Initial report: Discrete Diffusion for Anatomical Shape Generation

Miguel Ángel Rodríguez Fuentes
Télécom Paris, Institut Polytechnique de Paris
miguelangel.fuentes@telecom-paris.fr

Project repository: github.com/Marodriguezfu/PRIM-Discrete-Difusion-for-Anatomical-Shape-Generation

1 Introduction

This preliminary phase of the PRIM project, "Discrete Diffusion for Anatomical Shape Generation", aims to evaluate whether quantized vector autoencoder models can reliably learn discrete latent representations of anatomical structures, essential for the subsequent application of discrete diffusion to tokenized shapes. To this end, we have adapted a VQ-UNet-based architecture for prostate MRI image segmentation using volumetric data from the PROSTATE-DIAGNOSIS and Prostate-3T collections [1, 2], which provide high-quality T2-weighted axial scans with expertly annotated central gland and peripheral zone contours. Since the project's ultimate goal requires coherent 3D anatomical representations rather than isolated 2D slices, the experiments focus on full volumetric processing to capture in-plane continuity and morphological consistency. Using an existing public repository, we implemented and trained the VQUNet3Dposv3 model, integrating a U-Net 3D base architecture, residual blocks, positional encoding, and a vector quantization bottleneck [3] to generate discrete embeddings. This report summarizes the adapted implementation, experimental setup, and initial segmentation results, and serves as an assessment of the feasibility of using vector quantization-based tokenization as a foundation for future diffusion-based anatomical shape generation [4].

2 Baseline Code: Description and Functioning

The initial implementation used in this work is based on Ainkaran Santhi's public repository *Vector-Quantisation-for-Robust-Segmentation*, which implements vector-quantized autoencoder architectures for medical image segmentation. The repository was originally designed for binary segmentation tasks, where anatomical subregions (such as the transition zone and peripheral zone in prostate MRI) are merged into a single foreground class. This configuration naturally leads to higher Dice scores compared to multiclass segmentation, as the network only needs to distinguish the foreground from the background.

2.1 Repository Overview

The codebase contains four main components:

- **Dataset and Preprocessing:** A data loading flow for volumetric medical images that includes intensity normalization, spatial resampling, fill/crop to a fixed input size, and basic geometric zooms. Labels are converted to a binary mask.
- **Model Architecture:** A 3D VQ-UNet integrating a U-Net architecture, ResNet blocks, a convolutional bottleneck, and a vector quantization module. The latter discretizes the encoder features using a learned code dictionary, enabling compact latent representations.
- **Training Loop:** A PyTorch Lightning training script that combines segmentation losses with the embedding loss produced by the vector quantizer. Validation is performed periodically, and the best model is saved based on its Dice validation score.
- **Evaluation Metrics:** The original repository reports the Dice coefficient, Hausdorff distance (HD), and average surface distance (ASD), all calculated in binary format, which simplifies the segmentation process.

Division of the original dataset. In the original repository, the training and validation volumes came from BMC (BioMed Central), while the test volumes were obtained exclusively from RUNMC (Radboud University Nijmegen Medical Centre). Specifically, the provided `train.csv` and `validation.csv` files contained only cases from BMC, and the `test.csv` file contained only cases from RUNMC. This division may be problematic since the acquisition protocols for each database and/or domain are unknown and could differ. Additionally, the number of cases per division was not explicitly balanced with respect to the anatomical characteristics of the prostate regions.

2.2 Main Components of the Baseline

DataLoader and preprocessing. The dataloader reads 3D medical image volumes and their corresponding masks, applies normalisation and spatial preprocessing, and performs data augmentation during training. All labels are treated as a single prostate class, combining TZ and PZ.

Model architecture. The backbone follows an encoder–quantiser–decoder structure. The encoder is built from convolutional and residual blocks with progressive downsampling. The latent features are projected into an embedding dimension and passed through the `VectorQuantizer2` module [3], which replaces each latent vector with its nearest neighbour in the learned codebook. The decoder upsamples the quantised latents, integrates skip connections, and reconstructs the segmentation map.

Losses and metrics. The training objective combines segmentation loss with the embedding loss that aligns encoder outputs with codebook entries. The original repository evaluates segmentation using Dice, HD, and ASD metrics in the binary setting.

Training and inference pipeline. The pipeline includes forward propagation, loss computation, gradient updates, validation loops, and model checkpointing. During inference, the best-performing checkpoint is loaded and used to generate binary segmentation masks saved as NIfTI volumes.

3 Modifications Implemented in This Work

This section summarizes all the modifications made compared to the original repository *Vector-Quantisation-for-Robust-Segmentation* [5]. The changes encompass the data flow, training procedure, and recording and visualization workflow. These adaptations were necessary to support full three-class volumetric prostate segmentation, ensure the correct calculation of metrics, optimize GPU memory usage, and enable reproducible experiments for subsequent discrete diffusion modeling.

3.1 Pipeline Modifications

Preprocessing Standardization Preprocessing was standardized for training, validation, and testing:

- Normalization of intensity using `NormalizeIntensityd(nonzero=True,channel_wise=True)`,
- Harmonized data augmentation for volumetric inputs,
- Correction of inconsistencies present in the base implementation, such as the removal of data that did not have all the required labels.

Class Management Unlike the binary configuration of the original repository, the adapted pipeline uses a **three-class segmentation** (Background, TZ, PZ). The masks are maintained in full multiclass format and converted to one-hot encoding during training.

Reorganization of the dataset splits. To obtain a more balanced and statistically robust evaluation protocol, the original center-based split (BMC for training/validation and RUNMC for testing) was replaced with a new split generated using a custom script `make_balanced_split_by_volume.py`. This script:

- reads one or more existing CSV files (`train.csv`, `validation.csv`, `test.csv`, or a master list),
- loads the corresponding segmentation masks and calculates the voxel count and volumes (in mm^3) for the TZ (label 1) and PZ (label 2),
- assigns each case a patient identifier inferred from the file path,
- creates volume-based intervals for the TZ and PZ using quantiles,
- performs stratified splitting by patient, aiming for a 55/12/12 ratio for training/validation/testing.

The resulting `train.csv`, `validation.csv`, and `test.csv` files therefore contain a combination of BMC and RUNMC cases, balanced between the TZ/PZ volumes and grouped by patient. This avoids center-specific biases and ensures that all results presented in this work are based on a consistent and balanced data partitioning.

3.2 Architecture Modifications

Codebook and Embedding Dimension Adjustments. To balance expressiveness and GPU feasibility [3]:

- The codebook size was set to $n_{\text{embed}} = 1024$,
- The embedding dimension was set to 256,
- Quantization Convolution was aligned with the latent channel depth.

GPU Memory Optimization. To avoid out-of-memory errors during 3D training, the Transformer module was dropped and replaced with positional encoding, which provides global spatial context at a minimal memory cost.

