A. Different Configurations of QDM

We present the detailed configurations of our proposed QDM models as well as their number of parameters in Table 6.

Table 6. **Details of QDM models.** Our configurations for Small (S), Base (B) and Large (L) models are based on those of the DiT model [23]. To achieve a comparable parameter count, we halve the number of layers from the original DiT configurations.

Model	Layers N	Hidden size	Heads	#Parameters(M)
QDM-S	6	384	6	50.87
QDM-B	6	768	12	199.06
QDM-L	12	1024	16	691.94

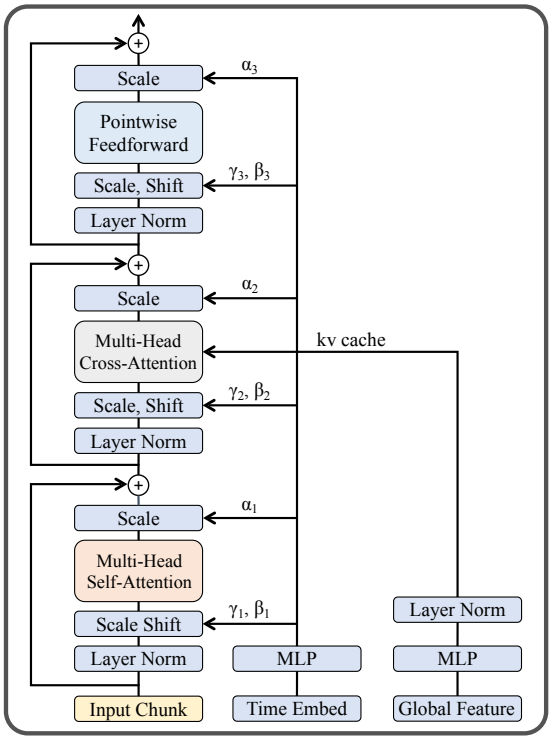

Figure 6. **Architecture of the cross-attention block.** Each block sequentially processes features through: (1) a multi-head self-attention (MSA) layer for local intra-window interactions, (2) a multi-head cross-attention (MCA) layer to fuse global context from upstream features, and (3) a two-layer feedforward network with GELU activation. All outputs from MSA, MCA, and feedforward are combined with residual connections to their respective inputs. Time-step conditioning is applied via adaLN-Zero layers, while the upstream global features are pre-projected into reusable key-value (kv) pairs to minimize redundant computations.

B. Implementation Details

Data Processing. For the real-world SR task, we adopt fixed-size 512×512 image crops during training. Following recent work [42], training images exceeding this resolution undergo either: (1) direct random cropping, or (2) resizing the shorter edge to 512 pixels before random cropping. We synthesize LR/HR pairs using the degradation pipeline from Real-ESRGAN [36].

Beyond real-world test benchmarks, we construct *LSDIR-Test* using center-cropped 512×512 images from the LSDIR test set [17], processed through identical degradation parameters as the training data.

For the medical CT SR task, we use the first degradation stage of Real-ESRGAN pipeline [36] simulating clinical noise artifacts. Both training and test images are first downsampled to 512×512 resolution as HR references, then processed through the adopted pipeline to obtain the corresponding LR image. Thus, we can create two benchmarks: *Med-SR4* ($\times 4$ super-resolution task) and *Med-SR8* ($\times 8$ super-resolution task).

Evaluation Metrics. The performance of different methods was evaluated using both reference-based and non-reference metrics. For reference-based assessment, we employed PSNR, SSIM [38], and LPIPS [46]. To better align with human perceptual judgments in generative super-resolution, we also incorporated non-reference metrics: CLIPIQA [33] and MUSIQ [15]. All LPIPS, CLIP-IQA, and MUSIQ results were computed following the official implementations provided in IQA-PyTorch [4][1].

For the *LSDIR-Test* and *RealSR* datasets, both reference and non-reference metrics were applied to ensure comprehensive evaluation. In contrast, only non-reference metrics were used for *RealSet80* due to the lack of ground truth images. Notably, for real-world super-resolution tasks, PSNR and SSIM were computed in the luminance (Y) channel of the YCbCr color space, while other metrics were calculated directly in RGB space. Evaluations on the *Med-SR4* and *Med-SR8* datasets relied solely on reference metrics, as non-reference perceptual scores were deemed less relevant.

