Diffusion-Based Quality Control of Medical Image Segmentations across Organs

Vincenzo Marcianò, Hava Chaptoukaev, Virginia Fernandez, M. Jorge Cardoso,
Sébastien Ourselin, Michela Antonelli, Maria A. Zuluaga

***Abstract*—Medical image segmentation using deep learning (DL) has enabled the development of automated analysis pipelines for large-scale population studies. However, state-of-the-art DL methods are prone to hallucinations, which can result in anatomically implausible segmentations. With manual correction impractical at scale, automated quality control (QC) techniques have to address the challenge. While promising, existing QC methods are organ-specific, limiting their generalizability and usability beyond their original intended task. To overcome this limitation, we propose *no-new Quality Control (nnQC)*, a robust QC framework based on a diffusion-generative paradigm that self-adapts to any input organ dataset. Central to nnQC is a novel *Team of Experts (ToE)* architecture, where two specialized *experts* independently encode 3D spatial awareness, represented by the relative spatial position of an axial slice, and anatomical information derived from visual features from the original image. A weighted conditional module dynamically combines the pair of independent embeddings, or *opinions* to condition the sampling mechanism within a diffusion process, enabling the generation of a spatially aware pseudo-ground truth for predicting QC scores. Within its framework, nnQC integrates *fingerprint* adaptation to ensure adaptability across organs, datasets, and imaging modalities. We evaluated nnQC on seven organs using fifteen publicly available datasets. Our results demonstrate that nnQC consistently outperforms state-of-the-art methods across all experiments, including cases where segmentation masks are highly degraded or completely missing, confirming its versatility and effectiveness across different organs.**

***Index Terms*—Quality Control, Generative Modeling, Self-adapting Framework, Medical Image Segmentation.**

I. Introduction

ADVANCES in deep learning (DL) have demonstrated unprecedented capabilities in automating and expediting medical image segmentation [1]. Despite their high accuracy, DL techniques can still predict anatomically implausible segmentations [2]. As a result, their translation to real-world clinical applications requires visual quality control (QC). Visual QC process involves inspecting each segmented image for spurious results, followed by manual correction or discarding, which is unfeasible at scale [3], [4].

Automated QC techniques have emerged as a mechanism to bypass visual QC of predicted segmentations [5], [6]. These methods involve the definition of a normative model of high-quality segmentations, which is then used to infer a qualitative [5] or quantitative [4], [6]–[16] score reflecting the quality of a given predicted image segmentation. However, while medical image segmentation frameworks are increasingly general [17] or easy to adapt and apply [18] across diverse modalities and organs, automated QC methods have not followed the same generalization trend. Current automatic QC methods are limited by their metric-specific [5], [10], [12] or organ-specific design [9], [13], [15], restricting their use across anatomical structures and imaging modalities and making their seamless use across applications difficult [1]. As a result, the efficiency and scalability achieved with general-purpose segmentation models are often undermined by the need to design and adapt dedicated QC pipelines for each new organ or application. Enabling large-scale population studies requires robust and self-adapting QC frameworks capable of jointly working with state-of-the-art segmentation methods [18] and able to assess segmentations of varying quality and degradation from different organs.

In this work, we introduce *no-new Quality Control* (nnQC), a self-adapting, metric- and model-agnostic framework designed for robust QC of medical image segmentations. nnQC is built upon a Latent Diffusion Model (LDM) backbone and is designed to handle diverse segmentation qualities while remaining adaptable across a wide range of organs, datasets, and imaging modalities. nnQC follows a state-of-the-art 2D reconstruction-based QC approach that generates a pseudo-ground-truth (pGT) mask associated with the predicted segmentation, enabling the estimation of any segmentation quality scores. It introduces a novel sampling strategy, denoted the Team of Experts (ToE), designed to inject 3D contextual information into the pGT reconstruction process. This strategy dynamically balances two independent sources of anatomical insight - referred to as *opinions* - obtained from two separate *experts* that encode anatomical information from the input image and spatial location derived from the segmentation mask. These *opinions* are fused into a conditional vector that guides the LDM's sampling process, enabling anatomically informed generation of the pGT.

Furthermore, inspired by the nnU-Net framework [18], nnQC incorporates the extraction of dataset-specific attributes, or *fingerprints*, enabling seamless self-adaptation across different organs, datasets, and imaging modalities. As such, the name *nnQC* (no-new QC) reflects the fact that the framework does not require designing a new QC model when adapting to different organs, modalities, or datasets. We perform an extensive validation of nnQC across 15 diverse scenarios, covering 15 datasets, four imaging modalities or techniques, and seven different organs, demonstrating its generalizability and robustness. To promote reproducibility and encourage broader adoption, we publicly release our open-source code and pre-trained model weights at github.com/robustml-eurecom/nnQC.

TABLE I
STRUCTURAL COMPARISON BETWEEN NNQC AND RECENT MANIFOLD-BASED QC METHODS

Model	Stochastic	Image Context	Conditioning	3D Context	Self-Configuring	Metric-Agnostic
Liu et al. [16]	✓	×	×	✓	×	×
Jin et al. [19]	✓	×	×	✓	×	×
Galati et al. [13]	×	×	×	×	×	✓
Wang et al. [15]	✓	✓	✓	×	×	✓
nnQC (Ours)	✓	✓	✓	✓	✓	✓

II. RELATED WORKS

A. Automatic QC

Automatic QC methods can be categorized into three main classes: embedded, semi-detached, and independent. Embedded QC methods are integrated within the segmentation model itself, allowing the model to self-evaluate its predicted output [4], [7], [8], [20], [21]. Semi-detached methods work separately but are specifically tailored for a particular family of segmentation approaches [9], [22], [23]. In contrast, independent QC methods are fully detached from any segmentation model, which makes them versatile and applicable across various segmentation frameworks [6], [10]–[16], [19], [24]. We focus on independent QC approaches due to their flexibility and adaptability, as they can be used without being tied to a specific model.

Among detached QC, metric-specific approaches typically focus on either classifying segmentation masks using qualitative scores (e.g., good/bad) [5], or regressing quantitative scores, such as the Dice Score [8], [10], [12], [16], [21], [24]. However, these methods are limited by the difficulty of gathering a sufficiently representative set of annotations covering the full spectrum of varying segmentation qualities [8], [10], and their inability to handle unbounded metrics (e.g., the Hausdorff Distance) [12]. Aiming to address the reliance on large annotated datasets, a subset of metric-specific approaches [16], [19] exploits the observation that high-quality segmentations inherently share common shape properties, and that these properties can be captured within a learned latent space. To this end, they employ a variational autoencoder (VAE) to learn the manifold of high-quality segmentation shapes, and a downstream regressor then operates on the learned embeddings to directly predict a quality score. While this design leverages the representational power of generative reconstruction to encode segmentation shape priors, it remains inherently metric-specific: the regressor is trained to predict a specific metric. Therefore, extending it to another measure requires retraining of the regression head from scratch.

Reconstruction-based QC techniques [6], [13], [15] are detached methods that circumvent the limitations of metric-specific approaches. These techniques generate a pseudo-ground-truth (pGT) mask associated with a given image and its corresponding predicted segmentation, enabling the estimation of quality scores for the predicted segmentation. Early reconstruction-based QC approaches [3], [6], relied on atlas propagation strategies. These registration-based methods assess segmentation quality by measuring the spatial overlap between the predicted mask and a set of reference atlas images. The underlying assumption is that a high-quality prediction will align well with at least one of the atlas images. However, the strategy depends on accurate image registration, which can be computationally expensive [13] and is prone to failure. Moreover, it requires access to annotated ground truth data at inference time.

More recent reconstruction-based-QC approaches also assume that high-quality segmentations lie in a common space. However, instead of assuming spatial alignment [3], [6] (i.e., Euclidean space), similarly to earlier metric-specific techniques [16], these methods [13], [15] build on the assumption that high-quality ground truth masks lie within a learnable latent manifold. While more efficient and robust, these methods suffer from two critical limitations. First, because they rely on distance-based retrieval to find the closest sample point to the predicted segmentation in the learned space, issues can arise when the segmentation to be controlled is very poor and is far from the underlying normative distribution of high-quality segmentations. In such cases, this distance-based matching may fail, resulting in pseudo-ground truths (pGTs) that no longer resemble the actual ground truth, ultimately leading to unreliable quality estimates. The latter problem may be exacerbated by the fact that state-of-the-art learning-based QC techniques operate in 2D [12], [13], [15], [19], [24]. Previous studies [12] have shown that performing QC at the slice level yields better results and provides finer granularity. However, the loss of three-dimensional information, which carries relevant geometric properties of a segmentation mask, can be detrimental to the sampling process. For example, segmented 2D masks of the heart's left ventricle should appear larger in the basal slices compared to the axial slices.

