Defining Robust Ultrasound Quality Metrics via an Ultrasound Foundation Model

Ziyang Huang[1†], Bingyan Li[1†], Chen Ma[1†], Tianyi Liu[2,3], Yihui Zhai[2,3], Hong Xu[2,3], Yi Guo[1], Zeju Li[1✉], and Yuanyuan Wang[1✉]

[1] College of Biomedical Engineering, Fudan University, Shanghai, China
zejuli@fudan.edu.cn; yywang@fudan.edu.cn
[2] Department of Nephrology, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China
[3] Shanghai Kidney Development and Pediatric Kidney Disease Research Center, Shanghai, China

Abstract. Clinicians lack a principled framework to quantify diagnostic utility in ultrasound reconstructions. Existing standards like PSNR and VGG-LPIPS are inadequate, failing to account for modality-specific physics or the structural nuances of acoustic imaging. We close this gap with a TinyUSFM-based evaluation framework featuring two distinct metrics: TinyUSFM-uLPIPS, a full-reference perceptual distance based on multi-layer token relations, and TinyUSFM-NRQ, a deployable no-reference quality score utilizing clean-manifold modeling and worst-region aggregation to detect localized harmful artifacts. We demonstrate that the presented metrics have four unique advantages: 1) Task-linked quality, where TinyUSFM-uLPIPS achieves superior calibration with semantic task damage, accurately reflecting Dice-score drops in segmentation where VGG-based metrics fail; 2) Cross-organ comparability, maintaining stable scoring scales and consistent severity rankings across diverse anatomical sites and domain-shifted data; 3) PSNR-consistent sensitivity, with TinyUSFM-NRQ providing a reliable quality score without ground-truth images that remains consistent with traditional fidelity benchmarks (i.e. PSNR); and 4) Clinical utility, improving the prediction of expert preference from 47.2% to 72.8% accuracy and producing super-resolution reconstructions preferred by sonographers. By integrating these advantages into a unified assessment and optimization loop, this work establishes a modality-aligned standard that finally bridges the gap between algorithmic performance and diagnostic utility. Our code is available at https://github.com/sextant-fable/US-Metrics.

Keywords: Ultrasound, Image Quality Assessment, Perceptual Metric

1 Introduction

Ultrasound is ubiquitous in clinical imaging due to its portability and safety. However, evaluating the perceptual and clinical usefulness of ultrasound re-

† Equal contribution; ✉ corresponding author.

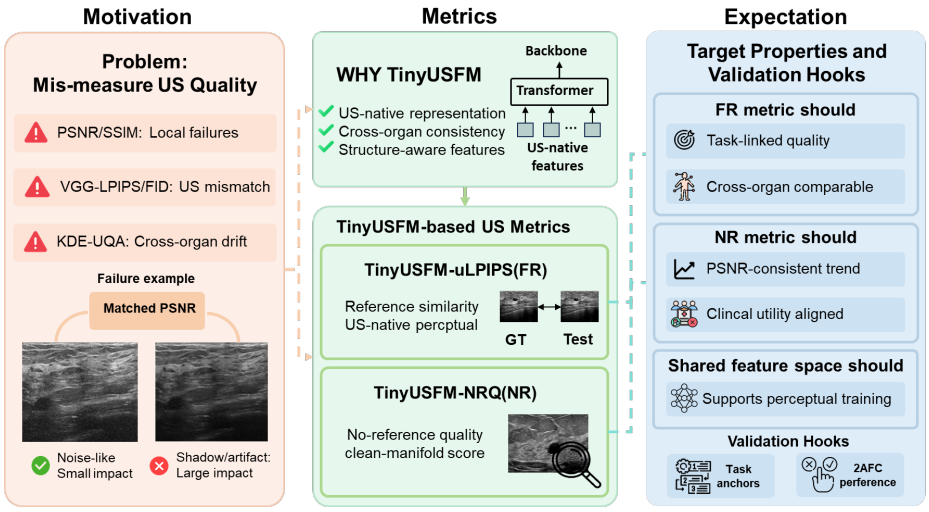

Fig. 1. Overview of the TinyUSFM-based ultrasound quality evaluation framework, including metric motivation (left), proposed full-reference (FR) and no-reference (NR) metrics (middle), and target properties with shared-feature downstream utility (right).

constructions and generative outputs remains poorly served by commonly used image-quality metrics, especially across organs, devices, and acquisition settings.