3.3 Training Modifications

Correct Calculation of Evaluation Metrics. The calculation of metrics was redesigned using validated MONAI utilities:

- `Dice Metric` for overlap accuracy,
- `calculate Hausdorff distance`,
- `calculate average surface distance`.

All metrics are calculated in separate batches, with `include_background=-False` to focus on prostate zones.

Improved Control Points. The control point filenames were expanded to include the epoch and metric values, for example, `best_dice_epochXX_dice0.XXXX`, allowing for clearer tracking of experiments.

Corrections to `validation_epoch_end`. Several issues were corrected:

- Correct average across all validation samples,
- Proper reset of the Dice accumulator,
- Consistent tracking of the best Dice and Multi metrics.

EPS Generation. The plotting routines were adapted to allow exporting training curves in EPS format for inclusion in scientific reports.

3.4 Logging and Visualization Modifications

Improved TensorBoard Logging. Additional logs include:

- training and validation losses,
- Dice, HD, ASD, and multimetric curves,
- epoch-wise metric evolution.

Visual Outputs. During testing, the implementation now exports [6]:

- input volumes,
- predicted segmentation masks,
- reference labels,
- all saved as NIfTI files for 3D inspection.

Structured test results. Test metrics are aggregated into a CSV file, and the rebuilt volumes are stored in a well-organized directory structure, facilitating both qualitative and quantitative analysis.

4 Network Architecture

The model used in this work is a volumetric U-Net with vector quantization, implemented as `VQUNet3Dposv3`. The architecture follows an encoder-quantizer-decoder design that combines 3D convolutional feature extraction, residual learning, discrete latent representation via a vector quantization bottleneck [3], and residual connections to preserve fine spatial details. With this configuration, it is possible to evaluate whether VQ-based tokenization can maintain the 3D anatomical structure of the prostate while remaining computationally viable for subsequent discrete diffusion modeling [4].

Encoder The encoder consists of four subsampling levels, each composed of a 3D convolution followed by a ResNet block. The number of channels progressively doubles at each stage ($16 \to 32 \to 64 \to 128 \to 256$), allowing the network to capture increasingly abstract volumetric features. The deepest layer applies a final ResNet block before projecting the feature tensor into an embedding dimension of 256 using a $1 \times 1 \times 1$ convolution.

Positional Encoding Before reaching the quantization bottleneck, the latent tensor is processed by a 3D positional encoding module. This component injects explicit spatial information into the embedded features, helping the model preserve geometric consistency across slices. Since the architecture does not include the internal Transformer present in other VQ-VAE variants, positional encoding serves as a lightweight mechanism for encoding the global spatial context without significantly increasing GPU memory consumption.

Vector Quantization Bottleneck. The integrated latent representation is discretized using the `VectorQuantizer2` module [3] with a dictionary of $n_{\text{embed}} = 1024$ entries of dimension 256. Each latent vector is replaced by the nearest dictionary entry, generating a spatial grid of discrete tokens. The loss of integration encourages stable dictionary usage and alignment between encoder features and dictionary prototypes.

Decoder. The decoder replicates the encoder through four oversampling stages, concatenating the oversampled features with the corresponding encoder activations at each stage via jump connections. This structure allows the model to combine high-level quantized information with low-level spatial details, facilitating accurate reconstruction of the multiclass prostate segmentation mask. Residual blocks at each decoding stage improve representational capacity and stabilize training.

Output Layer A final $1 \times 1 \times 1$ convolution assigns the decoded features to three output channels, corresponding to the background, transition zone (TZ), and peripheral zone (PZ). While the original repository was designed for binary segmentation (foreground vs. background) [5], the adapted model performs full three-class segmentation, which naturally results in lower segmentation metrics compared to the binary configuration.

Justification for Architectural Decisions The removal of the Transformer and the inclusion of 3D positional encoding were due to GPU memory limitations and the goal of achieving a stable, lightweight, and complete quantized 3D representation. The chosen codebook size and embedding dimension offer a balance between expressive power and ease of training. Overall, the architecture represents a practical solution: expressive enough to model the anatomy of the prostate in 3D, yet compact enough to serve as an initial feasibility study for future discrete diffusion experiments [4].

5 Evaluation Metrics

To evaluate the performance of the `VQUNet3Dposv3` model in volumetric prostate segmentation, four complementary evaluation metrics were used: the Dice coefficient (DICE), the Hausdorff distance (HD), the average surface distance (ASD), and a weighted multimetric score. Together, these metrics capture both regional overlap and boundary accuracy, providing a comprehensive view of segmentation quality.

5.1 DICE Coefficient (DICE)

The DICE coefficient measures the spatial overlap between the predicted mask and the actual segmentation:

$$\text{DICE} = \frac{2|A \cap B|}{|A| + |B|}.$$

It is the most widely used metric in medical image segmentation because it directly quantifies the proportion of the anatomical region that is correctly reconstructed. DICE is particularly appropriate for the prostate, where

the Transition Zone (TZ) and Peripheral Zone (PZ) may have unbalanced sizes. However, DICE is sensitive to small structures: misclassifying even a few voxels in a small region can lead to a large decrease in the score. This sensitivity is essential for detecting segmentation failures that might not be visible solely through loss values.

5.2 Hausdorff Distance (HD)

The Hausdorff Distance evaluates the maximum surface discrepancy between predicted and actual boundaries. Formally, it calculates the greatest distance from any point on one surface to the nearest point on the other. Because the HD is strongly influenced by outliers, it highlights errors such as poorly segmented isolated voxels or jagged boundary peaks. Therefore, the HD provides a measure of robustness and maximum boundary error, which is crucial in clinical applications where localized boundary deviations can indicate anatomically implausible predictions.

5.3 Average Surface Distance (ASD)

The Average Surface Distance complements the HD by measuring the average distance between surfaces instead of the maximum deviation:

$$\text{ASD} = \frac{1}{|S_A|} \sum_{p \in S_A} d(p, S_B),$$

where $d(p, S_B)$ is the distance from point p on the prediction surface S_A to the reference surface S_B. ASD is less sensitive to outliers and captures the overall geometric similarity between surfaces. In prostate segmentation, ASD is useful for detecting smooth and systematic contour changes that, while they may not significantly affect DICE or HD individually, do degrade the quality of the reconstruction.

5.4 Multimetric Score

Since DICE, HD, and ASD capture different aspects of segmentation performance, a weighted multi-objective score was used [3, 4]:

$$\text{multi} = w_D \cdot \text{DICE} + w_{HD} \cdot \frac{1}{1 + \text{HD}} + w_{ASD} \cdot \frac{1}{1 + \text{ASD}}.$$

This formulation ensures that lower HD and ASD values contribute positively to the final score, keeping all terms on comparable scales. The multimetric criterion is crucial when selecting control points, as it avoids the risk of choosing a model with good performance in terms of region overlap (DICE) but with low geometric accuracy (HD or ASD). In the context of generating tokenized anatomical representations for diffusion models [4], it is especially important to prioritize segmentations with smooth and anatomically consistent boundaries.

