive precision, density and coverage metrics, while the worse recall and FID point to inherent trade-offs in position-free representations.

Future work could explore improved codebook utilisation, extension to higher resolutions, and application to other spatially-aligned domains beyond faces. Permutation-invariant discrete representation learning setting may also prove suitable for other modalities with global, compositional features. Our work aims to provide a useful foundation for further exploration of position-free discrete representation learning.

References

1. Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., Van Gool, L.: Soft-to-hard vector quantization for end-to-end learning compressible representations. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 1141–1151. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
2. Arthur, D., Vassilvitskii, S.: K-means++ the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035 (2007)
3. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6713–6722 (2018)
4. Baykal, G., Kandemir, M., Unal, G.: Edvae: Mitigating codebook collapse with evidential discrete variational autoencoders. Pattern Recognition **156**, 110792 (2024)
5. Bond-Taylor, S., Hessey, P., Sasaki, H., Breckon, T.P., Willcocks, C.G.: Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII. p. 170–188. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20050-2_11
6. Bruff, D.: The assignment problem and the hungarian method. Notes for Math **20**(29-47), 5 (2005)
7. Caselles-Dupré, H., Garcia Ortiz, M., Filliat, D.: Symmetry-based disentangled representation learning requires interaction with environments. Advances in Neural Information Processing Systems **32** (2019)
8. Child, R.: Very deep vaes generalize autoregressive models and can outperform them on images. arXiv preprint arXiv:2011.10650 (2020)
9. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv preprint arXiv:2307.08698 (2023)
10. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 5154–5163 (2020)
11. Donahue, J., Simonyan, K.: Large scale adversarial representation learning. Advances in neural information processing systems **32** (2019)
12. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021)
13. Fifty, C., Junkins, R.G., Duan, D., Iyengar, A., Liu, J.W., Amid, E., Thrun, S., Ré, C.: Restructuring vector quantization with the rotation trick. arXiv preprint arXiv:2410.06424 (2024)
14. Flajolet, P., Sedgewick, R.: Analytic combinatorics. cambridge University press (2009)
15. Gauthier, J.: Conditional generative adversarial nets for convolutional face generation. Class project for Stanford CS231N: convolutional neural networks for visual recognition, Winter semester **2014**(5), 2 (2014)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
17. Goswami, N., Mukuta, Y., Harada, T.: Hypervq: Mlr-based vector quantization in hyperbolic space. arXiv preprint arXiv:2403.13015 (2024)
18. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
19. Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., Lerchner, A.: Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230 (2018)
20. Higgins, I., Matthey, L., Pal, A., Burgess, C.P., Glorot, X., Botvinick, M.M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2016)
21. Kalra, S., Adnan, M., Taylor, G., Tizhoosh, H.R.: Learning permutation invariant representations using memory networks. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 677–693. Springer International Publishing, Cham (2020)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. IEEE Transactions on Pattern Analysis & Machine Intelligence **43**(12), 4217–4228 (dec 2021). https://doi.org/10.1109/TPAMI.2020.2970919
23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8107–8116. IEEE Computer Society, Los Alamitos, CA, USA (jun 2020). https://doi.org/10.1109/CVPR42600.2020.00813, https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00813
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
25. Keysers, D., Schärli, N., Scales, N., Buisman, H., Furrer, D., Kashubin, S., Momchev, N., Sinopalnikov, D., Stafiniak, L., Tihon, T., et al.: Measuring compositional generalization: A comprehensive method on realistic data. arXiv preprint arXiv:1912.09713 (2019)

26. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
27. Łańcucki, A., Chorowski, J., Sanchez, G., Marxer, R., Chen, N., Dolfing, H.J., Khurana, S., Alumäe, T., Laurent, A.: Robust training of vector quantized bottleneck models. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2020)
28. Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., Teh, Y.W.: Set transformer: A framework for attention-based permutation-invariant neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3744–3753. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/lee19d.html
29. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
30. Liu, R., Zhe, T., Fu, Y., Xia, F., Wang, D., et al.: Permutation-invariant representation learning for robust and privacy-preserving feature selection. arXiv preprint arXiv:2510.05535 (2025)
31. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
32. Lu, Y., Tai, Y.W., Tang, C.K.: Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European conference on computer vision (ECCV). pp. 282–297 (2018)
33. Luo, Y., Zhu, J., He, K., Chu, W., Tai, Y., Wang, C., Yan, J.: Styleface: Towards identity-disentangled face generation on megapixels. In: European Conference on Computer Vision. pp. 297–312. Springer (2022)
34. Murphy, R.L., Srinivasan, B., Rao, V., Ribeiro, B.: Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. arXiv preprint arXiv:1811.01900 (2018)
35. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning. pp. 7176–7185. PMLR (2020)
36. Ning, M., Sangineto, E., Porrello, A., Calderara, S., Cucchiara, R.: Input perturbation reduces exposure bias in diffusion models. arXiv preprint arXiv:2301.11706 (2023)
37. van den Oord, A., Vinyals, O., kavukcuoglu, k.: Neural discrete representation learning. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf
38. Phung, H., Dao, Q., Tran, A.: Wavelet diffusion models are fast and scalable image generators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10199–10208 (June 2023)
39. Quirke, L., Shabalin, S., Belrose, N.: Binary sparse coding for interpretability. arXiv preprint arXiv:2509.25596 (2025)
40. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Advances in neural information processing systems **32** (2019)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
42. Sajjadi, M.S., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. Advances in neural information processing systems **31** (2018)
43. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022)
44. Tan, S., Ji, B., Bi, M., Pan, Y.: Edtalk: Efficient disentanglement for emotional talking head synthesis. In: European Conference on Computer Vision. pp. 398–416. Springer (2024)
45. Tang, Y., Ha, D.: The sensory neuron as a transformer: Permutation-invariant neural networks for reinforcement learning. Advances in Neural Information Processing Systems **34**, 22574–22587 (2021)
46. Teng, J., Zheng, W., Ding, M., Hong, W., Wangni, J., Yang, Z., Tang, J.: Relay diffusion: Unifying diffusion process across resolutions for image synthesis. arXiv preprint arXiv:2309.03350 (2023)
47. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
48. Verine, A., Negrevergne, B., Pydi, M.S., Chevaleyre, Y.: Precision-recall divergence optimization for generative modeling with gans and normalizing flows. arXiv preprint arXiv:2305.18910 (2023)
49. Wang, X., Chen, H., Tang, S., Wu, Z., Zhu, W.: Disentangled representation learning. arXiv preprint arXiv:2211.11695 (2022)
50. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the generative learning trilemma with denoising diffusion gans. arXiv preprint arXiv:2112.07804 (2021)
51. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved VQGAN. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=pfNyExj7z2`
52. Zadorozhnyy, V., Cheng, Q., Ye, Q.: Adaptive weighted discriminator for training generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4781–4790 (2021)
53. Zhang, B., Zheng, W., Zhou, J., Lu, J.: Preventing local pitfalls in vector quantization via optimal transport. arXiv preprint arXiv:2412.15195 (2024)
54. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
55. Zhao, J., Xiong, L., Karlekar Jayashree, P., Li, J., Zhao, F., Wang, Z., Sugiri Pranata, P., Shengmei Shen, P., Yan, S., Feng, J.: Dual-agent gans for photorealistic and identity preserving profile face synthesis. Advances in neural information processing systems **30** (2017)
56. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. In: Conference on Neural Information Processing Systems (NeurIPS) (2020)