Ultrasound appearance is shaped by modality-specific physics and sampling. Coherent interference produces speckle, specular responses are view dependent, attenuation and shadowing create structured dropouts, and scan conversion introduces anisotropic resolution and sampling artifacts [3,2,18]. Clinically meaningful failures are often localized: a small region of clutter or shadowing may erase a critical boundary even when global intensity changes slightly. In contrast, adjustments to global amplification or depth-dependent amplification alter pixel values across the image without degrading its diagnostic interpretability.

Existing metrics are poorly calibrated for ultrasound due to three primary limitations: 1) Fidelity vs. Clinical Utility: Pixel-level measures (PSNR and SSIM [20]) prioritize global intensity agreement, making them overly sensitive to benign brightness shifts while failing to detect localized, clinically harmful artifacts. 2) Domain Mismatch: Natural-image perceptual metrics (VGG-LPIPS and FID [24,6]) rely on photographic priors that misinterpret ultrasound-specific textures as noise and under-penalize structured occlusions. 3) Lack of Generalization: While generic no-reference IQA metrics such as BRISQUE [14] and NIQE [15], and ultrasound-specific learned IQA/QA methods including fetal ultrasound IQA [21,23], expert-agnostic ultrasound IQA [17], and KDE-based QA [8], have been proposed, most existing approaches are tied to restricted clinical protocols, organ-specific assumptions, or task-specific labels. In contrast, our goal is to define both FR and NR metrics in a shared ultrasound foundation-

model feature space and evaluate whether this space supports task-linked, cross-organ, and clinically aligned quality assessment.

As illustrated in Fig. 1, we turn to USFM [7] and its lightweight variant, TinyUSFM [11] to develop a principled, ultrasound-native evaluation framework that bridges the gap between image-quality metrics and diagnostic utility. By leveraging the foundation model's feature space, we introduce two metrics—the full-reference (FR) TinyUSFM-uLPIPS and the no-reference (NR) TinyUSFM-NRQ—which offer superior calibration with clinical tasks and stable performance across diverse anatomical sites. We validate both metrics using task-based semantic anchors and clinician two-alternative forced choice (2AFC) studies. Furthermore, we employ the same feature space as a perceptual loss for super-resolution, establishing a validated pipeline for clinical image enhancement.

2 Methodology

2.1 Ultrasound-native representations from TinyUSFM

For an input image x, let $f_\ell(x) \in \mathbb{R}^{(1+T_\ell)\times C_\ell}$ denote the TinyUSFM output at layer ℓ, where T_ℓ is the number of spatial patch tokens and C_ℓ is the channel dimension. After removing the global [CLS] token, we denote the remaining patch-token matrix by $F_\ell(x) \in \mathbb{R}^{T_\ell \times C_\ell}$. For TinyUSFM-NRQ, each layer is further summarized by average pooling its patch tokens to obtain $u_\ell(x) \in \mathbb{R}^{C_\ell}$, and the pooled features from layers $\mathcal{L}$ are concatenated and L_2-normalized to form a global descriptor $z(x)$. Thus, $F_\ell(x)$ is used for full-reference comparison, whereas $z(x)$ is used for no-reference quality modeling.

2.2 TinyUSFM-uLPIPS: a full-reference quality metric

Given a reference image x and a test image y, TinyUSFM-uLPIPS compares their TinyUSFM patch-token features across layers $\mathcal{L} = \{3, 5, 7, 11\}$, where $F_\ell(x), F_\ell(y) \in \mathbb{R}^{T_\ell \times C_\ell}$ denote the [CLS]-removed patch-token matrices at layer ℓ. In all experiments, we use $\mathcal{L} = \{3, 5, 7, 11\}$, $r = 3$, and $\tau = 20$.

Let $\hat{F}_\ell(x)$ and $\hat{F}_\ell(y)$ denote the row-wise L_2-normalized versions of $F_\ell(x)$ and $F_\ell(y)$. For each token location $q \in \{1, \ldots, T_\ell\}$, let $\mathcal{N}_r(q)$ denote its local neighborhood with radius r, and let $A_{\ell,q}(x) \in \mathbb{R}^{|\mathcal{N}_r(q)|\times C_\ell}$ be the matrix formed by stacking the normalized token vectors in that neighborhood. We then define

$$d^{\mathrm{s}}_\ell(x, y) = \frac{1}{T_\ell} \sum_{q=1}^{T_\ell} \mathrm{MeanAbs}\big(\mathrm{Softmax}\big(\tau A_{\ell,q}(x) A_{\ell,q}(x)^\top\big) - \mathrm{Softmax}\big(\tau A_{\ell,q}(y) A_{\ell,q}(y)^\top\big)\big)\,, \tag{1}$$

where Softmax$(\cdot)$ is applied row-wise and τ is a temperature parameter.

