This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

TargetIQA: Anatomy-Aware Ultrasound Image Quality Assessment

George Lansdown[1,2(✉)], Sotirios A. Tsaftaris[2], Alison Q. Smithard[1,2], and Antanas Kascenas[1]

[1] Canon Medical Research Europe, Edinburgh, UK
[2] The University of Edinburgh, Edinburgh, UK
george.lansdown@mre.medical.canon

1 Supplementary Material

1.1 Full Experiment 2 Results

We detail the results of experiment 2 across all organs in Table 1.

Table 1. Experiment 2: Synthetic degradation sensitivity by organ (test set), using **LPIPS** as the reference metric. Per-organ and macro-averaged SRCC between score delta and LPIPS perceptual distance. LPIPS increases with degradation, so more *positive* = stronger sensitivity. 95% cluster bootstrap CI (clusters on original images). **Bold**: best; underline: second best per column.

Method	Heart	Liver	Kidney	Macro (↑)
General (no anatomy)				
BRISQUE	-0.132 $^{+0.168}_{-0.193}$	0.111 $^{+0.203}_{-0.189}$	0.142 $^{+0.118}_{-0.174}$	0.041 $^{+0.092}_{-0.114}$
NIQE	0.085 $^{+0.162}_{-0.107}$	0.418 $^{+0.226}_{-0.143}$	<u>0.360</u> $^{+0.158}_{-0.091}$	0.287 $^{+0.118}_{-0.060}$
CLIP-IQA	0.048 $^{+0.319}_{-0.233}$	0.399 $^{+0.167}_{-0.138}$	0.195 $^{+0.175}_{-0.190}$	0.214 $^{+0.139}_{-0.104}$
TinyUSFM-NRQ (mixture)	-0.043 $^{+0.125}_{-0.138}$	0.001 $^{+0.104}_{-0.112}$	-0.071 $^{+0.131}_{-0.128}$	-0.038 $^{+0.069}_{-0.071}$
Organ-specific				
TinyUSFM-NRQ (matched)	-0.025 $^{+0.107}_{-0.115}$	-0.033 $^{+0.101}_{-0.112}$	-0.106 $^{+0.109}_{-0.108}$	-0.054 $^{+0.058}_{-0.066}$
Qwen3.6-27B	-0.002 $^{+0.204}_{-0.163}$	0.346 $^{+0.134}_{-0.125}$	0.265 $^{+0.111}_{-0.096}$	0.203 $^{+0.094}_{-0.077}$
Ours				
TargetIQA	<u>0.644</u> $^{+0.090}_{-0.083}$	**0.627** $^{+0.054}_{-0.050}$	**0.665** $^{+0.180}_{-0.147}$	**0.645** $^{+0.067}_{-0.061}$
TargetIQA (no-anatomy)	**0.732** $^{+0.049}_{-0.053}$	<u>0.619</u> $^{+0.086}_{-0.085}$	0.327 $^{+0.303}_{-0.237}$	<u>0.559</u> $^{+0.105}_{-0.090}$
TargetIQA (wrong-anatomy)	0.151 $^{+0.060}_{-0.075}$	0.180 $^{+0.072}_{-0.081}$	0.023 $^{+0.030}_{-0.022}$	0.118 $^{+0.031}_{-0.036}$

1.2 Full Experiment 3 Results

We detail the results of experiment 3 across all organs in Table 2.

Table 2. Experiment 3: Paired accuracy by organ (test set). Fraction of pairs correctly ranked higher-quality first. 95% bootstrap CI. Chance level: 50%. **Bold**: best; underline: second best per column.