Training Process & Hyperparamters. In real-world SR task, following prior work (LDM [26], ResShift [44]), our architecture operates in latent space using a Vector Quantized GAN (VQGAN, [7]) with a downsampling factor of 4. The model was trained for 150,000 iterations using the dual-stream training objective (Eq. 6), with a batch size of 64 on eight NVIDIA A100 80GB PCIe GPUs. We employed the Adam optimizer with a learning rate of 5×10^{-5}, followed by a 50,000-iteration fine-tuning phase using the perceptual training objective (Eq. 7). For all real-world experiments, we set the quadtree threshold $s = 0.00$.

For medical CT scan super-resolution, we first trained

[1] https://github.com/chaofengc/IQA-PyTorch

a KL-regularized VAE [16, 26] on the medical dataset with a 4× downsampling factor. Our proposed model was then trained in this latent space for 50,000 iterations. To ensure fair comparison, ResShift and LDM were reimplemented and trained under identical conditions (50k iterations), while BSRGAN [45] and SwinIR [18] were trained for 100,000 iterations to account for their distinct optimization requirements. For all medical SR experiments, we set the quadtree threshold $s = 0.15$. For both tasks, the upstream patch size and downstream window size were set to 8. The low-resolution (LR) image was concatenated to the latent feature map before being passed to QDM, with a maximum of 64 chunks processed in parallel. Loss weightings λ_1 and λ_2 were both set to 1, and λ_3 was set to 0.1.

Besides, Figure 6 provides a detailed structure of cross-attention block presented in Figure 2. Each cross-attention block processes features sequentially through three components: a local intra-window self-attention, a cross-attention mechanism that integrates global context from upstream features, and a feedforward network. It's important to note that the upstream global features are pre-projected into reusable key-value (kv) pairs to minimize redundant computations. The MLP ratio is set to the default value of 4.

Ultra-High-Resolution Inference. To handle ultra-high-resolution (UHR) images beyond GPU memory limits, we adopt a patch-based inference strategy with Gaussian-weighted fusion proposed in [43]. The input image $I_{\text{LR}} \in \mathbb{R}^{B\times C\times H\times W}$ is divided into overlapping patches of size $P \times P$ and stride $S \leq P$. Each patch is super-resolved individually to $\hat{p}_k \in \mathbb{R}^{C\times(s_f P)\times(s_f P)}$ and accumulated into a global canvas with corresponding Gaussian weights $w_k(x)$ generated by separable 1D kernels. The final output is obtained by normalized fusion:

$$I_{\text{HR}}(x) = \frac{\sum_k w_k(x)\,\hat{p}_k(x)}{\sum_k w_k(x)}.$$

This Gaussian overlap-add formulation ensures seamless blending between neighboring patches while enabling scalable processing of arbitrarily large images. In practice, we set $P = 128$, $S = 112$, and accumulate in `float64` precision for numerical stability.

C. Additional Experimental Results

Table 7. Downstream-only and prior-based methods on Med-8

Metric	Downstream-only	Prior-based		QDM-L
		DiffBIR	SeeSR	
PSNR↑	31.75	23.62	26.87	**33.05**
SSIM↑	0.9685	0.6553	0.6665	**0.9743**
LPIPS↓	0.0181	0.3686	0.0686	**0.0159**

Table 8. Quantitative comparisons of the proposed QDM-L equipped with different thresholds, on the benchmark *LSDIR-Test*.

Threshold	Density	MACs(T)	Reference Metrics			Non-Reference Metrics	
			PSNR ↑	SSIM ↑	LPIPS ↓	CLIPIQA↑	MUSIQ↑
Full Mask	100.00	18.32	22.16	0.5958	0.2452	0.6444	69.1535
$s = 0.00$	99.41	18.32	22.16	0.5958	0.2452	0.6444	69.1535
$s = 0.03$	90.42	18.19	22.34	0.6059	0.2658	0.5646	68.4705
$s = 0.06$	80.18	17.98	22.50	0.6100	0.2997	0.5016	65.9053
$s = 0.09$	70.37	17.73	22.62	0.6105	0.3340	0.4493	62.3469

Downstream-only Design on *Med-SR8*. We designed the dual-stream architecture to leverage the complementary strengths of both branches. The downstream branch focuses on local, detail-rich regions while the upstream branch captures global context and applies minimum refinement on homogeneous regions. Here, we conducted experiments to validate that without the upstream branch, the model cannot capture the global view and is prone to overlook large-scale structure, leading to degraded performance, as shown in Table 7. More importantly, a downstream-only design lacks the ability to allocate computation adaptively across different regions of the image.