In this work, we leverage the advantages of 2D learning-based reconstruction-based QC techniques while addressing their limitations. We propose a novel sampling strategy that learns to generate high-quality pGTs from a diffusion-based generative model, guided by a rich embedding of visual and spatial cues. By injecting 3D contextual information, the proposed distance-based retrieval with conditional sampling is better suited to recover pGTs from poor segmentations. At the same time, it provides a scalable, model-agnostic QC solution. Table I summarizes the main characteristics of our proposed approach and compares them to state-of-the-art QC methods that build upon the principle that high-quality ground truth masks lie within a learnable latent manifold.

B. Image Synthesis

Image synthesis, powered by generative modeling, is a powerful tool in medical imaging that is used in numerous applications [25]–[29]. While earlier approaches primarily

relied on Variational Autoencoders (VAEs) [30] and Generative Adversarial Networks (GANs) [31], recent trends favor diffusion models (DMs) due to their superior training stability (better than GANs) and high-fidelity sample quality [32]–[34] (better than VAEs). However, the high computational demands of DMs, operating in the image space, limit their scalability in medical imaging applications. Latent Diffusion Models (LDMs) [34] overcome this by performing the diffusion process in a learned latent space, typically using a spatially-aware VAE or VAE-GAN. This approach enables LDMs to sample more effectively than traditional VAEs while preserving essential structural information in a compact space. This is particularly useful for reconstruction-based QC, where severely corrupted masks may deviate from plausible segmentations. LDMs sampling process helps guide the output towards realistic, high-quality reconstructions, avoiding the risk of producing overly smoothed or implausible results [6], [34].

Only a few previous works have explored the usage of DMs for segmentation mask generation. Fernández et al. [28] use a VAE-GAN–based mask generator to condition an LDM for image synthesis. Gupta et al. [29] propose a DM to generate topologically accurate masks for subsequent image generation. In both scenarios, the generated masks are an intermediate step towards the final goal of image synthesis.

In this work, we build on the LDM framework for image synthesis proposed by [28] and we extend it and adapt it to address a slightly different setup. In our case, we aim at generating segmentation masks (i.e. pseudo ground truths) guided by an input segmentation mask, whose quality is to be assessed, and the original input image.

C. Generalist and Specialist Frameworks

Recent advances in medical image segmentation have led to the emergence of generalist models capable of segmenting a wide range of anatomical structures from different protocols and imaging modalities with minimal manual intervention (e.g., prompts, scribbles, or bounding boxes) [17], [35], [36].

Alongside them, specialist models, most notably nnU-Net [18], remain highly competitive, often surpassing generalist models. nnU-Net exemplifies a "one-for-all" model paradigm, where its architecture is self-configuring and retrained from scratch for each new dataset. Despite requiring full retraining, its strong performance, automation of preprocessing and hyperparameter tuning, and ease of use have made it a de facto standard in the field [37]. Both generalist and specialist approaches have enabled fast deployment at a large scale of medical image segmentation.

In this work, we take inspiration from nnU-Net's self-adaptation strategy, integrating dataset-specific fingerprints and, thus, removing the need for manual tuning. In this way, we address the bottleneck that QC represents at the moment to medical image segmentation pipeliness, by offering a robust, scalable solution to QC that can be easily adapted across organs, datasets, and imaging techniques.

III. Method

Given an image $I \in \mathbb{R}^{H\times W}$, with H its height and W its width, and its associated segmentation *S* generated by an arbitrary segmentation model, we aim to perform segmentation QC by generating a pseudo-ground truth segmentation, pGT_I^S that approximates the real but unknown ground truth segmentation, GT_I. The pGT then serves as a reference for computing the quality score of *S* using an arbitrary quality metric $M(S, pGT_I^S)$, such that $M(S, pGT_I^S) \simeq M(S, GT_I)$.

We address the QC problem by learning to sample from a learned manifold of GT segmentations (Sec. III-A). To generate pGT_I^S, we rely on a latent diffusion process that is formulated as a restoration task, where a latent diffusion model (LDM) is trained to denoise corrupted masks under the guidance of S (Sec. III-B). Central to nnQC, the learning process is conditioned by a set of embeddings, referred to as *opinions*, which are generated by a conditioning mechanism, denoted the Team of Experts (ToE) module (Sec. III-C). The ToE introduces 3D spatial awareness, represented by the relative spatial position of the axial slice (referred to as the *slice ratio*), and anatomical information derived from visual features extracted from the image I.

Within the training and inference of the proposed framework (Sec. III-D), we integrate the usage of fingerprint adaptation to ensure adaptability across organs, datasets, and imaging modalities (Sec. III-E) Figure 1 presents an overview of the proposed nnQC framework.

A. Manifold of Good Quality Segmentations

nnQC builds on the hypothesis that good-quality segmentations lie on a common manifold [13], [15], [16]. We learn such a high-quality manifold from a Variational Autoencoder (VAE) trained in an adversarial fashion, i.e., a VAE-GAN [25], [28], [34], using ground truth (GT) masks.

The 2D spatial VAE-GAN acts as a shallow autoencoder, applying a non-aggressive downsampling to the input GT mask dimension by a factor of 3: The encoder VAE_E learns to compress the high-dimensional input mask $x \in \mathbb{R}^{H\times W\times 1}$ into a low-dimensional latent representation $z = \mathcal{E}(x)$ where $z \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$. This dimensionality follows standard LDM compression configurations [25], [34], ensuring a compact latent space for efficient sampling while preserving sufficient high-frequency spatial details for accurate mask reconstruction. We set the latent space channel $C = 2$ to accommodate the expression of high-level segmentation features, while preserving a good trade-off of spatial relevance in the compressed latent space. As in [28], the spatial VAE is optimized with the following loss:

$$\begin{aligned}\mathcal{L}_{\text{VAE}} = \lambda_{\text{KLD}}\mathcal{L}_{\text{KLD}}\big(\text{VAE}_E(S) \parallel \mathcal{N}(0,1)\big) \\ + \lambda_{perc}\mathcal{L}_{perc}\big(S,\hat{S}\big) \\ + \lambda_{adv}\mathcal{L}_{adv}\big(D(S), D(\hat{S})\big) \\ + \lambda_{Dice}\mathcal{L}_{Dice}\big(S,\hat{S}\big)\end{aligned} \tag{1}$$

where S is an input segmentation mask (i.e., a GT mask), $\hat{S}$ is the reconstructed segmentation, and VAE_E the encoder of the VAE. $\mathcal{L}_{\text{KLD}}$ is the Kullback-Leibler divergence loss that

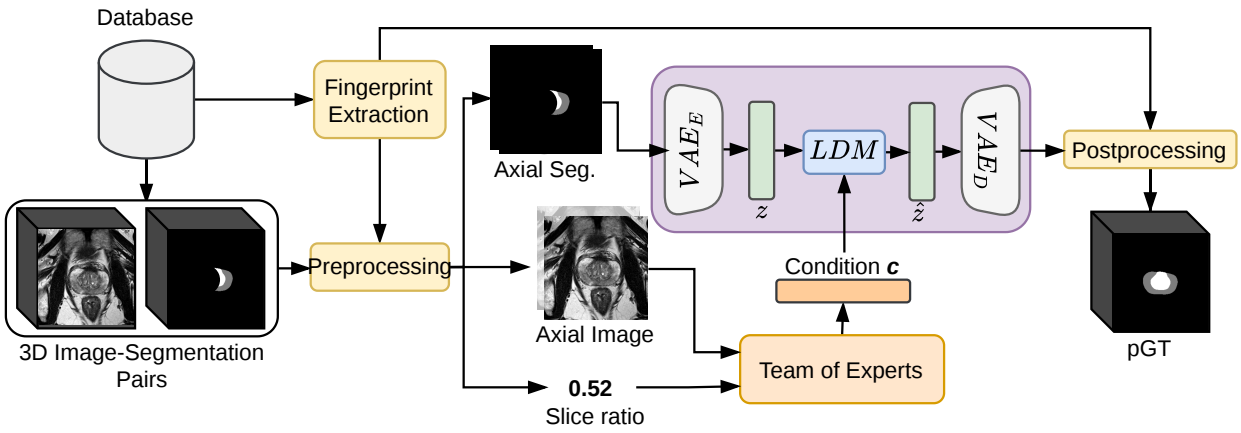

Fig. 1. The nnQC framework. For a 3D image–segmentation pair, dataset-specific *fingerprints* are extracted and used to preprocess it. Each axial segmentation slice and its corresponding 2D image are passed to the Team of Experts (ToE), which produces conditioning embeddings c for the latent diffusion process. A VAE-GAN maps the 2D segmentation to be quality checked into a latent space of high-quality segmentations from which a DDIM-based Latent Diffusion Model (LDM) generates a pseudo-ground truth (pGT). A postprocessing restores the pGT to its original space.