6 Hyperparameters and Experimental Setup

This section details the hyperparameters used during training and explains the rationale for each choice. All experiments were performed using the `VQUNet3Dposv3` architecture [3] on volumetric MRI data of the prostate [1, 2], with three segmentation classes (Background, Transition Zone [TZ], and Peripheral Zone [PZ]). The selected configurations balance expressive capacity, memory feasibility on the Télécom Paris GPU cluster, and training stability. For clarity, we distinguish between the reference binary model and two training regimes of the proposed three-class model.

6.1 Data Splitting

All experiments described in this work use the balanced split generated by the script `make_balanced_split_by_volume.py` described in the section 3 (Pipeline Modifications). The final split follows a ratio of **55/12/12** for training, validation, and testing, respectively, stratified by TZ and PZ volume and grouped by patient. Unlike the original repository, which used BMC for training/validation and RUNMC only for testing [5], the new split combines both centers in all subsets. As a result, the evaluation protocol is more robust against center-specific biases, and the three-class results presented are based on a consistent and statistically balanced data split.

6.2 Training Hyperparameters

Batch size. A batch size of **1** was used in all runs. Due to the high memory consumption of the 3D convolutional layers, jump connections, and the vector quantization bottleneck, larger batch sizes would lead to out-of-memory (OOM) errors. Using a single-volume batch ensures stable training on a single GPU.

Number of epochs. For the proposed three-class `VQUNet3Dposv3` model [3], two training regimes were considered:

- a **50-epoch** run, used as the initial baseline in the three-class setup;
- a **100-epoch** run was performed, designed to further refine the vector quantized representation and segmentation performance.

The original reference binary model from the reference repository was also trained for 50 epochs with its default configuration.

Validation frequency (`check_val`). In the 50-epoch run of the proposed model, validation was performed at each **epoch** (`check_val = 1`), with the best Dice checkpoint occurring around epoch 49. Two 100-epoch runs were also performed, with validation at **5 epochs** (`check_val = 5`) and **1 epoch** (`check_val = 1`), achieving the best performance at epochs 54 and 79, respectively. The reference binary model performed validation similarly at each epoch, with a final checkpoint recorded after 50 epochs. The corresponding quantitative results are summarized Section 7.

Learning rate and optimizer. In all runs of `VQUNet3Dposv3`, the **AdamW** was used with a learning rate of `1e-4` and an L2 regularization of`1e-5`. AdamW improves stability compared to Adam by decoupling the L2 regularization from gradient updates, which is especially beneficial in architectures that include quantization losses.

Loss function. The segmentation loss is a combined loss of **Dice** + **Cross-Entropy** (`DiceCELoss` from MONAI), applied to one-hot encoded labels. This combination improves performance in unbalanced classes while maintaining stable gradients.

6.3 Model Hyperparameters

Channels. The base number of channels was set to **16**, doubling at each encoder level (up to 256). This provides sufficient representation capacity while controlling memory usage in the deeper 3D layers.

Embedding dimension. The embedding dimension was set to **256** [3]. This choice balances the expressiveness of the latent representations with the computational cost of quantization and codebook search.

Codebook size (n_{embed}). The codebook contains **1024 entries** [3]. This provides a large set of prototype vectors for 3D feature discretization without causing a memory overflow during training.

Dropout. The dropout was set to **0.0**. Empirically, the dropout technique degraded stability at the quantization bottleneck and was therefore disabled.

Number of classes. The model generates three classes: Background, Transition Zone (ZT), and Peripheral Zone (ZP).

6.4 Spatial Configuration

Input Volume Size. All volumes were resampled and cropped to a uniform spatial size of $(192, 192, 64)$. This ensures compatibility with the U-Net 3D structure and maintains anatomical coverage of the prostate.

Voxel Spacing. Images and labels were resampled to a unified voxel spacing of $(0.5mm, 0.5mm, 1.5mm)$ using MONAI's `Spacingd`. This harmonized the differences between acquisitions.

6.5 Multimetric Weightings

A weighted multi-objective score was used for model selection [3, 4]:

$$\text{multi} = w_D \cdot \text{Dice} + w_{HD} \cdot \frac{1}{1+HD} + w_{ASD} \cdot \frac{1}{1+ASD}.$$

Two different configurations of (w_D, w_{HD}, w_{ASD}) were considered:

- **50-epoch run (three-class model):** $w_D = 1.0$, $w_{HD} = 0.0$, $w_{ASD} = 0.0$. In this case, the multimetric score coincides with the Dice coefficient, and only the model with the best Dice coefficient was selected.

- **100-epoch run (three-class model):** $w_D = 8.0$, $w_{HD} = 1.0$, $w_{ASD} = 1.0$. This configuration prioritizes the Dice index and explicitly incorporates HD and ASD into the selection criteria, favoring models with good overlap and precise boundaries [4].

For the 100-epoch run, the control points for best Dice index and best Multi model coincide at epochs 54 and 79. A specific best Multi model was not calculated for the 50-epoch run with $w_{HD} = w_{ASD} = 0$.

6.6 Rationale for Hyperparameter Selection

In general, the selected hyperparameters reflect a balance between anatomical fidelity, computational feasibility, and training stability. The combination of a small batch size, conservative codebook dimensions, and a controlled validation frequency ensures that the model remains manageable while providing valuable insights into the behavior of vector-quantized 3D representations [3].

With the balanced split described above, the reference binary model, trained over 50 epochs, achieved average test metrics of approximately 0.7542 ± 0.0664 for the Dice index, 37.70 ± 15.09 mm for head height, and 6.74 ± 3.33 mm for anatomical diameter. For the three-class proposal `VQUNet3Dposv3`, running 50 epochs with validation in each epoch and selection based on pure Dice ($w_D = 1, w_{HD} = w_{ASD} = 0$) produced average test values of Dice 0.7407 ± 0.0772, HD 18.43 ± 6.13 mm and ASD 2.02 ± 0.63 mm. By extending the training to 100 epochs, validating every 5 epochs, and adopting the multimetric weighting ($w_D = 8, w_{HD} = 1, w_{ASD} = 1$), performance was further improved, achieving a Dice index of 0.7729 ± 0.0508, and HD of 16.74 ± 6.59 mm and an ASD of 1.86 ± 0.40 mm. Finally, training over 100 epochs, validating each epoch and using the same multimetric weighting ($w_D = 8, w_{HD} = 1, w_{ASD} = 1$), minimally reduced the performance of the metrics, achieving a Dice index of 0.7386 ± 0.0721, an HD of 17.92 ± 7.22 mm, and an ASD of 1.97 ± 0.66 mm. These results (see also the section 7) confirm that the chosen configurations allow the three-class model to match and even exceed the reference Dice index, while drastically reducing the HD and ASD, thus improving boundary accuracy in a more complex and clinically relevant segmentation environment.