Pretrained-Prior-Based Methods on *Med-SR8*. We've included detailed results on natural images comparing to several prior-based methods in Table 1. To test their generalization abilities, we evaluated two methods (DiffBIR and SeeSR) on the Med-8 benchmark. Neither method yielded satisfactory reconstructions in the medical domain, with notably lower performance compared to QDM (Table 7). These results highlight a substantial generalization gap, suggesting that such methods may require significant domain-specific data and fine-tuning.

Ablation study of quadtree threshold on *LSDIR-Test*. We also conducted an ablation study on the quadtree threshold using real-world benchmarks. However, real-world images often contain significant noise and lack large homogeneous regions, making them less suited for adaptive computation. As shown in Table 4, at s=0 (lossless mode), QDM-L achieves performance identical to the full-mask baseline but exhibits minimal computation reduction, as the mask remains nearly dense. Increasing s to 0.09 reduces computation by 3.22% compared to the baseline, but this comes at a significant performance cost. This suggests that higher quadtree thresholds are not well-suited for super-resolution tasks involving noisy and complex low-resolution inputs.

D. Qualitative comparisons of different methods.

We also include more qualitative comparisons of different methods in Figure 7, Figure 8 and Figure 9.

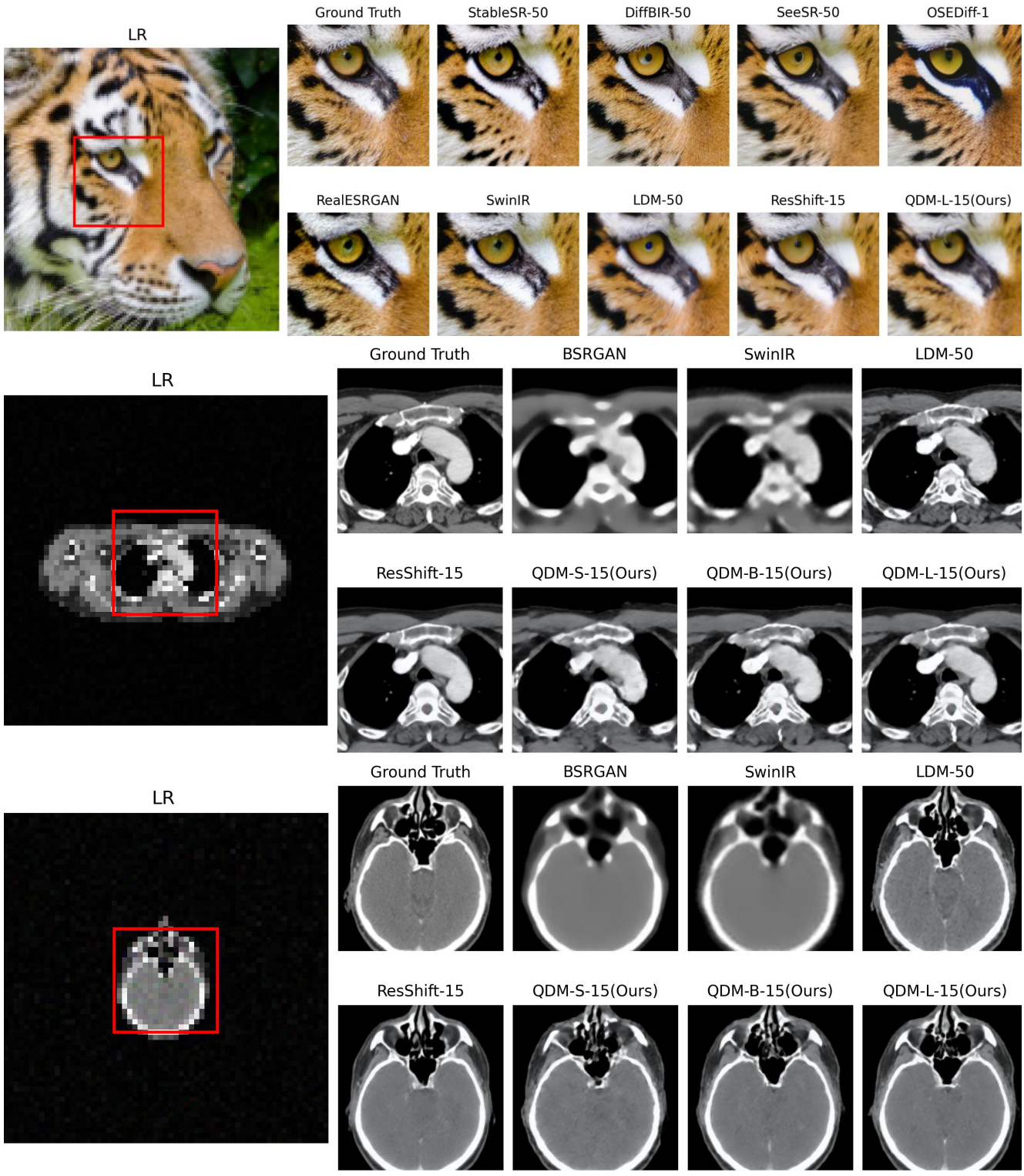

Figure 7. Visual comparison of different methods on real-world images and medical CT datasets. Zoom in for finer details.

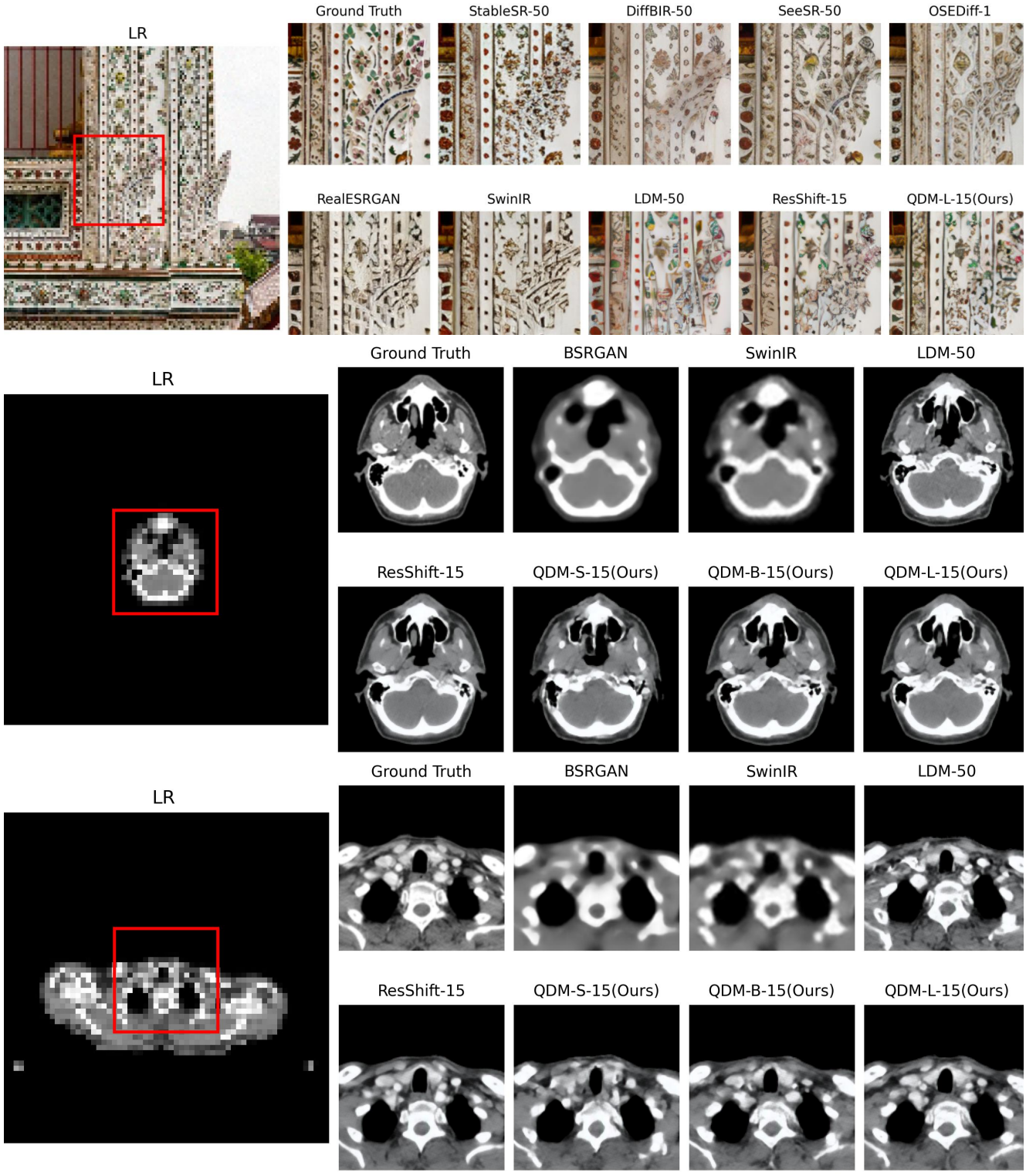

Figure 8. Visual comparison of different methods on real-world images and medical CT datasets. Zoom in for finer details.