forces the latent space $\text{VAE}_E(S)$ to be normally distributed, $\mathcal{L}_{perc}$ denotes a perceptual loss [38], $\mathcal{L}_{Dice}$ represents the generalized Dice Loss, and $\mathcal{L}_{adv}$ is a patch-GAN adversarial loss [39] obtained by forwarding synthetic and real segmentations through a patch-GAN discriminator D [28]. We choose $\mathcal{L}_{Dice}$ as it allows the VAE to learn the spatial relationships among different classes in the input segmentation [13]; $\mathcal{L}_{perc}$ and $\mathcal{L}_{adv}$ are also included due to their proven effectiveness in improving reconstruction quality [28], [34]. The different λ coefficients serve as weights modulating the contribution of individual loss to $\mathcal{L}_{\text{VAE}}$.

B. Latent Diffusion Models for Pseudo Ground Truth Generation

Once the manifold of high-quality segmentations Z is learned, current approaches generate the pseudo-ground truths pGT_I^S by decoding Z in a deterministic fashion [13] or through iterative search of the learned latent space [15]. In nnQC, pGT_I^S is generated through a latent diffusion process that operates in the compressed, normative latent space Z learned by the VAE-GAN (Sec. III-A).

We cast the diffusion process as a *restoration* task [27]: a latent diffusion model (LDM) [34] is trained to iteratively denoise corrupted masks under the guidance of an auxiliary signal, namely the segmentation mask S to be quality controlled. For this purpose, the initial latent representation $z_0 \in Z$ is corrupted by injecting an *imperfect* segmentation.

To simulate a wide range of realistic *imperfect* segmentations, we synthetically corrupt the available GT masks using random morphological perturbations (see Sec. IV).

Following the objective function in [32], for a given timestep $t \in [0, T]$ of the reverse diffusion process, the LDM is trained to minimize

$$\mathcal{L}_{\text{LDM}} = \|\epsilon - \epsilon_\theta(z_{t,S}; c)\|_2^2, \tag{2}$$

where ϵ_θ is the learned function to predict the true noise $\epsilon \sim \mathcal{N}(0, 1)$ from $z_{t,S}$, the latent representation z_t corrupted with an imperfect segmentation S, given a condition c. In nnQC, we design the condition c to encode 3D spatial information and visual anatomical features derived from I to guide the denoising process. The mechanism to build c, which we refer to as *Team of Experts*, is presented in the following.

C. Team of Experts: Dual Embeddings for LDM Conditioning

We enforce nnQC to sample from the learned high-quality segmentations manifold by conditioning the LDM on two complementary feature sets extracted from I, guiding the sampling process for generating pGTs. Each feature set, or *opinion*, is derived from a specialized feature extractor, or *expert*. We denote the set of features as the *Team of Experts* (ToE) (Figure 2).

nnQC deals with 2D image–segmentation pairs extracted from 3D volume pairs, which have been proven to yield better results [12], but lack three-dimensional information that may be important to the sampling process. We address this limitation by injecting 3D contextual information into the conditioning.

To this end, we introduce Expert E_1 to encode the relative spatial position of the axial slices, expressed as a *slice-to-volume ratio* in the range $[0, 1]$, into a fixed-dimensional embedding vector, o_1. This is achieved through a lightweight Multi-Layer Perceptron (MLP) that receives the slice-to-volume ratio as input and outputs the positional embedding, o_1. E_1 is jointly trained with the LDM (Eq. 2) rather than separately. In this way, we ensure that the learned positional embedding space remains semantically aligned with the latent manifold learned by the VAE-GAN. Moreover, the shared optimization scheme prevents the collapse of the latent space and promotes spatial conditioning throughout the generation process.

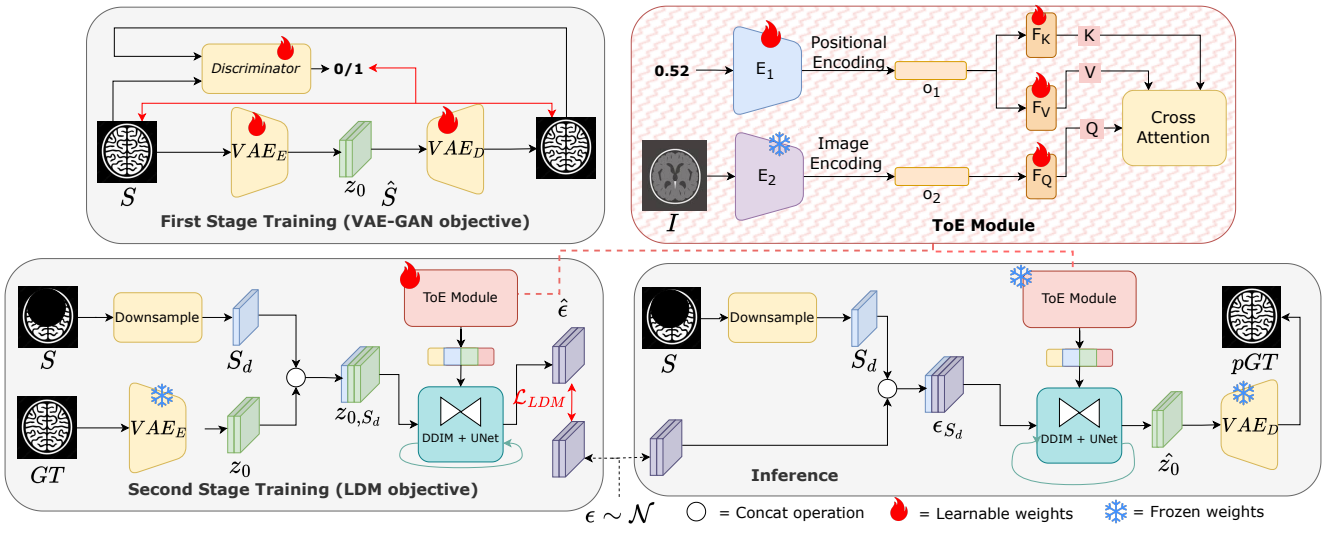

Fig. 2. Two-stage training and inference workflow. At the first stage (top left), the VAE is trained adversarially (i.e., a VAE-GAN) to learn a rich latent space of high-quality segmentations (i.e., GTs). During the training's second stage (bottom left), the LDM learns to reconstruct noise conditioned by embeddings from the *Team of Experts* (ToE) module (top right). The ToE's positional embedding is jointly optimized with the LDM. At inference (bottom right), Gaussian noise and S_d are fed into the LDM; the ToE-generated condition c guides the LDM to recover z_0, which is decoded by VAE_D to generate the pGT.

Previous works [15] have demonstrated that utilizing information from I yields better reconstruction than relying solely on the information conveyed by S [13]. nnQC follows a similar approach. However, rather than integrating information from I by reconstructing the (I, S) pair [15], which can add complexity to the training process, nnQC encodes visual features from I, within a vector o_2. This is achieved through the definition of Expert E_2, which leverages a pretrained CLIP-like vision encoder to extract high-level semantic features from I. By employing UniMedCLIP [40], a vision encoder pretrained in a large set of medical and clinical data, we expect to have anatomical information well-encapsulated in the resulting embedding, o_2. Unlike E_1, E_2 is used in a pre-trained fashion, thus freezing the vision encoder weights during the joint optimization of E_1 and the LDM.

To dynamically balance the opinions from the ToE, we utilize a Cross-Attention mechanism [41], [42] that serves as a *dynamic switch*. It assigns appropriate importance to each opinion, and it generates a unified conditioning vector c. This is achieved by projecting both o_1 and o_2 using linear layers F_Q, F_K, F_V to produce a query $Q = F_Q(o_1)$, key $K = F_K(o_2)$, and value $V = F_V(o_2)$. Afterwards, the conditioning vector c (Eq. 2) is thus obtained as the Cross-Attention vector

$$c = \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \qquad (3)$$

where d_k is the dimensionality of the keys, and fed as a condition to the diffusion process.

D. Two-stage Training and Inference Workflows

The nnQC framework follows a two-stage training design. Figure 2 illustrates the two-stage training and inference workflows.