7 Quantitative Results

This section presents the quantitative performance of the proposed model `VQUNet3Dposv3` añadir [3] for prostate segmentation into three classes (Fundus, Transition Zone, Peripheral Zone) and compares it with the original binary reference model from the reference repository [5]. All results were calculated on the balanced division described in Section 6 and are summarized in terms of mean and standard deviation for the Dice index, Hausdorff distance (HD), and average surface distance (ASD).

Three configurations of the proposed three-class model are considered:

- a **50-epoch run** with validation at each epoch (`check_val = 1`) and model selection based solely on the Dice index ($w_D = 1, w_{HD} = w_{ASD} = 0$); the best checkpoint is found at epoch 49;
- a **100-epoch run** with validation every 5 epochs (`check_val = 5`) and a multimetric selection criterion ($w_D = 8, w_{HD} = 1, w_{ASD} = 1$); the best Dice and Multimetric checkpoints coincide at epoch 54;
- a **100-epoch run** with validation every epoch (`check_val = 1`) and a multimetric selection criterion ($w_D = 8, w_{HD} = 1, w_{ASD} = 1$); the best Dice and Multimetric checkpoints coincide at epoch 79;
- the original **binary baseline** model (prostate vs. background), trained for 50 epochs under the original repository configuration [5].

7.1 Summary of Mean and Standard Deviation Metrics

Table 1 reports the mean $\pm$ standard deviation of Dice, HD and ASD for the binary baseline and for the proposed three-class model under the two training regimes. For the 100-epoch run, the best-Dice and best-Multi checkpoints yield identical test statistics, since they correspond to the same epoch with the chosen weights.

7.2 Interpretation

As expected, the binary baseline model achieves a reasonably high Dice score (≈ 0.75), but at the cost of very large boundary errors, with HD ≈ 37.7 mm and ASD ≈ 6.7 mm. This reflects the fact that the baseline solves a simpler two-class problem (prostate vs. background), without explicitly separating TZ and PZ.

Setting	Dice	HD	ASD
Baseline (binary, 50 epochs)	0.7542 ± 0.0664	37.6952 ± 15.0929	6.7354 ± 3.3346
Best Dice (3-class, 50-5 epochs-val, epoch 49)	0.7407 ± 0.0772	18.4319 ± 6.1275	2.0209 ± 0.6272
Best Dice (3-class, 100-5 epochs-val, epoch 54)	0.7729 ± 0.0508	16.7430 ± 6.5931	1.8607 ± 0.4019
Best Multi (3-class, 100-5 epochs-val, epoch 54)	0.7729 ± 0.0508	16.7430 ± 6.5931	1.8607 ± 0.4019
Best Dice (3-class, 100-1 epochs-val, epoch 79)	0.7386 ± 0.0721	17.9175 ± 7.2166	1.9745 ± 0.6564
Best Multi (3-class, 100-1 epochs-val, epoch 79)	0.7386 ± 0.0721	17.9175 ± 7.2166	1.9745 ± 0.6564

Table 1: Mean $\pm$ standard deviation of Dice, Hausdorff Distance (HD) and Average Surface Distance (ASD) for the baseline binary model and the proposed three-class `VQUNet3Dposv3` under two training regimes. All results are computed on the balanced test split.

In the three-class setting, the 50-epoch run of `VQUNet3Dposv3` already produces competitive Dice values (≈ 0.74) while dramatically reducing both HD and ASD compared to the baseline. Extending training to 100 epochs and using the multi-metric selection criterion further improves all metrics: the best-Dice/best-Multi checkpoint at epoch 54 attains a higher mean Dice (≈ 0.77) and substantially lower HD and ASD (around 16.7 mm and 1.86 mm, respectively). Additionally, with the simulation of 100 epochs with 1 validation epoch per epoch, despite a slight drop in metrics, it still exhibits competitive performance.

Overall, these results indicate that the proposed three-class VQ-based model [3]is able to match and even surpass the baseline Dice score, while significantly improving boundary accuracy in a more challenging and clinically relevant segmentation scenario.

8 Training Dynamics and Learning Curves

This section analyses the evolution of the training loss and validation metrics over the 100-epoch run of the three-class `VQUNet3Dposv3` model [3] with validation every epoch and multi-metric weights $w_D = 8, w_{HD} = 1, w_{ASD} = 1$. The corresponding learning curves are shown in Figures 1a–1b.

8.1 Training and Validation Loss

Figure 1a displays the evolution of the training and validation loss over 100 epochs [1, 2]. Both curves show a rapid decrease during the first 10–20 epochs, followed by a slower but steady decay afterwards. The training loss starts above 1.5 and gradually converges towards values around 0.4, while the validation loss decreases from approximately 0.75 to about 0.25. Although small oscillations are visible, especially in the later stages of training, there is no indication of divergence between the two curves; the validation loss remains consistently lower than the training loss. This behaviour suggests that the optimisation is stable and that there is no clear evidence of overfitting in the current regime.

8.2 Validation Dice and Multi-score

The validation Dice curve in Figure 1c exhibits a pronounced improvement during the initial epochs. Starting from values close to zero, the Dice score quickly rises above 0.3 after only a few epochs and surpasses 0.6 within the first ten. Between epochs 20 and 30 the curve reaches the range 0.68–0.72 and then enters a plateau with small fluctuations around 0.7–0.73 for the remainder of training. This indicates that most of the volumetric overlap is learned early, and subsequent epochs mainly refine the details of the segmentation.

The validation multi-score (Figure 1d), which combines Dice, HD and ASD, follows a qualitatively similar pattern but with a smoother increase. It starts near zero, rises steeply during the first 5–10 epochs to values around 0.4–0.5, and then continues to grow more gradually, stabilising close to 0.6 towards the end of training. The fact that the multi-score continues to improve slightly even after the Dice has plateaued suggests that later epochs still contribute to better boundary alignment (lower HD and ASD), even if the global overlap changes only marginally.

8.3 Validation HD and ASD

The Hausdorff Distance curve in Figure 1e shows a noisy but overall decreasing trend. HD starts at very high values (around 70–90 mm) in the first epochs and then progressively drops towards 40–50 mm. Although there

are several local spikes throughout training, the envelope of the curve clearly moves downward, and in the final third of the run HD mostly fluctuates within a narrower band around 40–45 mm. This behaviour reflects the strong sensitivity of HD to isolated outliers, while still indicating a net improvement in the worst-case boundary errors.

The ASD curve in Figure 1f displays an even more pronounced reduction. After an initial peak above 20 mm, ASD rapidly falls below 10 mm within the first 10 epochs, and continues to decrease towards values around 4–6 mm. Occasional spikes are visible, but they are short-lived, and the overall level remains low for most of the training. This confirms that, on average, the predicted surfaces approach the ground-truth boundaries quite closely once the model has passed the initial learning phase.

8.4 Joint View of All Metrics

Figure 1b summarises the simultaneous evolution of Dice, multi-score, HD and ASD on a single plot. Three main observations can be made:

- the sharp increase in Dice during the first 10–20 epochs is mirrored by a strong improvement in the multi-score;
- the largest reductions in HD and ASD occur in the same early epoch range, coinciding with the period where Dice and multi-score rise most rapidly;
- after roughly 30 epochs, all metrics enter a regime of slower change, with Dice and multi-score fluctuating around their asymptotic values and HD/ASD oscillating within a relatively narrow band without systematic degradation.