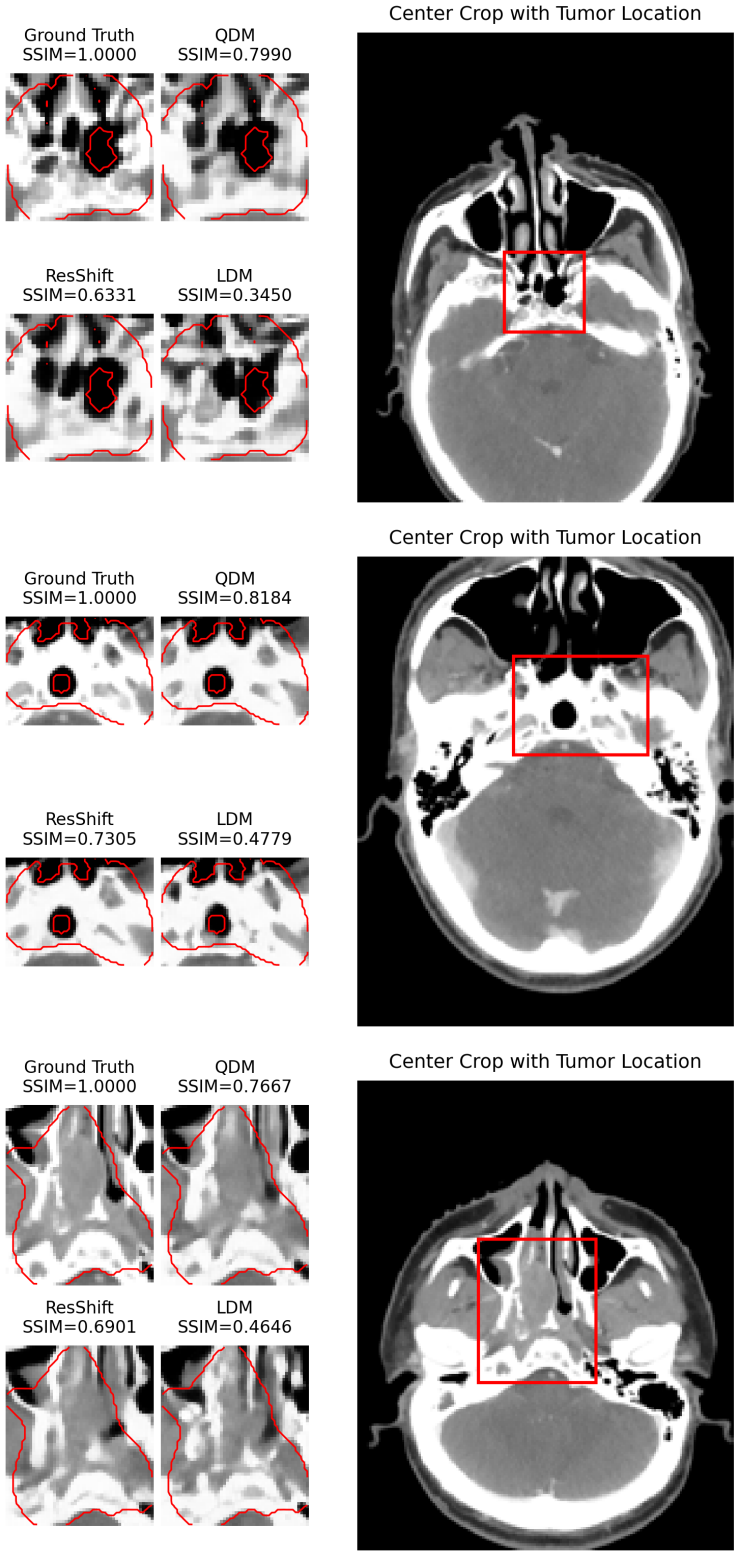

Figure 9. Qualitative comparison of CECTs in the tumor region.