1) Training: In the first stage, we train the VAE-GAN using adversarial training [28], [34]. The frozen VAE's encoder, i.e., VAE_E, is then used as part of the LDM's training, during the second stage.

We adopt a Denoising Diffusion Implicit Model (DDIM) [33], which constructs a non-Markovian forward process that preserves the same training objective as Denoising Diffusion Probabilistic Models (DDPMs) [32], but allows for a more efficient deterministic sampling procedure. For the LDM's internal model, we use a conditional UNet architecture [34], [43] as the network that learns the denoising process. During the second stage of training, we set the number of diffusion steps to $T = 1000$. We set $\epsilon \in \mathbb{R}^{2\times H/3\times W/3}$ to be consistent with the dimensionality of Z, as defined in Section III-A. Similarly, synthetically generated *imperfect* segmentations S are rescaled to the [0,1] range and downsampled to $S_d \in \mathbb{R}^{1\times H/3\times W/3}$. The resulting S_d is concatenated with the sampled z_0, forming the input $z_{0,S_d} \in \mathbb{R}^{3\times H/3\times W/3}$ for the diffusion UNet. Crucially, the optimization of the LDM is shared with the trainable components of the Team of Experts (ToE). Specifically, the gradients derived from minimizing $\mathcal{L}_{LDM}$ (Eq. 2) are backpropagated to update both the weights of the denoising U-Net and the parameters of the positional expert E_1, along with the linear projection layers of the cross-attention mechanism (F_Q, F_K, F_V). This joint training strategy ensures that the generated conditioning embeddings c are semantically aligned with the latent diffusion space, preventing latent collapse and enabling effective guidance for the restoration task.

2) Inference: We leverage the efficiency of DDIM sampling, which offers a flexible trade-off between sample quality and generation speed, substantially reducing computational costs [33]. While DDIMs typically use 50 steps [33], we empirically reduce the number of sampling steps to $T = 20$ as two factors mitigate the generation complexity: 1) the

concatenation of the input mask with the input noise, which injects a strong bias into the process, and 2) limiting the image domain to binary masks with pixels constrained to $\{0, 1\}$.

During the sampling process, inference mirrors training: a randomly sampled Gaussian noise $\epsilon \in \mathbb{R}^{2\times H/3\times W/3}$ is concatenated with the rescaled and downsampled segmentation mask S to be quality controlled, yielding the input noise $\epsilon_{S_d} \in \mathbb{R}^{3\times H/3\times W/3}$. This input is denoised by the trained UNet to reconstruct the latent representation z_0 by reversing the diffusion process. Finally, the denoised latent sample is decoded by the VAE decoder, VAE_D, to produce pGT_I^S.

E. Fingerprints for Self-Adaptable QC

Inspired by nnUNet [18], we use fingerprints to enable our framework to self-adapt to various data types and conditions. We define the *fingerprints* as a set of key characteristics that describe the input dataset: the median voxel spacing of subject volumes, the median size of foreground regions, image orientation, intensity ranges specific to each modality, and the number of unique segmentation classes. These fingerprints form the basis for dataset-specific adaptations during both data pre-processing and post-processing, as well as at the network/model level. At pre-processing, the fingerprints guide image rescaling. The median voxel spacing and the median cropped volume size are used to standardize the image dimensions (256×256), while image contrast is scaled based on the 0.5 and 99.5 percentile intensity values within the foreground regions [18], ensuring modality-specific normalization. The rescaled images are aligned to a predefined orientation (right, anterior, superior, *"RAS"*). During post-processing, the fingerprints serve to restore the original resolution.

At the network level, the fingerprints allow the selection of the number of input and output channels of the VAE's first and last layers using the number of segmentation labels in the dataset. Unlike the intensive fingerprint-based adaptation process in nnUNet [18], we leverage the intrinsic adaptability of LDMs to operate within a predefined image space [34]. As a result, we use a homogeneous latent size across all datasets, which simplifies training while preserving flexibility.

IV. Experiments and Results

A. Experimental design and setup

1) Datasets: We conduct experiments on 15 datasets covering seven organ types, three imaging modalities - magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound (US) - and a varying number of annotated structures (labels). We use six datasets from the Medical Segmentation Decathlon (MSD) challenge [44], encompassing Spleen (61 CT scans, 1 label), Prostate (48 MRI volumes, 2 labels), Heart (30 MRI volumes, 1 label), Liver (210 CT volumes, 1 label), Pancreas (420 CT volumes, 1 label), and Hippocampus (394 MRI volumes, 2 classes); the ACDC (150 MRI volumes, 3 labels), M&M-2 (360 MRI volumes, 3 labels), and CAMUS datasets (500 US scans, 3 labels) for heart segmentation; KiTS 2021 [45] (300 CT scans, 1 label) for kidney segmentation; CHAOS 2021 [46] for abdominal organ segmentation from MR images (40 MRI volumes, 1 label per organ: kidney, spleen, and liver); PROSTATE-X [47] (346 MRI volumes, 2 labels) for prostate segmentation; and AbdomenCT-1K [48] (Abd1K-CT; 1000 CT volumes, 1 label for spleen and for liver).

2) Benchmarks: We consider three reconstruction-based QC baselines for comparison: *(1) Galati et al.* [13] a deterministic reconstructor based on a Convolutional Autoencoder, which reconstructs segmentation masks to restore their original shape; *(2) Liu et al.* [16] a two-stage regressor that uses a VAE trained to learn the normative good-quality manifold of GTs and an MLP that processes the features generated by the latent space obtained from the reconstructed segmentation to predict the a pseudo Dice score; and *(3) Wang et al.* [15], a VAE that processes the channel-wise concatenation of image-segmentation pairs, and adjusts their compressed embeddings in the latent-space using a stochastic iterative search. Unlike other baselines which show missing training setups and architectural choices, which fundamentally limited our ability to directly reproduce and benchmark against their method in our experimental setup, our benchmark selection is motivated by the public availability of implementation details.

3) Evaluation Metrics: We assess performances using the Pearson correlation (*r*) and the Mean Absolute Error (MAE) between the predicted pseudo-quality scores (using a pGT) and real quality scores (using the available GT). We use the Dice-Sørensen Coefficient (DSC) and the 95% Hausdorff Distance (HD95) as quality metrics. In the cross-dataset experiment, we use the MAE between predicted and real scores.

4) Setup & Implementation Details: We adopt an 80–20% training–testing split at the subject level. We use GT labels to learn the manifold of GT segmentations (Sec. III-A) at the first stage of training. During the second stage, we simulate segmentations of varying quality by corrupting the GT from the considered datasets through synthetic degradations [22], [23], allowing the LDM to learn how to recover good-quality segmentation masks from degraded ones. This choice is motivated by the need to generate a diverse and unbiased training distribution of mask quality levels. While one could in principle use the outputs of a specific segmentation model to obtain masks at varying quality, this would introduce an architectural bias, as the framework would be exposed only to the failure modes and artefact patterns characteristic of that particular model, ultimately compromising its ability to generalise across different segmentation sources [10], [12]. By contrast, corrupting ground-truth masks directly through synthetic degradations produces a broad and model-agnostic spectrum of quality levels, an approach already adopted in the medical imaging literature [22], [23]. The GTs are degraded to five distinct levels, corresponding to uniformly spread DSC intervals of [0.05-0.10), [0.10-0.25), [0.25-0.50), [0.50-0.75), and [0.75-0.95]. We implement the degradation pipeline including: (i) *Morphological Perturbations* with random erosion and dilation operations (kernel sizes 3-7) simulate systemic under- and over-segmentation errors, mirroring the boundary uncertainty often observed in nnU-Net and MedSAM [17] when tissue contrast is low; (ii) *Cutout and Random Holes* by randomly mask out regions within the organ to simulate false negatives and mimic the 'missed region' artifacts; (iii)

Additive Noise (False Positives) with the injection of random mask blobs into the background to simulate false positives, a frequent failure mode in Transformer-based models (e.g., SwinUNETR [49]). During testing, GTs are also subject to degradation through the same procedure. The resulting test set comprises a total of 9,370 2D slices. We aggregate the 2D predictions on 3D volumes to compute the evaluation metrics at the subject level.