Channel co-activation statistics are compared using Gram matrices computed from the normalized token matrices [5], with $G_\ell(x) = \frac{1}{T_\ell C_\ell} \hat{F}_\ell(x)^\top \hat{F}_\ell(x)$

and $d_\ell^{\mathrm{g}}(x,y) = \mathrm{MeanAbs}(G_\ell(x) - G_\ell(y))$. The per-layer distance is $d_\ell(x,y) = d_\ell^{\mathrm{s}}(x,y) + d_\ell^{\mathrm{g}}(x,y)$, and the final TinyUSFM-uLPIPS score is

$$\text{TinyUSFM-uLPIPS}(x,y) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} d_\ell(x,y). \tag{2}$$

2.3 TinyUSFM-NRQ: a no-reference quality metric

TinyUSFM-NRQ scores a single ultrasound image without a reference by combining organ-aware clean-manifold likelihood with worst-region aggregation. This design models valid appearance variation across devices and settings while remaining sensitive to localized failures.

For an image x, we extract overlapping patches $\{x_i\}_{i=1}^{N_x}$ of size 224×224 with stride 112, where N_x is the number of extracted patches. Each patch x_i is encoded as $\hat{z}_i = z(x_i)$ using the global TinyUSFM descriptor from Sec. 2.1. For each organ o, clean patch descriptors are projected to a PCA space of dimension $d = 128$ and modeled by a diagonal-covariance Gaussian mixture model p_o with $K = 4$ components. If the organ label is available, patch x_i is scored by $s_i(x) = \log p_o(\hat{z}_i)$; otherwise, we use a uniform mixture over all O organ models, i.e., $s_i(x) = \log\Big(\frac{1}{O}\sum_{o=1}^{O} p_o(\hat{z}_i)\Big)$. To emphasize localized failures, we average the lowest-scoring patches, where $\kappa = \max(1, \mathrm{round}(0.15\, N_x))$ and $\mathcal{W}_\kappa(x)$ denotes the index set of the κ worst patches. We then define

$$\text{TinyUSFM-NRQ}(x) = \frac{1}{\kappa} \sum_{i \in \mathcal{W}_\kappa(x)} s_i(x), \tag{3}$$

where higher values indicate better quality.

2.4 TinyUSFM features as a perceptual loss for image restoration

For SR optimization, we use a differentiable token-distance loss in the same TinyUSFM feature space: Let $\hat{F}_\ell(\cdot)$ denote the row-wise L_2-normalized [CLS]-removed token matrix at layer ℓ. Given a reconstruction $\hat{y}$ and its target image y, we define

$$\mathcal{L}_{\text{TinyPerc}}(\hat{y}, y) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \frac{1}{T_\ell} \sum_{t=1}^{T_\ell} \Big\| \hat{F}_{\ell,t}(\hat{y}) - \hat{F}_{\ell,t}(y) \Big\|_2^2. \tag{4}$$

For 512×512 images, the TinyUSFM perceptual loss is computed on sliding 224×224 windows and averaged across windows to avoid resizing artifacts.

3 Experiments and Results

3.1 Evaluation Protocol

Task-Linked Quality (FR): We assess metric ability to capture task-linked quality by using performance degradation of downstream models as an objective

proxy for clinical utility. We systematically apply distortions using a binary search to ensure each reaches an identical PSNR-matched magnitude relative to the clean baseline. By evaluating a pre-trained model across these controlled degradations, we use the resulting drop in task performance (e.g., segmentation or classification accuracy) as a "ground truth" for how much a specific distortion actually interferes with diagnostic information. The rank correlation (Spearman's ρ and Kendall's τ) between TinyUSFM-uLPIPS and these accuracy drops serves as a primary indicator of whether the metric prioritizes task-relevant structural integrity over benign global fluctuations like intensity shifts.