Taken together, these learning curves indicate that the model converges to a stable solution: losses decrease, Dice and multi-score improve and then remain high, and the boundary-based metrics do not deteriorate over time. The multi-score behaves as intended, providing a smooth summary of overlap and boundary quality that reflects the joint evolution of Dice, HD and ASD throughout training.

9 Qualitative Results

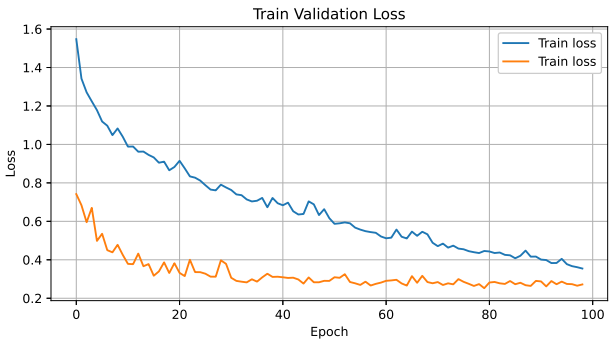
In addition to the global metrics presented in the section 7, a qualitative inspection of several 3D test cases was performed, converting both the reference labels and the predicted segmentations into surface meshes [6]. In all representations, the transition zone (ZT) is shown in beige and the peripheral zone (ZP) in green. For consistency, all the examples illustrated in this section come from the configuration of **100 epochs, validation at every epoch**, whose extreme values per case (best, average, worst) are summarized in the table 2, along with the corresponding values for the other configurations (50 epochs, validation at every epoch and 100 epochs, validation every 5 epochs).

	Dice ↑			HD [mm] ↓			ASD [mm] ↓		
Configuration	Best	Medium	Worst	Best	Medium	Worst	Best	Medium	Worst
50 epochs, val=1	0.81	0.77	0.56	8.16	18.05	29.42	1.44	1.72	3.09
100 epochs, val=5	0.84	0.80	0.67	7.27	14.79	31.53	1.35	1.72	2.55
100 epochs, val=1	0.82	0.75	0.56	10.15	14.69	30.98	1.34	1.84	3.21

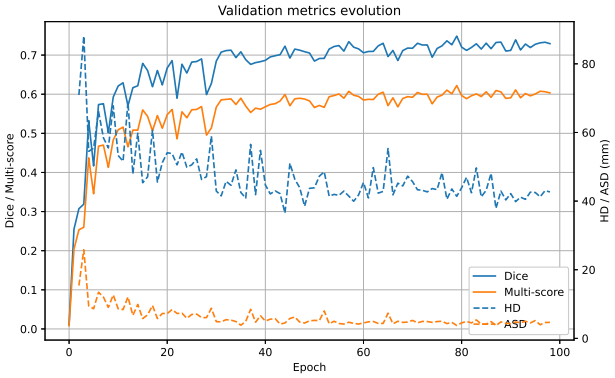
Table 2: Best, medium and worst per–case values of Dice, Hausdorff Distance (HD) and Average Surface Distance (ASD) for each training configuration.

BSince Dice, HD, and ASD capture different aspects of performance, a single test volume can be optimal for one metric and suboptimal for another. Within the 100 epochs / val=1 setting, we selected seven representative subjects [1, 2]:

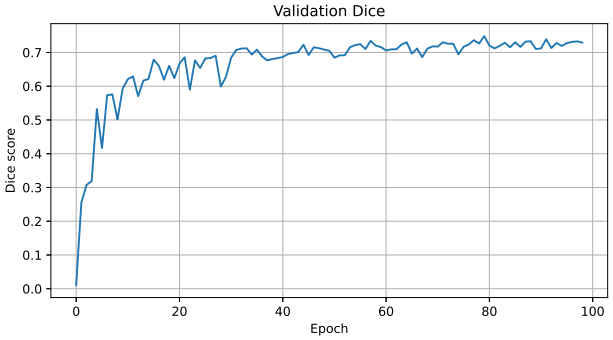
- Test case 2 with **best Dice**, which combines a very high overlap (Dice ≈ 0.82);
- Test case 6 with **best HD**, achieving the lowest boundary errors (HD ≈ 10.15 mm);
- Test case 11 with **best ASD and average HD**, achieving low average boundary errors but close to the median HD of the test distribution (ASD ≈ 1.34 mm, HD ≈ 14.69 mm);
- Test case 10 with **medium Dice and ASD**, with average limit accuracy and acceptable volumetric overlap (Dice ≈ 0.75, ASD ≈ 1.84 mm);