For the benchmark models, we follow the guidelines of the respective studies. For [13] and [15], we rely on the publicly available codebase provided by the authors, while for [16], we implemented their pipeline and model architecture as described in their study. For both nnQC and benchmark models, we train one model per organ, i.e., for CHAOS we do not train a single model across all abdominal organs, but rather a separate model for each organ. All code is developed with Python 3.10, along with Python's MONAI library and PyTorch 2.0 for the implementation of the end-to-end pipeline. Training experiments are run on an 80 GB NVIDIA A100, brought by the French national cluster IDRIS on Jean-Zay machines with a 12.4 CUDA version (average GPU memory consumption with a batch size of 32 is around 32Gb).

B. Results

1) Benchmark Study: We assess the performance of nnQC and compare it against benchmark methods on the synthetically degraded marks from ten datasets (PROSTATE-X and M&M-2 are excluded). Figure 3 reports the Pearson correlation coefficient (r) and Figure 4 the obtained MAE.

Models relying solely on mask information, such as [13], perform poorly with an average DSC MAE of 0.36 ± 0.10, HD95 MAE of 19.2 ± 3.3, and low correlations (DSC $r = 0.21 \pm 0.23$, HD95 $r = 0.13 \pm 0.12$). Liu et al. [16] achieves a better, but yet limited performance with an average DSC MAE of 0.25 ± 0.05 and moderate correlations (mean DSC $r = 0.62 \pm 0.11$). Both approaches exhibit broad, heavy-tailed error distributions, reflecting limited robustness across organs. Instead, models that also use information from the original image report a competitive performance, as observed in the results from Wang et al. [15] (mean DSC $r = 0.77 \pm 0.09$, mean HD95 $r = 0.78 \pm 0.12$, average DSC MAE of 0.16 ± 0.15 and average HD95 MAE of 11.2 ± 3.84), confirming the importance of also encoding information from the original image.

Nonetheless, nnQC consistently reports a better performance across organs, both in terms of r (mean DSC $r = 0.89 \pm 0.03$ and HD95 $r = 0.94 \pm 0.02$) and MAE (DSC MAE 0.12 ± 0.06 and HD95 MAE 9.54 ± 2.33) outperforming the baseline models. For instance, Wang et al. achieve the best correlations in heart-related datasets, such as ACDC and CAMUS; however, their performance degrades in other datasets, including MSD Pancreas and MSD Liver. This can be explained by the fact that Wang's original model [15] has been conceived for heart segmentation QC. Instead, nnQC has been designed to be easily adapted across organs, datasets, and imaging techniques, which is reflected in its consistent performance across different scenarios.

TABLE II
CROSS-DATASET PERFORMANCE. MODELS ARE TRAINED ON MSD PROSTATE, ACDC, MSD SPLEEN AND MSD LIVER. BOLD DENOTES BEST.

		DSC MAE	HD95 MAE (mm)
PROSTATEx	Wang	0.20 ± 0.06	9.84 ± 4.12
	nnQC	$\mathbf{0.09 \pm 0.03}$	$\mathbf{5.23 \pm 2.67}$
M&M-2	Wang	0.15 ± 0.08	7.62 ± 3.88
	nnQC	$\mathbf{0.10 \pm 0.03}$	$\mathbf{4.41 \pm 2.03}$
Abd1K-CT Spleen	Wang	0.17 ± 0.09	12.43 ± 5.21
	nnQC	$\mathbf{0.08 \pm 0.04}$	$\mathbf{6.87 \pm 3.14}$
Abd1K-CT Liver	Wang	0.14 ± 0.14	18.62 ± 7.83
	nnQC	$\mathbf{0.13 \pm 0.08}$	$\mathbf{9.24 \pm 4.56}$

2) Cross-dataset generalization: We assess nnQC's generalization capabilities through a cross-dataset evaluation by using an out-of-distribution (OOD) testing dataset, i.e., different from those used for training. Models trained on MSD Prostate and ACDC are tested on PROSTATEx and M&M-2 heart, with 380 and 160 subjects, respectively, comprising a total of 9226 axial slices. Additionally, to further assess generalization across different organs and imaging modalities, we evaluate models trained on MSD Spleen and MSD Liver on the corresponding OOD splits from the Abd1K-CT. Table II presents the obtained results in terms of the MAE between predicted and real DSC and HD95, and compares them against Wang et al. [15], the best competing model in the benchmark study.

nnQC reports low MAE values that closely match those achieved in the in-distribution (ID) dataset (see Section IV-B1), where a MAE of 0.14 ± 0.04 was recorded for the MSD Prostate dataset, and 0.07 ± 0.04 for the ACDC dataset. Notably, we found that the MAE in the PROSTATEx dataset is lower than that of the ID data, underscoring nnQC's ability to generalize to unseen OOD data. In contrast, Wang et al. [15] show a drop in performance when exposed to OOD data, as evidenced by an increase in MAE compared to the values obtained from the ID data (0.16 ± 0.07 for MSD Prostate and 0.06 ± 0.02 for ACDC). The generalization trend is confirmed on the abdominal datasets: nnQC achieves a DSC MAE of 0.08 ± 0.04 and 0.13 ± 0.08 on Abd1K-CT Spleen and Liver, respectively, consistently outperforming Wang et al. across both organs and both metrics. The larger HD95 MAE observed for the Liver dataset reflects the inherent complexity of liver boundary delineation, which is a well-known challenge in abdominal CT segmentation. The superior generalization capabilities of nnQC can be attributed to the use of encoder-derived representations from UniMedCLIP's vision encoder, which has been pre-trained on large, diverse medical datasets. The pre-training makes nnQC more robust to domain shifts. Furthermore, the use of relative positional encodings helps disambiguate spatial structures, ensuring consistent performance even in the presence of OOD data.

3) Model ranking: We assess whether the pseudo-quality scores produced through nnQC can be used for model ranking. To that end, we consider three state-of-the-art medical image segmentation frameworks, nnUNet [18], MedSAM [17], and SwinUNETR [49], along with two reference baselines emulating a perfect model and a low-performance one. For

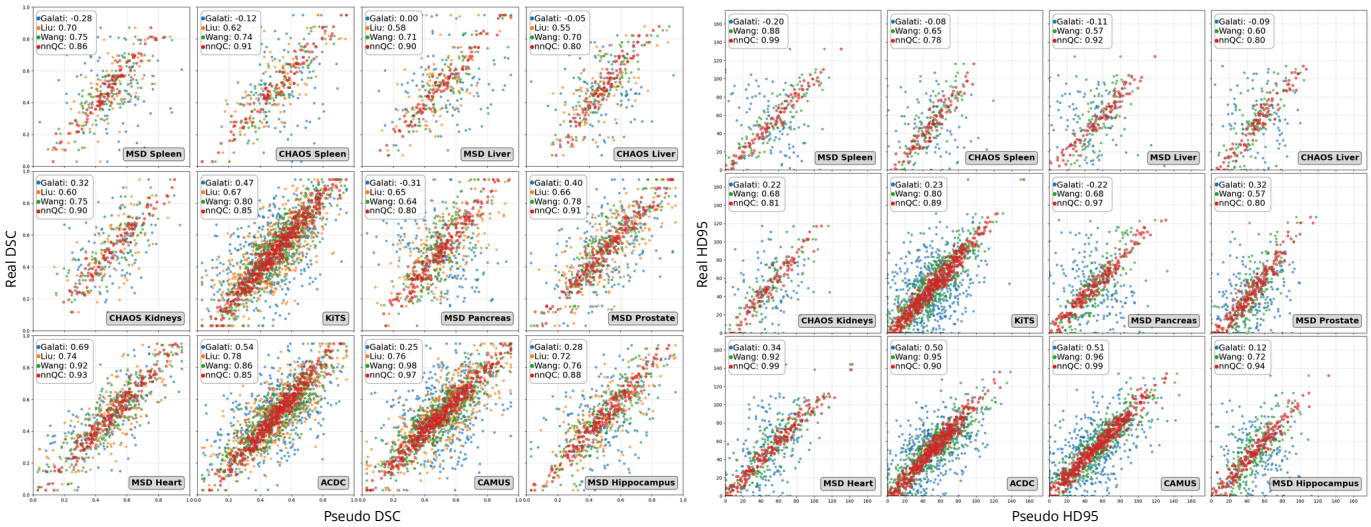

Fig. 3. Pearson correlation (r) between the predicted pseudo-quality scores and real scores (DSC and HD95) across different organs, modalities, and datasets. HD95 is not estimated for Liu et al [16] as their model is designed to predict pseudo DSCs.