Cross-Organ Comparability (FR): To evaluate cross-organ comparability, we assess the consistency of degradation rankings derived by TinyUSFM-uLPIPS across diverse anatomical sites using Kendall's W, ensuring that the relative severity of distortions is penalized uniformly regardless of the organ. We further analyze score-scale stability using the interquartile range (IQR), which measures the statistical spread of the middle 50% of the data. A lower IQR indicates that the metric produces more consistent and predictable scores for a given quality level, suggesting that the numerical scale remains stable and is not skewed by anatomical variations.

PSNR-Consistent Sensitivity (NR): To evaluate our NR metric for TinyUSFM-NRQ, we utilize a ground-truth ranking protocol. We synthesize six variants of each clean image by applying distortions of known, increasing severity. Since the true quality hierarchy is established by design, we assess the metric's effectiveness by measuring how closely its predicted scores correlate with these pre-defined ranks. High correlation indicates that the framework can accurately perceive diagnostic degradation without requiring a pristine reference image.

Clinician Utility (NR): To validate the metric's practical utility, we conducted a blinded 2AFC study. Clinicians evaluated PSNR-matched image pairs featuring different degradation types and selected the more clinically acceptable version in each case. These expert preferences were then compared against TinyUSFM-NRQ scores across 540 cross-degradation pairs to determine if the metric aligns with professional diagnostic standards when traditional pixel-fidelity measures fail to distinguish quality.

3.2 Full-Reference Evaluation (TinyUSFM-uLPIPS)

Experimental Setup: We use thyroid and kidney ultrasound images for Dice-based segmentation evaluation, including thyroid nodule segmentation from DDTI [16] and kidney capsule segmentation from the Open Kidney Ultrasound Data Set [19]. We degrade the images to matched PSNR levels (20–25) and run the same segmentation model to obtain Dice scores. We then compare VGG-LPIPS and TinyUSFM-uLPIPS on the degraded images and test which metric better reflects the true severity of task damage.

The evaluation covers 8 distortions categorized into four primary groups: noise and texture corruption, blur and resolution loss, ultrasound-specific artifacts (such as shadowing, specular clipping, and missing scanlines), and geometric deformation.

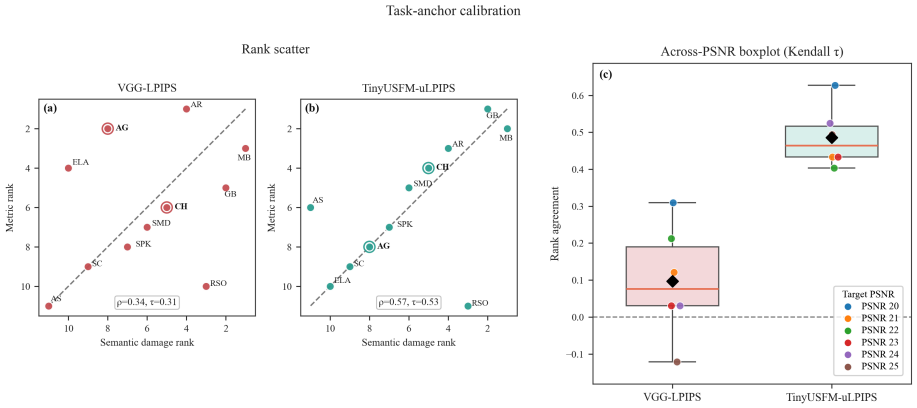

Fig. 2. Task-anchor calibration of FR metrics under PSNR-aligned evaluation. Left: representative DDTI rank–rank scatter at PSNR= 20 for VGG-LPIPS and TinyUSFM-uLPIPS. Right: Kendall's τ summarized across segmentation anchors and PSNR targets; colored dots indicate individual anchor–PSNR settings.

Task-Linked Quality: Fig. 2 shows a representative DDTI rank-scatter example and Kendall's τ summaries across segmentation anchors and PSNR targets. TinyUSFM-uLPIPS aligns better with semantic damage than VGG-LPIPS on the thyroid and kidney ultrasound datasets, and it shows consistently higher and more stable Kendall's τ across PSNR targets (Fig. 2). At low PSNR, the difference is most clear because distortion effects are easier to separate; at higher PSNR, rank comparison becomes harder because Dice drops are smaller, but TinyUSFM-uLPIPS remains more stable than VGG-LPIPS.