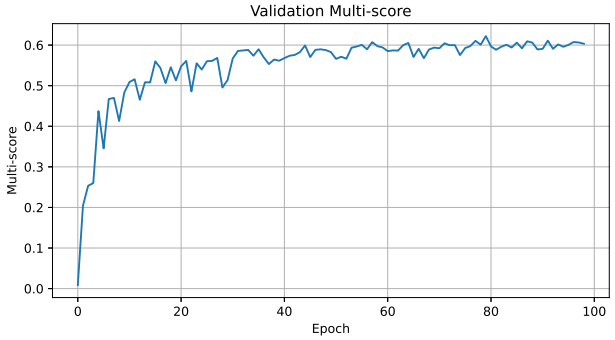
(a) Training loss



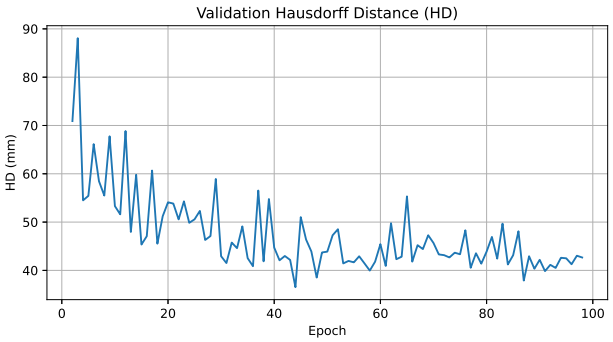
(b) Joint evolution of all metrics



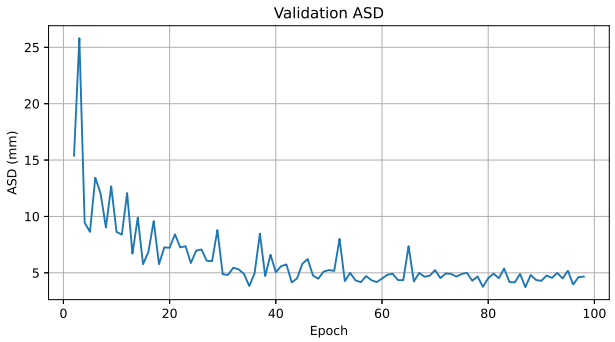
(c) Validation Dice



(d) Validation multi-score



(e) Validation HD



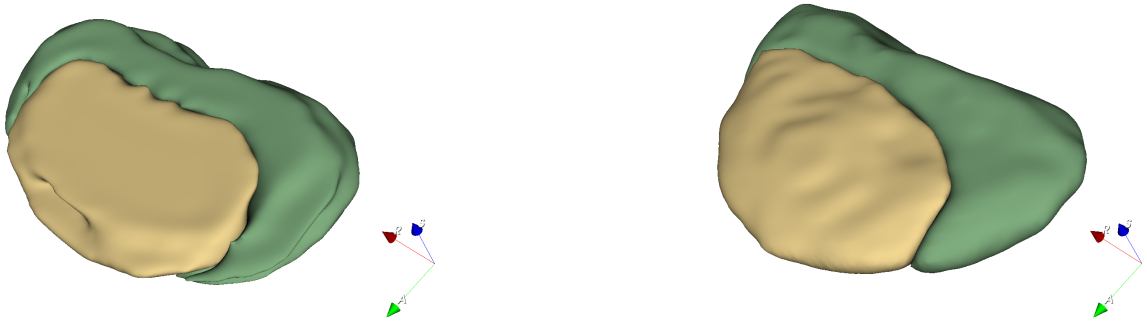
(f) Validation ASD

Figure 1: Training and validation curves for the three-class `VQUNet3Dposv3` model over 100 epochs with validation every epoch.

- Test case 7 with very **worst Dice**, illustrating a clear failure (Dice ≈ 0.56).
- Test case 1 with very **worst HD**, where the maximum surface distance is large (HD ≈ 30.98 mm);
- Test case 4 with very **worst ASD**, again illustrating a clear failure (ASD ≈ 3.21 mm).

9.1 Best Dice (test case 02)

Figures 2a and 2b show the ground-truth and predicted meshes for test case 02 [1, 2], which attains the best Dice score in the 100 epochs / val=1 configuration (Dice ≈ 0.82). The predicted TZ and PZ almost perfectly overlap with the reference volumes: the global shape of the gland is well recovered, and the interface between zones follows the anatomical boundary with only very small deviations. The external contour is smooth and anatomically plausible, without spurious components. Although HD and ASD are not the very best in the test set, they remain low, confirming that the excellent volumetric overlap is accompanied by good boundary accuracy.

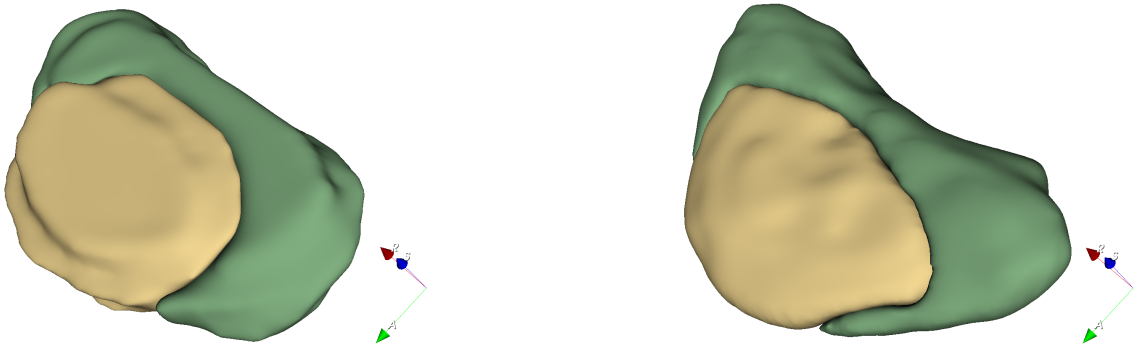


(a) Ground-truth TZ/PZ (test case 02, best Dice). (b) Prediction (test case 02).

Figure 2: Best Dice example: excellent volumetric overlap.

9.2 Best HD (test case 06)

Figures 3a and 3b correspond to test case 06 [1, 2], which exhibits the lowest Hausdorff Distance (HD $\approx$ 10.15 mm). In this subject, the prediction closely follows the ground-truth surface over the whole prostate. Both TZ and PZ are correctly located, and the PZ cap is accurately reproduced without large local protrusions or missing regions. The very small maximal discrepancy between surfaces explains the excellent HD, while Dice and ASD are also high-quality, though not necessarily extremal, showing that this case is close to an "ideal" segmentation from a geometric standpoint.



(a) Ground-truth TZ/PZ (test case 06, best HD). (b) Prediction (test case 06).

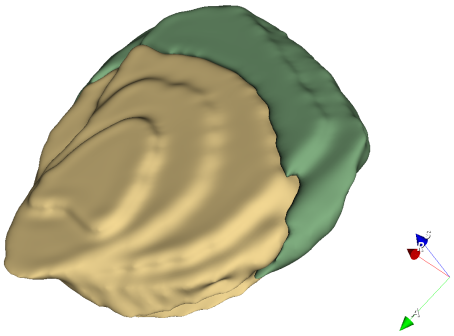
Figure 3: Best HD example. TZ is shown in beige and PZ in green.

9.3 Best ASD and medium HD (test case 11)

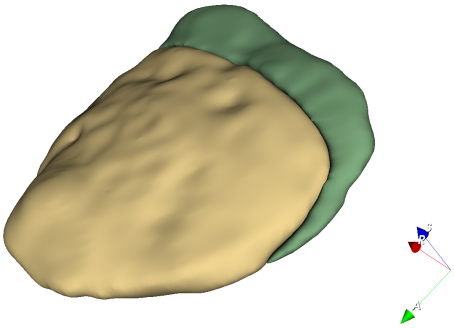
The meshes in Figures 4a and 4b illustrate test case 11 [1, 2], which achieves the best Average Surface Distance (ASD $\approx$ 1.34 mm) but only a medium HD. The prediction and ground truth are extremely close over most of the surface: the TZ and PZ volumes largely overlap, and the boundary between zones is smooth and well aligned. However, a few localised regions show slightly larger deviations that affect the maximum distance and thus increase HD to a value close to the median of the distribution. This case clearly demonstrates the complementarity between ASD and HD: ASD captures the fact that most of the surface is very well matched, while HD remains sensitive to isolated outliers.

9.4 Medium Dice and medium ASD (test case 10)

Figures 5a and 5b present test case 10 [1, 2], a representative example with Dice and ASD close to the median values (Dice $\approx$ 0.75, ASD $\approx$ 1.84 mm). The global morphology and orientation of the prostate are correctly recovered: TZ and PZ have similar volumes and relative positions as in the ground truth. Nonetheless, the prediction shows mild shape differences, such as slightly flattened or thickened regions and small systematic offsets along parts of the boundary. These discrepancies are not large enough to drastically reduce Dice, but they are reflected in the intermediate ASD, illustrating what a "typical" segmentation looks like for this configuration.