Method	Breast	Carotid	Kidney	Liver	Thyroid	Macro PA % (↑)
General (no anatomy)						
BRISQUE	**72.1** $^{+9.8}_{-11.5}$	81.4 $^{+8.1}_{-8.1}$	78.4 $^{+13.5}_{-13.5}$	71.1 $^{+13.3}_{-13.3}$	44.8 $^{+10.3}_{-10.3}$	69.6 $^{+5.1}_{-5.5}$
NIQE	44.3 $^{+13.1}_{-13.1}$	**95.3** $^{+3.5}_{-4.6}$	**94.6** $^{+5.4}_{-8.1}$	**97.8** $^{+2.2}_{-4.4}$	85.1 $^{+6.9}_{-8.1}$	83.4 $^{+3.5}_{-3.8}$
CLIP-IQA	68.8 $^{+11.5}_{-11.5}$	61.6 $^{+10.5}_{-10.5}$	86.5 $^{+10.8}_{-10.8}$	91.1 $^{+6.7}_{-8.9}$	36.8 $^{+10.3}_{-10.3}$	69.0 $^{+4.7}_{-4.8}$
TinyUSFM-NRQ (mixture)	34.4 $^{+11.5}_{-13.1}$	30.2 $^{+9.3}_{-9.3}$	62.2 $^{+13.5}_{-16.2}$	68.9 $^{+13.3}_{-13.3}$	27.6 $^{+10.3}_{-9.2}$	44.7 $^{+5.2}_{-5.6}$
Organ-specific						
TinyUSFM-NRQ (matched)	37.7 $^{+11.5}_{-13.1}$	36.0 $^{+10.5}_{-9.3}$	32.4 $^{+16.2}_{-13.5}$	60.0 $^{+13.3}_{-15.6}$	58.6 $^{+10.3}_{-10.3}$	45.0 $^{+5.8}_{-5.7}$
Qwen3.6-27B	62.3 $^{+13.1}_{-13.1}$	22.1 $^{+9.3}_{-8.1}$	21.6 $^{+16.2}_{-10.9}$	35.6 $^{+13.4}_{-13.3}$	52.9 $^{+10.4}_{-10.3}$	38.9 $^{+5.4}_{-5.1}$
Ours						
TargetIQA	57.4 $^{+13.1}_{-13.1}$	87.2 $^{+5.8}_{-7.0}$	94.6 $^{+5.4}_{-8.1}$	95.6 $^{+4.4}_{-6.7}$	**87.4** $^{+6.9}_{-8.1}$	**84.4** $^{+3.6}_{-3.8}$
TargetIQA (no-anatomy)	62.3 $^{+11.5}_{-13.1}$	72.1 $^{+9.3}_{-9.3}$	73.0 $^{+13.5}_{-16.2}$	93.3 $^{+6.7}_{-8.9}$	63.2 $^{+9.2}_{-10.4}$	72.8 $^{+4.9}_{-4.5}$
TargetIQA (wrong-anatomy)	52.5 $^{+13.1}_{-13.1}$	80.2 $^{+8.1}_{-8.1}$	78.4 $^{+10.8}_{-13.5}$	64.4 $^{+13.3}_{-13.3}$	77.0 $^{+8.1}_{-10.3}$	70.5 $^{+5.2}_{-5.3}$

Table 3. Top 10 positive prompt templates by Exp. 3 paired accuracy. `{anatomy}` is replaced with the organ-specific anatomy term. † denotes our chosen template.

ID	Prompt template
P0	A well-penetrating ultrasound of {anatomy} with uniform field brightness
P1	An ultrasound showing {anatomy} with well-defined echotexture
P2	A high-quality ultrasound of {anatomy} with uniform echotexture and good penetration
P3	A well-obtained ultrasound of {anatomy} with clear tissue detail
P4	A high-quality ultrasound of {anatomy} with homogeneous echotexture and good depth
P5†	A well-focused ultrasound showing {anatomy} parenchyma with clear echotexture
P6	An ultrasound of {anatomy} with uniform echogenicity
P7	A well-focused ultrasound of {anatomy} with good gain and clear echotexture
P8	A high-quality ultrasound of {anatomy}
P9	A diagnostically adequate ultrasound of {anatomy}

Table 4. Negative prompt candidates used in the sweep. N6 and N7 are our chosen negatives.

ID	Prompt
N1	A low-quality ultrasound with reduced clarity
N2	An ultrasound with significant image degradation
N3	An ultrasound with poor penetration and loss of detail
N4	An ultrasound with heavy noise and degraded resolution
N5	A poorly optimised ultrasound with reduced diagnostic quality
N6†	A technically inadequate ultrasound
N7†	A poorly optimised ultrasound with significant artifacts

† Denotes our chosen configuration.

Table 5. Best-balanced negative-set configuration per positive prompt (**BiomedCLIP** embeddings), selected by equal-weight percentile rank of E1, E2, and E3 on the validation split. E2 is the Spearman ρ between score delta and **LPIPS** (a perceptual distance, so more positive = stronger degradation sensitivity). **Bold**: best; underline: second best per column. † denotes our chosen template.

ID	Best neg set	E1	E2	E3 %
P0	N6 + N7	0.663	+0.420	**83.0**
P1	N6 + N7	0.790	+0.426	80.5
P2	N6 + N7	0.781	+0.408	78.1
P3	N6 + N7	0.790	+0.454	78.5
P4	N6 + N7	0.765	+0.432	78.4
P5†	N6 + N7	**0.804**	**+0.524**	82.0
P6	N6 + N7	0.718	+0.381	77.0
P7	N6 + N7	0.793	+0.473	77.6
P8	N6 + N7	0.747	+0.440	73.3
P9	N6 + N7	0.772	+0.474	73.9

Table 6. Best-balanced negative-set configuration per positive prompt (**CLIP** embeddings), selected by equal-weight percentile rank of E1, E2, and E3 on the validation split. E2 is the Spearman ρ between score delta and **LPIPS** (a perceptual distance, so more positive = stronger degradation sensitivity). **Bold**: best; underline: second best per column. † denotes our chosen template.

ID	Best neg set	E1	E2	E3 %
P0	N4 + N7	0.581	+0.062	62.7
P1	N7	**0.647**	-0.141	65.6
P2	N4	0.500	-0.096	49.5
P3	N7	0.628	-0.048	**69.3**
P4	N4	0.491	-0.086	50.0
P5†	N6 + N7	0.479	-0.119	58.1
P6	N7	0.578	+0.017	63.0
P7	N4 + N7	0.518	-0.020	51.1
P8	N4 + N7	0.584	+0.019	65.4
P9	N7	0.615	**+0.228**	55.2