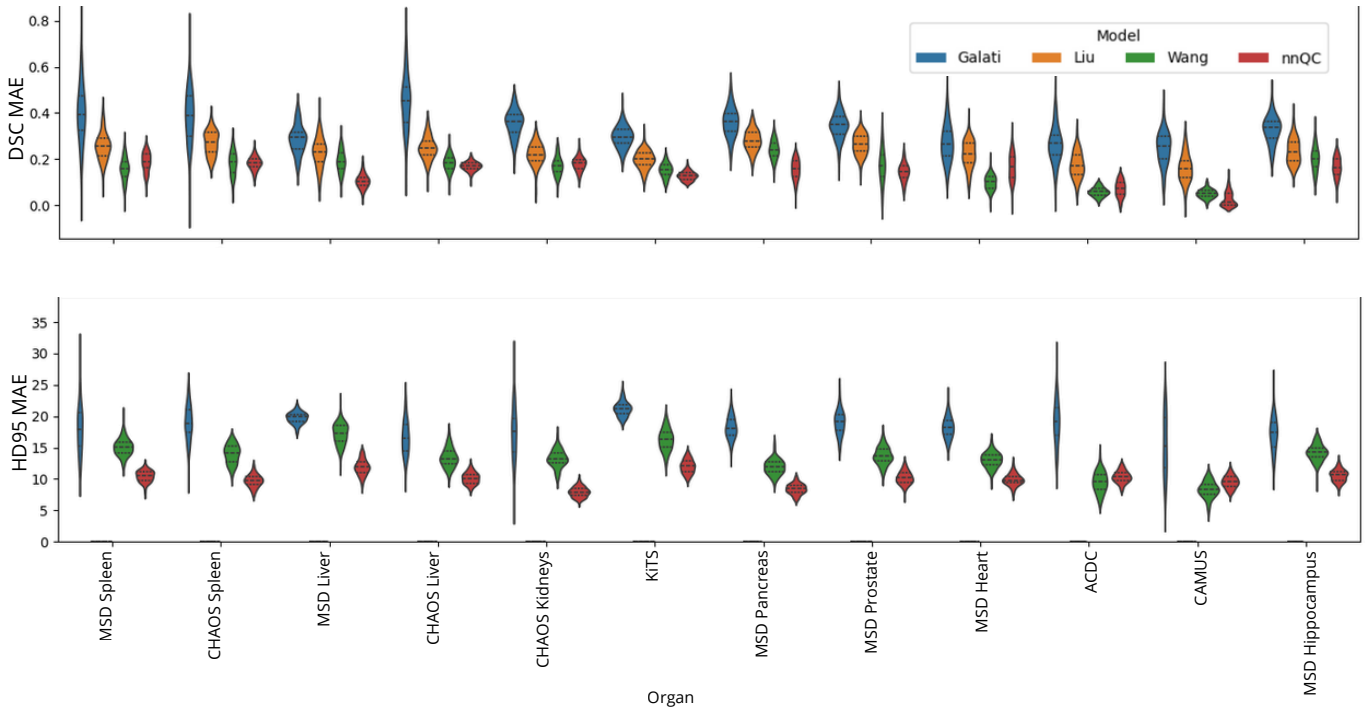

Fig. 4. Mean Absolute Error (MAE) distribution across different organs, modalities, and datasets. The MAE is measured as the difference between the predicted pseudo-quality scores and real scores (DSC and HD95). As in Fig 3, HD95 is not estimated for Liu et al [16].

the first one, we use the GT masks. For the second one, we rely on an atlas-based segmentator using ANTs [50] with five image-segmentation pairs from the training set as atlases, employing a joint-fusion policy to segment the unseen images. We generate segmentations across three cardiac datasets (MSD Heart, ACDC, and CAMUS) encompassing different imaging techniques and semantic labels. We use the pseudo-DSC obtained from nnQC to rank the five models and compare these rankings with those obtained using the GT. Ranking agreement is measured with Kendall's τ test (Table III).

In MSD Heart (late gadolinium enhancement MRI), nnQC perfectly reproduces the ranking that would be obtained using the ground truth. For ACDC and CAMUS(MRI and US), the rankings are reproduced with the exception of two swaps between MedSAM and SwinUNETR (ACDC) and between nnUNet and SwinUNETR (CAMUS), corresponding to a Kendall's $\tau = 0.80$. These discrepancies are likely due to subtle differences in performance between the models. To validate this hypothesis, we performed a t-tests on the real DSC distributions for each pair of models involved in a rank swap to assess whether there is a significant difference between the performance (in terms of the DSC) of the two models. The

TABLE III
NNQC-BASED MODEL RANKING VS GT-BASED RANKING ON THREE CARDIAC DATASETS FROM THREE DIFFERENT IMAGE MODALITIES. KENDALL'S τ MEASURES THE SIMILARITY OF THE TWO RANKINGS.

Dataset	Model	nnQC Rank	GT Rank	τ
MSD Heart	GT	1	1	1.00
	nnUNet	2	2	
	MedSAM	3	3	
	SwinUNETR	4	4	
	ANTs	5	5	
ACDC	GT	1	1	0.80
	nnUNet	2	2	
	MedSAM	4	3	
	SwinUNETR	3	4	
	ANTs	5	5	
CAMUS	GT	1	1	0.80
	nnUNet	3	2	
	SwinUNETR	2	3	
	MedSAM	4	4	
	ANTs	5	5	
Average Kendall's τ				0.87

TABLE IV
ABLATION STUDY ON THE CHAOS LIVER AND CAMUS DATASETS. BOLD DENOTES BEST PERFORMANCE.

Dataset	ToE Configuration	DSC r	HD95 r	DSC MAE
CHAOS Liver	no conditioning	0.66	0.59	0.32 ± 0.18
CHAOS Liver	with Image Encoding	0.72	0.75	0.18 ± 0.08
CHAOS Liver	with Positional Encoding	**0.85**	0.75	0.20 ± 0.04
CHAOS Liver	Full Model	0.80	**0.80**	**0.17 ± 0.03**
CAMUS	no conditioning	0.71	0.43	0.27 ± 0.25
CAMUS	with Image Encoding	0.86	0.88	0.12 ± 0.07
CAMUS	with Positional Encoding	**0.90**	0.92	0.11 ± 0.04
CAMUS	Full Model	0.89	**0.97**	**0.05 ± 0.04**

t-test yielded p-values of 0.704 (MedSAM vs SwinUNETR in ACDC) and 0.112 (nnUNet vs SwinUNETR in CAMUS), indicating that the observed rank swaps occur in settings where the performance differences are not statistically significant.

4) Ablation study: We conduct an ablation study to understand the role of the *opinions* from the ToE module in the framework's performance. In particular, we study performance as we remove the cross-attention module and disable one expert at a time to condition the LDM. For the study, we consider two datasets: CAMUS, where nnQC performs best, and CHAOS Liver, where nnQC performance is the lowest (Figures 3 and 4). Table IV reports the obtained results.

The ablation studies reveal consistent results across datasets, highlighting the importance of 3D spatial information from positional encodings for model performance. Using positional encodings alone yields the highest DSC r, while image encodings have a lower performance on their own. This is likely due to subtle changes in appearance (i.e., texture and intensity), making image encodings less informative. Nonetheless, the full model performs best, indicating that the information from both *experts* is complementary and enhances nnQC's performance.

5) Qualitative latent-retrieval analysis: Lastly, we study the learned latent representations across nnQC and the baselines. Figure 5 shows 2D projections of the different normative learned manifolds and their respective centroids in ACDC. Using a randomly selected sample for QC, we visualize its location and that one of the reference GT in the latent space, as well as the corresponding reconstructed pGT. Additionally, we display the reconstructed centroid, as it provides insights into the model's implicit *idea* of the represented domain [52], or, in this case, its average understanding of anatomical variability.

In Galati et al. [13] and Liu et al. [16] the centroids exhibit abnormal reconstructions, which may reflect on the quality of the learned latent representation. Specifically, in [13], the reconstucted consists of a flat mask dominated by one class with scattered artifacts from other classes, lacking any relevant semantic information. As a result, when faced with a poor-quality segmentation, the model collapses into a blank pGT. Similarly, in [16], the reconstructed centroid mask displays fragmented and inconsistent contours, leading to an incomplete and erroneous pGT in the example. Instead, Wang et al. [15] present a centroid that corresponds to a segmentation mask with a well-defined anatomical shape, indicating that the latent space encodes a strong anatomical prior, which in turn enables the model to generate anatomically plausible shapes. Nonetheless, this smooth "average" shape suggests a learned latent representation that cannot fully capture the high variability across shapes, which may stem from the limited size of the latent encoding (i.e., $\mathbb{R}^{16}$). The plausible but *anatomically incorrect* pGT in Figure 5 (where the right ventricle class is not generated) can be further explained by the iterative sampling mechanism implemented in [15], which stops once it retrieves a plausible shape. This behavior suggests that the model's conditioning on the intensity image is insufficient to guide the sampling process effectively.