Under matched PSNR, VGG-LPIPS shows a systematic bias: it tends to over-penalize noise-like degradations and under-penalize ultrasound-specific structural artifacts. For example, at PSNR= 20 on the thyroid dataset, VGG-LPIPS ranks AdditiveGaussian as more severe than ClutterHaze/ROIShadow Occlusion, even though the latter causes larger Dice drop, while TinyUSFM-uLPIPS gives the correct ordering.

Cross-Organ Comparability: TinyUSFM-uLPIPS produces more consistent cross-organ degradation rankings than VGG-LPIPS (Fig. 3(a)) and shows a smaller inter-organ IQR on several representative degradations, such as speckle-related and resolution-related corruptions (Fig. 3(b)).

3.3 No-Reference Evaluation (TinyUSFM-NRQ)

Experimental Setup. We evaluate TinyUSFM-NRQ on public ultrasound datasets covering multiple organs and acquisition domains, including CUBS [13], UF1990 [4], STMUS [12], AUL [22], BUSI [1], MMOTU [25], DDTI [16], KidneyUS [19] and CAMUS [9]. Robustness analysis for TinyUSFM-NRQ and baselines is conducted on 9 organs following the shared protocol in Sec. 3.1, yielding

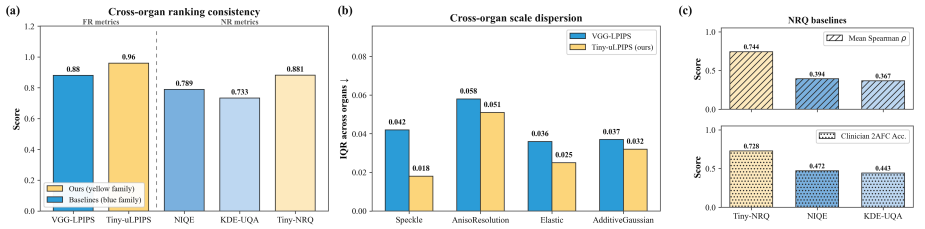

Fig. 3. Cross-organ robustness and NRQ baseline comparison under PSNR-aligned evaluation. (a) Cross-organ ranking consistency for FR and NRQ metrics (Kendall's W, PSNR 20–25 dB). (b) Cross-organ scale dispersion for FR metrics (IQR across organs) on representative degradations (aggregated over PSNR 20–25 dB). (c) NRQ baseline comparison: within-organ correlation (Spearman ρ) and clinician 2AFC agreement.

(9×8) organ×distortion conditions. We compare TinyUSFM-NRQ with NIQE, BRISQUE, and KDE-UQA (Fig. 3(c)).

PSNR-Consistent Sensitivity. TinyUSFM-NRQ achieves a mean Spearman $\rho = 0.744$ over PSNR levels, indicating strong within-organ monotonicity. NIQE and KDE-UQA fail systematically on ultrasound-typical structural degradations, especially scanline missing and specular clipping. At PSNR= 25, these metrics become numerically ill-conditioned for structural artifacts, and several distortion families demonstrate near-zero or negative correlations with clinical quality across different organs. This makes them unreliable as a unified cross-organ quality control scale under heterogeneous ultrasound artifacts and is consistent with their lower monotonicity and weaker clinician alignment relative to TinyUSFM-NRQ (Fig. 3(c)).

Using the same PSNR-aligned degradation settings on unseen organs (stomach, multifidus muscle, and neck nerve), TinyUSFM-NRQ maintains strong monotonicity and reliable severity ranking under domain shift. Its degradation scoring trends are consistent with the in-distribution results, showing similar relative severity patterns across major artifact families.

Clinician Utility. TinyUSFM-NRQ predicts clinician preference with 72.8% accuracy (95% CI [0.689, 0.764], Wilson), significantly above chance (two-sided binomial test, $p<10^{-20}$). In embedded sanity checks (same degradation, different PSNR) and duplicate pairs, the readers achieved 100% accuracy and consistency.

3.4 Ultrasound Super-Resolution

Experimental Setup. We train the same SwinIR backbone [10] with identical splits under three objectives: $\mathcal{L}_1$, $\mathcal{L}_1 + \mathcal{L}_{\text{VGG}}$, and $\mathcal{L}_1 + \mathcal{L}_{\text{TinyUSFM}}$ (Sec. 2.4). Training uses a combined breast+echocardiography ultrasound dataset for 100 epochs (LR 10^{-4}) with $\lambda_{\text{percep}}/\lambda_1 = 0.1/1$ for both perceptual losses.