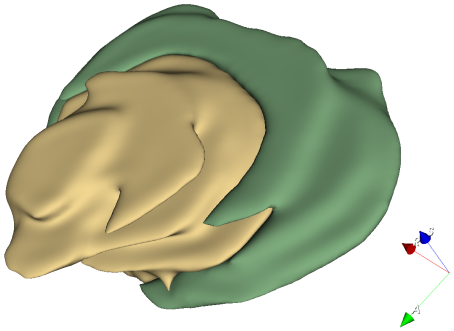


(a) Ground-truth TZ/PZ (test case 11, best ASD / medium HD).

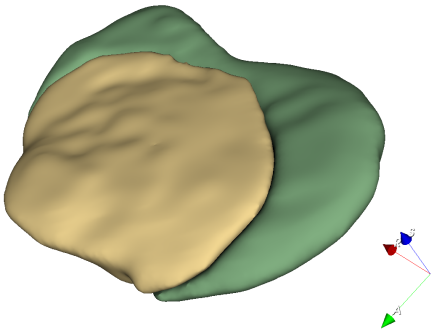


(b) Prediction (test case 11).

Figure 4: Best ASD / medium HD example. good global shape, but local boundary irregularities.



(a) Ground-truth TZ/PZ (test case 10, medium Dice / medium ASD).

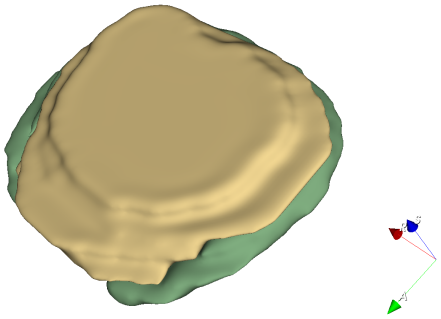


(b) Prediction (test case 10).

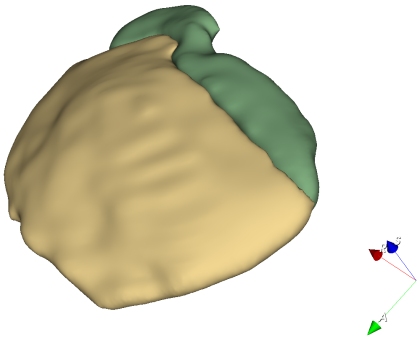
Figure 5: Medium Dice and ASD: reasonable global alignment with local protrusions and maximum surface distance.

9.5 Worst Dice (test case 07)

The example in Figures 6a and 6b corresponds to test case 07 [1, 2], which attains the lowest Dice score in the test set (Dice ≈ 0.56). Here the predicted TZ and PZ depart noticeably from the ground truth: the overall gland is deformed, with an over-extended TZ and PZ regions that are partially missing or misplaced. Although some parts of the surface still coincide with the reference, large portions of the volume are either under- or over-segmented. This leads to a substantial loss of overlap while HD and ASD also deteriorate, illustrating a genuine failure case in terms of volumetric agreement.



(a) Ground-truth TZ/PZ (test case 07, worst Dice).

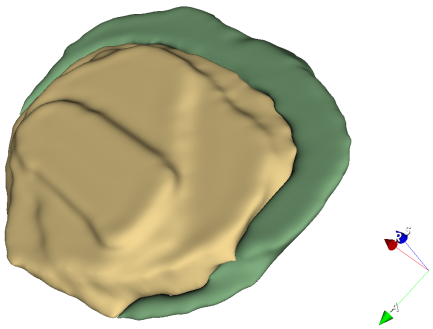


(b) Prediction (test case 07).

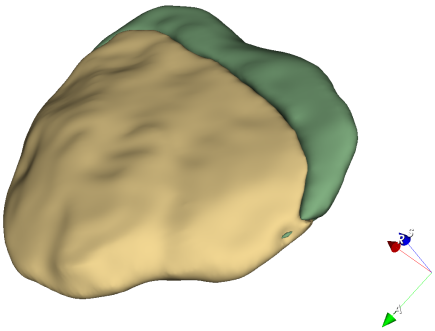
Figure 6: Worst Dice example: clear mismatch in overall shape, illustrating a failure case of the model.

9.6 Worst HD (test case 01)

Figures 7a and 7b show test case 01 [1, 2], which presents the worst Hausdorff Distance (HD ≈ 30.98 mm). In this subject, most of the prostate is reasonably well aligned with the ground truth, and the Dice score remains in an acceptable range. However, the prediction contains pronounced local protrusions and elongated structures, particularly near the apex, that are not present in the reference mesh. These outlying regions contribute little to the global overlap but induce very large pointwise distances, dramatically increasing HD. This example highlights the importance of HD for detecting isolated but potentially clinically relevant segmentation errors [7].



(a) Ground-truth TZ/PZ (test case 01, worst HD).

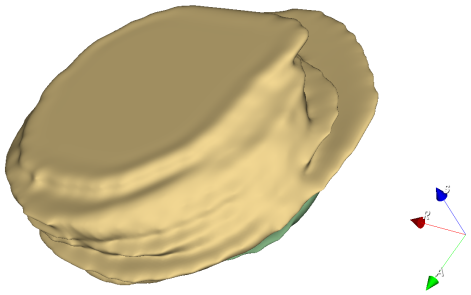


(b) Prediction (test case 01).

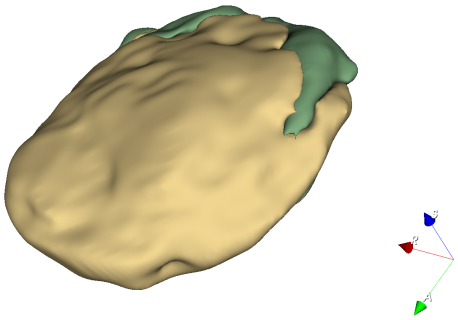
Figure 7: Worst HD: outliers that strongly affect the maximum surface distance.

9.7 Worst ASD (test case 04)

Finally, Figures 8a and 8b depict test case 04 [1, 2], which exhibits the highest ASD (ASD ≈ 3.21 mm). Unlike the worst-HD example, the errors here are not restricted to a few extreme outliers. Instead, there is a systematic offset of the predicted surface over a relatively large portion of the gland: the TZ appears globally shifted and thickened, and the PZ band does not follow the reference contour closely. As a consequence, the average distance between surfaces is high, even though the maximal distance (HD) is not the worst inthe set. This case illustrates how ASD complements both Dice and HD by penalising widespread but moderate boundary misalignments.



(a) Ground-truth TZ/PZ (test case 04, worst ASD).



(b) Prediction (test case 04).

Figure 8: Worst ASD example: clear mismatch zone delineation, illustrating a failure case of the model.

Overall, the qualitative analysis for the 100 epochs / val=1 configuration confirms the quantitative findings: most test cases exhibit anatomically plausible segmentations with good overlap and reasonable boundary accuracy, while a small number of outliers show clear artefacts that are effectively captured by HD and ASD [7]. The selected examples also emphasise that Dice, HD and ASD provide complementary information and should be considered jointly when assessing model performance and selecting checkpoints for downstream clinical applications.