In contrast, nnQC's reconstructed centroid can be described as a topological template of the considered anatomy. Although the contours are noisier, the reconstructed centroid preserves the spatial relationship between anatomical structures (e.g., left ventricle enclosed by myocardium and myocardium adjacent to right ventricle). Unlike [15], ours captures a more abstract concept of the anatomy, encompassing anatomical variability rather than a concrete shape instance, as a direct consequence of the richer 2D latent space. This latent representation offers a meaningful starting point for the ToE-conditioned diffusion process, which then refines this anatomical template into subject-specific reconstruction variations, where the sampled pGT closely resembles the corresponding GT (Figure 5).

6) Computational cost analysis: We compare the training and inference time of nnQC against the baseline methods. Training time is reported per epoch (with consistent batch sampling across datasets), while inference time is measured by processing all slices of 10 randomly selected subjects per dataset and averaging the results (Table V).

On an NVIDIA A100 GPU, the total training time per epoch for nnQC is 204.5s, which is higher than Wang et al. (92.0s) and Liu et al. (148.1s). This increased training cost reflects the added complexity of our Latent Diffusion Model (LDM) architecture and two-stage training pipeline (where the training time refers to the total of both stages). This reflects a trade-off with the improved performance and robustness observed in our results across diverse organs. For inference, while nnQC has higher latency (404ms per sample) than Wang et al. (32ms) and Galati et al. (12ms) due to iterative sampling, it remains well

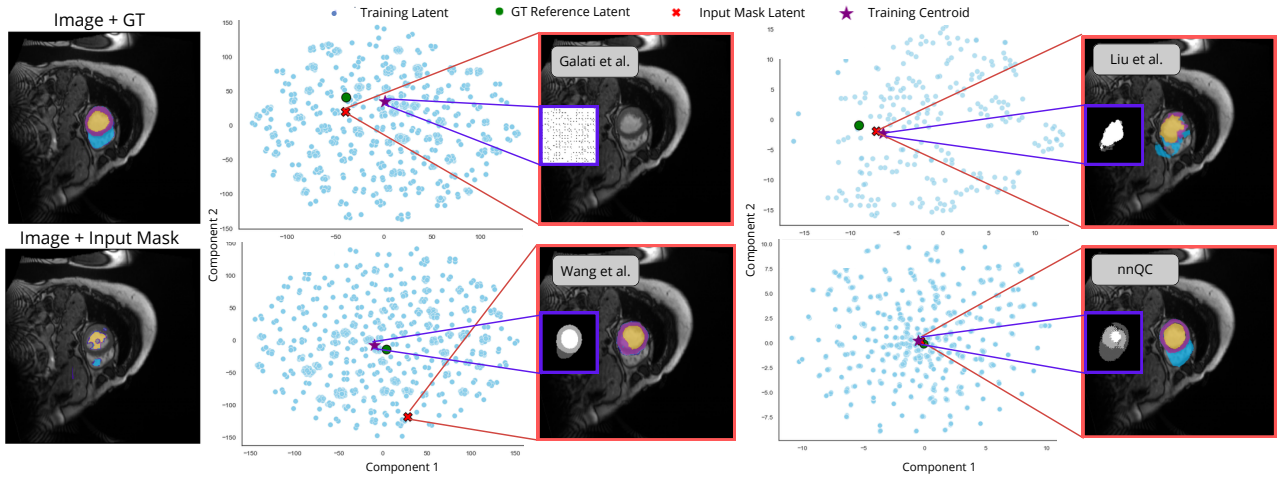

Fig. 5. Learned normative manifolds and generated pGTs from a low-quality input segmentation from the ACDC dataset. The first column shows the GT and a low-quality segmentation overlaid in the original image. The following blocks display the latent spaces learned by different QC methods, the reconstructed pGTs (red box), and the reconstructed centroids (purple box). The projected manifolds are obtained using t-SNE [51].

TABLE V
COMPUTATIONAL COST COMPARISON AMONG THE PROPOSED BENCHMARKS AND NNQC.

Model	Liu et al.	Galati et al.	Wang et al.	nnQC (Ours)
Training time per epoch (s)	148.1	23.6	92.0	204.5
Inference time per sample (s)	0.226 ± 0.167	0.012 ± 0.009	0.032 ± 0.017	0.404 ± 0.221

within practical subsecond latency for offline quality control workflows. This reflects the trade-off between computational cost and reconstruction quality.

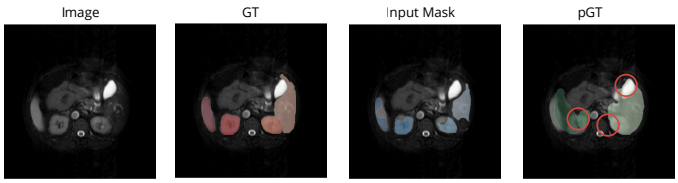

Fig. 6. Failed pGT sampling when multiple organ classes are present in the input, in this case, liver, spleen, left and right kidneys. Red circles indicate anatomical inconsistencies in the pGT.

V. CONCLUSION

In this work, we introduced *nnQC*, a model- and metric-agnostic quality control framework for segmentation masks that generates reliable pseudo-ground truths through a novel sampling strategy. At its core, nnQC features a *Team of Experts (ToE)* module that independently processes the input image and relative axial position by using cross-attention as a dynamic mechanism to balance their contributions Furthermore, nnQC extracts dataset-specific *fingerprints* that allow for automatic adaptation to a wide range of anatomical structures and imaging modalities. Extensive experiments across twelve datasets, seven organs and three image modalities demonstrated that nnQC outperforms state-of-the-art methods, confirming itself as a versatile QC solution, that can robustly handle high- and low-quality segmentations across organs and imaging modalities.

We have, however, identified some pending limitations. First, external experiments indicate that nnQC struggles with complex multi-organ segmentations, where recovering accurate inter-class topological relationships becomes difficult. For instance, Figure 6 illustrates a pGT failure on a multi-organ scenario (CHAOS dataset). We hypothesize this behavior arises from the large spatial separation among organ classes, which hinders nnQC from forming a coherent, normative segmentation *template*. Currently, we circumvent this by using separate models for each organ, but it would be desirable to have a single model to handle QC across all organs in an image. Second, the current evaluation excludes highly heterogeneous structures, such as tumors or vascular structures. This choice stems from the inherent difficulty of embedding such structures within a learned normative manifold, as their irregular shapes and heterogeneity prevent including them in a single "good-quality" latent representation. Finally, we acknowledge that the relative spatial position encoding employed in nnQC assumes a certain consistency in the Field-of-View (FOV) across training and test acquisitions. In scenarios where a significant FOV mismatch exists between datasets, for example, when one acquisition covers the full anatomy, and another captures only a partial slab, the positional encoding may become inconsistent, representing a boundary condition that could affect the reliability of the generated pseudo-ground-truth masks. To address these limitations, future work may explore: (1) the incorporation of a topological interaction loss to better capture inter-class spatial dependencies;and (2) the extension of nnQC toward a fully 3D formulation, enabling the model to leverage volumetric context for more reliable spatial reasoning.

REFERENCES

[1] S. K. Zhou, H. Greenspan, C. Davatzikos, J. S. Duncan, B. Van Ginneken, A. Madabhushi, J. L. Prince, D. Rueckert, and R. M. Summers, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, 2021.

[2] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, and et al., "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved?," *IEEE Transactions on Medical Imaging*, 2018.

[3] R. Robinson, V. V. Valindria, W. Bai, and et al., "Automated quality control in image segmentation: application to the UK Biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance*, 2019.

[4] B. Billot, C. Magdamo, Y. Cheng, S. E. Arnold, S. Das, and J. E. Iglesias, "Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets," *Proceedings of the National Academy of Sciences*, 2023.

[5] T. Kohlberger, V. Singh, C. Alvino, C. Bahlmann, and L. Grady, "Evaluating segmentation error without ground truth," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2012.

[6] V. V. Valindria, I. Lavdas, W. Bai, K. Kamnitsas, E. O. Aboagye, A. G. Rockall, D. Rueckert, and B. Glocker, "Reverse classification accuracy: predicting segmentation performance in the absence of ground truth," *IEEE Transactions on Medical Imaging*, 2017.

[7] J. Kalkhof and A. Mukhopadhyay, "M3D-NCA: Robust 3d segmentation with built-in quality control," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, vol. 14220, pp. 169–178, 2023.