10 Technical Challenges

The work carried out so far required several technical adjustments to make the original repository usable on the prostate MRI dataset and on the available hardware. This section summarises the main difficulties encountered and the solutions implemented.

GPU memory limitations (OOM). The original architecture, which combines a deep 3D U-Net, a vector–quantisation bottleneck and, in some versions, an internal Transformer module, led to frequent out-of-memory (OOM) errors on the GPUs available at Télécom Paris, even with small batch sizes. In practice, it was not possible to train the model with the original configuration (large codebook, high embedding dimension, Transformer enabled and multi-GPU). To address this, the network was reconfigured to use a single GPU, batch size 1, a reduced codebook ($n_{\text{embed}} = 1024$) and an embedding dimension of 256, and the internal Transformer block was removed. These changes significantly reduced the memory footprint while preserving the ability to learn meaningful latent representations.

Dataset inconsistencies and reorganisation. Adapting the original `ProstateDataset` to the TCIA PROSTATE DIAGNOSIS / Prostate-3T data [1, 2] required several corrections. Some volumes presented missing labels or absent classes (no PZ), and the original split used in the repository placed BMC cases in train/validation and RUNMC cases in test, introducing a strong centre bias. A dedicated script (`make_balanced_split_by_volume.py`) was developed to regenerate the CSV files with a 55/12/12 train/validation/test split, stratified by TZ/PZ volume and grouped by patient.

Metric computation and validation logic. During the first experiments, the reported Dice, HD and ASD values did not match visual impressions of the segmentations. This discrepancy was traced back to issues in the validation loop: metrics were aggregated incorrectly across batches and volumes, and the multi-metric score used for checkpoint selection was not consistent with the final evaluation. The `validation_epoch_end` method was therefore rewritten to: (i) accumulate per-volume metrics using an explicit counter (`val_number`), (ii) compute true averages over the validation set, and (iii) define a well-documented multi-score based on Dice, HD and ASD.

Adapting and extending the original repository. Porting the original Vector-Quantisation-for-Robust-Segmentation code [5] to the present setting required substantial refactoring. The data loading pipeline had to be adapted to a different folder structure and label convention; the network class (`VQUNet3Dposv3`) was integrated into a PyTorch Lightning module with additional logging; and new callbacks were added for saving checkpoints with epoch information and exporting EPS figures of the learning curves. Several environment-related issues (library versions, CUDA compatibility, MONAI transforms) also had to be resolved before reproducible training could be achieved.

11 Current Challenges and Future Work

The results obtained so far show that the adapted `VQUNet3Dposv3` model can produce competitive three-class prostate segmentations and significantly improve boundary metrics with respect to the original baseline [5]. Nevertheless, several technical and methodological challenges remain open and will guide the next steps of this project.

Improving codebook stability. Although the current configuration yields a functioning vector-quantised representation, some signs of codebook under–utilisation and sensitivity to hyperparameters are still observed. A more systematic exploration of the commitment cost, learning rate schedules and codebook update strategies [3] is required to obtain a more stable and expressive latent space, which is particularly important for the downstream discrete diffusion stage [4].

Tuning codebook size and architecture capacity. The current codebook size ($n_{\text{embed}} = 1024$) and embedding dimension (256) were chosen as a compromise between expressiveness and GPU constraints. Future work will explore smaller and larger codebooks, as well as alternative bottleneck designs [3], to better understand their impact on segmentation accuracy, codebook usage and, ultimately, the quality of the discrete shape representations used by the diffusion model [4].

Re-evaluating the Transformer module. The internal Transformer present in the original repository [5] had to be removed for memory reasons. If additional GPU resources become available, it would be interesting to reintroduce a lightweight attention module, or experiment with more parameter–efficient variants, to assess whether modelling long-range dependencies in the latent space can further improve anatomical coherence.

12 Conclusions of the First Report

This first report has focused on adapting and extending the original Vector-Quantisation-for-Robust-Segmentation repository [5] to the problem of three-class prostate MRI segmentation and on establishing a robust experimental framework for subsequent work on discrete diffusion models [4].

The dataset was reorganised using a balanced 55/12/12 train/validation/test split, stratified by TZ/PZ volume and grouped by patient, thereby removing centre-specific biases present in the original BMC/RUNMC split [1, 2]. The data preprocessing, loss functions, metrics and logging pipeline were standardised, and the validation logic was corrected so that Dice, HD, ASD and the multi-metric score are computed consistently and can be reliably used for model selection.

Quantitatively, the proposed three-class model matches and even surpasses the Dice of the original binary baseline [5] while substantially improving boundary metrics. The learning curves show stable convergence without obvious overfitting, and the qualitative analysis confirms that most predicted meshes are anatomically plausible, with only a small number of clear failure cases.

At the same time, several limitations remain. GPU memory constraints imposed restrictions on batch size, codebook dimension and the use of attention modules; some signs of codebook under-utilisation are still present; and a few test subjects exhibit significant local artefacts that degrade HD and ASD. These issues highlight the need for richer regularisation, improved codebook training strategies and more diverse data.

References

[1] B Nicolas Bloch, Ashali Jain, and C. Carl Jaffe. *Data From PROSTATE-DIAGNOSIS*. 2015. DOI: 10.7937/K9/TCIA.2015.FOQEUJVT. URL: https://www.cancerimagingarchive.net/collection/prostate-diagnosis/.

[2] Geert Litjens, Jurgen Futterer, and Henkjan Huisman. *Data From Prostate-3T*. 2015. DOI: 10.7937/K9/TCIA.2015.QJTV5IL5. URL: https://www.cancerimagingarchive.net/collection/prostate-3t/.

[3] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. "Neural Discrete Representation Learning". In: *CoRR* abs/1711.00937 (2017). arXiv: 1711.00937. URL: http://arxiv.org/abs/1711.00937.

[4] Minghui Hu et al. "Global Context with Discrete Diffusion in Vector Quantised Modelling for Image Generation". In: *CoRR* abs/2112.01799 (2021). arXiv: 2112.01799. URL: https://arxiv.org/abs/2112.01799.

[5] AinkaranSanthi. *Vector-Quantisation-for-Robust-Segmentation*. https://github.com/AinkaranSanthi/Vector-Quantisation-for-Robust-Segmentation. June 2022. URL: https://github.com/AinkaranSanthi/Vector-Quantisation-for-Robust-Segmentation.

[6] *Slicer Wiki*. [Online; accessed 17-November-2025]. 2019. URL: %5Curl%7Bhttps://www.slicer.org/w/index.php?title=Main_Page&oldid=62645%7D.

[7] B. Nicholas Bloch et al. *NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures (ISBI-MR-Prostate-2013)*. 2015. DOI: 10.7937/K9/TCIA.2015.ZF0VLOPV. URL: https://www.cancerimagingarchive.net/analysis-result/isbi-mr-prostate-2013/.