[8] P. Qiu, S. Chakrabarty, P. Nguyen, S. S. Ghosh, and A. Sotiras, "QCResUNet: Joint subject-level and voxel-level prediction of segmentation quality," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2023.

[9] Q. Lin, X. Chen, C. Chen, and J. M. Garibaldi, "A novel quality control algorithm for medical image segmentation based on fuzzy uncertainty," *IEEE Transactions on Fuzzy Systems*, vol. 31, no. 8, pp. 2532–2544, 2022.

[10] R. Robinson, O. Oktay, W. Bai, V. V. Valindria, M. M. Sanghvi, N. Aung, J. M. Paiva, F. Zemrak, K. Fung, E. Lukaschuk, A. M. Lee, V. Carapella, Y. J. Kim, B. Kainz, S. K. Piechnik, S. Neubauer, S. E. Petersen, C. Page, D. Rueckert, and B. Glocker, "Real-time prediction of segmentation quality," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, 2018.

[11] B. Audelan and H. Delingette, "Unsupervised quality control of image segmentation based on Bayesian learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, 2019.

[12] J. Fournel, A. Bartoli, D. Bendahan, M. Guye, M. Bernard, E. Rauseo, M. Y. Khanji, S. E. Petersen, A. Jacquier, and B. Ghattas, "Medical image segmentation automatic quality control: A multi-dimensional approach," *Medical Image Analysis*, 2021.

[13] F. Galati and M. A. Zuluaga, "Efficient model monitoring for quality control in cardiac image segmentation," in *Functional Imaging and Modeling of the Heart (FIMH 2021)*, vol. 12738, pp. 101–111, 2021.

[14] B. Specktor-Fadida, L. Ben-Sira, D. Ben-Bashat, and L. Joskowicz, "SegQC: a segmentation network-based framework for multi-metric segmentation quality control and segmentation error detection in volumetric medical images," *Medical Image Analysis*, vol. 103, p. 103638, 2025.

[15] S. Wang, G. Tarroni, C. Qin, Y. Mo, C. Dai, C. Chen, B. Glocker, Y. Guo, D. Rueckert, and W. Bai, "Deep generative model-based quality control for cardiac MRI segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, 2020.

[16] F. Liu, Y. Xia, D. Yang, A. L. Yuille, and D. Xu, "An alarm system for segmentation algorithm based on shape model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[17] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[18] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: Self-adapting framework for deep learning-based bioedical image segmentation," *Nature Methods*, 2021.

[19] Jin, Xiyao and Hao, Yao and Hilliard, Jessica and Zhang, Zhehao and Thomas, Maria A and Li, Hua and Jha, Abhinav K and Hugo, Geoffrey D, "A quality assurance framework for routine monitoring of deep learning cardiac substructure computed tomography segmentation models in radiotherapy," *Medical physics*, vol. 51, no. 4, pp. 2741–2758, 2024.

[20] Arega, Tewodros Weldebirhan and Bricq, Stéphanie and Legrand, François and Jacquier, Alexis and Lalande, Alain and Meriaudeau, Fabrice, "Automatic uncertainty-based quality controlled T1 mapping and ECV analysis from native and post-contrast cardiac T1 mapping images using Bayesian vision transformer," *Medical image analysis*, vol. 86, p. 102773, 2023.

[21] Qiu, Peijie and Chakrabarty, Satrajit and Nguyen, Phuc and Ghosh, Soumyendu Sekhar and Sotiras, Aristeidis, "QCResUNet: Joint subject-level and voxel-level segmentation quality prediction," *Medical Image Analysis*, 2025.

[22] Aresta, Guilherme and Bogunović, Haris, "FAZ Segmentation Quality Assessment in OCTA via Denoising Autoencoders and Segmentation Uncertainty Estimation," in *Medical Imaging with Deep Learning-Short Papers*, 2025.

[23] Jebril, Haneen and Pinetz, Thomas and Bogunović, Haris, "Shape Prior For Quality Assessment in OCTA via Denoosing Autoencoders at the Segmentation Level," *IEEE Access*, 2025.

[24] Li, Kang and Yu, Lequan and Heng, Pheng-Ann, "Towards reliable cardiac image segmentation: Assessing image-level and pixel-level segmentation quality via self-reflective references," *Medical Image Analysis*, 2022.

[25] W. H. Pinaya, P.-D. Tudosiu, J. Dafflon, P. F. Da Costa, V. Fernandez, P. Nachev, S. Ourselin, and M. J. Cardoso, "Brain imaging generation with latent diffusion models," in *MICCAI workshop on deep generative models*, 2022.

[26] P.-D. Tudosiu, W. H. Pinaya, P. F. Da Costa, J. Dafflon, A. Patel, P. Borges, V. Fernandez, M. S. Graham, R. J. Gray, P. Nachev, *et al.*, "Realistic morphology-preserving generative modelling of the brain," *Nature Machine Intelligence*, vol. 6, pp. 811–819, 2024.

[27] C. I. Bercea, M. Neumayr, D. Rueckert, and J. A. Schnabel, "Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models," in *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

[28] V. Fernandez, W. H. L. Pinaya, P. Borges, M. S. Graham, P.-D. Tudosiu, T. Vercauteren, and M. J. Cardoso, "Generating multi-pathological and multi-modal images and labels for brain MRI," in *Deep Generative Models (DGM4MICCAI), MICCAI Workshop*, vol. 13609, pp. 117–126, 2022.

[29] S. Gupta, D. Samaras, and C. Chen, "Topodiffusionnet: A topology-aware diffusion model," in *International Conference on Learning Representations (ICLR)*, 2025.

[30] D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," *International Conference on Machine Learning*, 2015.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014.

[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, 2020.

[33] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations (ICLR)*, 2021.

[34] Rombach, Robin and Blattmann, Andreas and Lorenz, Dominik and Esser, Patrick and Ommer, Björn, "High-resolution image synthesis with latent diffusion models," pp. 10684–10695, 2022.

[35] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca, "Universeg: Universal medical image segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6830–6839, 2023.

[36] H. E. Wong, M. Rakic, J. Guttag, and A. V. Dalca, "Scribbleprompt: fast and flexible interactive segmentation for any biomedical image," in *European Conference on Computer Vision*, vol. 15098, pp. 3–19, 2024.

[37] F. Isensee, T. Wald, C. Ulrich, M. Baumgartner, S. Roy, K. Maier-Hein, and P. F. Jaeger, "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, vol. 14999, pp. 488–498, 2024.

[38] Q. Yang, P. Yan, Y. Zhang, H. Yu, Y. Shi, X. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1348–1357, 2018.

[39] S. Gur, S. Benaim, and L. Wolf, "Hierarchical patch vae-gan: Generating diverse videos from a single sample," *Advances in Neural Information Processing Systems*, 2020.

[40] M. U. Khattak, S. Kunhimon, M. Naseer, S. Khan, and F. S. Khan, "Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities," *arXiv preprint arXiv:2412.10372*, 2024.

[41] C.-F. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the*

IEEE/CVF international conference on computer vision, pp. 357–366, 2021.

[42] D. Rebain, M. J. Matthews, K. M. Yi, G. Sharma, D. Lagun, and A. Tagliasacchi, "Attention beats concatenation for conditioning neural fields," *arXiv preprint arXiv:2209.10684*, 2022.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 2015.

[44] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, *et al.*, "The medical segmentation decathlon," *Nature communications*, 2022.

[45] N. Heller, F. Isensee, D. Trofimova, R. Tejpaul, N. Papanikolopoulos, and C. Weight, eds., *Kidney and Kidney Tumor Segmentation*. 2022.

[46] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar, D. Lachinov, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, D. Sheet, G. Dovletov, O. Speck, A. Nürnberger, K. H. Maier-Hein, G. Bozdağı Akar, G. Ünal, O. Dicle, and M. A. Selver, "CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation," *Medical Image Analysis*, 2021.

[47] S. G. Armato III, H. Huisman, K. Drukker, L. Hadjiiski, J. S. Kirby, N. Petrick, G. Redmond, M. L. Giger, K. Cha, A. Mamonov, *et al.*, "Prostatex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images," *Journal of Medical Imaging*, 2018.

[48] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, *et al.*, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6695–6714, 2021.

[49] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," in *Brain Lesion: Toward Natural Science for Brain Tumor Segmentation (BrainLes 2021)*, vol. 12962 of *Lecture Notes in Computer Science*, pp. 272–284, 2021.

[50] B. B. Avants, N. Tustison, G. Song, *et al.*, "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[51] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[52] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations*, 2016.