Perceptual-Clinical Alignment Loop. Compared with $\mathcal{L}_{\text{VGG}}$, using $\mathcal{L}_{\text{TinyUSFM}}$ as the perceptual loss improves PSNR and SSIM by 5.3% and 3.3%, respectively.

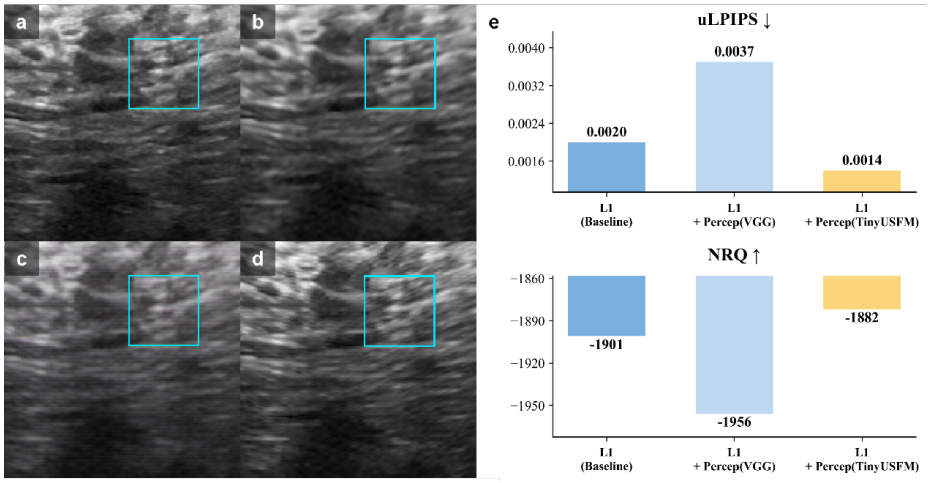

Fig. 4. SR comparison under three training objectives. (a–d) Representative breast ultrasound example (ground truth; $\mathcal{L}_1$; $\mathcal{L}_1 + \mathcal{L}_{\text{VGG}}$; $\mathcal{L}_1 + \mathcal{L}_{\text{TinyUSFM}}$). (e) Quantitative comparison using TinyUSFM-uLPIPS and TinyUSFM-NRQ.

Although $\mathcal{L}_1$ still gives the strongest PSNR, it produces over-smoothed reconstructions, while $\mathcal{L}_{\text{VGG}}$ reduces fidelity and introduces ultrasound-inconsistent artifacts. In contrast, $\mathcal{L}_{\text{TinyUSFM}}$ maintains competitive fidelity and better preserves diagnostically relevant acoustic texture with fewer conspicuous structural hallucinations. As shown in Fig. 4, the ultrasound-native perceptual metrics (TinyUSFM-uLPIPS and TinyUSFM-NRQ) also favor $\mathcal{L}_{\text{TinyUSFM}}$ reconstructions. These findings validate the use of ultrasound-native features for both model optimization and evaluation. By prioritizing acoustic-specific structures over pixel-level fidelity, the metric aligns more closely with clinical utility.

In a blinded clinician pairwise evaluation, expert sonographers compared reconstructions from the same case and selected the one more suitable for clinical diagnosis. Clinicians preferred $\mathcal{L}_{\text{TinyUSFM}}$ reconstructions over $\mathcal{L}_1$ in 87.5% of cases and over $\mathcal{L}_{\text{VGG}}$ in 100% of cases. Moreover, $\mathcal{L}_1$ was preferred over $\mathcal{L}_{\text{VGG}}$ in 97.5% of cases, indicating that visually sharper but hallucinated structures are clinically unacceptable in ultrasound. This is consistent with TinyUSFM-NRQ, which favors $\mathcal{L}_{\text{TinyUSFM}}$ reconstructions over the $\mathcal{L}_1$ and $\mathcal{L}_{\text{VGG}}$ baselines.

Runtime analysis confirms a functional trade-off between the two modes. TinyUSFM-uLPIPS is optimized for high-fidelity offline research evaluation, whereas TinyUSFM-NRQ provides significantly lower latency and higher throughput. This efficiency makes the NRQ variant well-suited for real-time, no-reference quality control within active clinical and system pipelines.

4 Conclusion

We introduced a TinyUSFM-based ultrasound quality evaluation framework with two ultrasound-native metrics: TinyUSFM-uLPIPS for FR perceptual assessment and TinyUSFM-NRQ for NR quality scoring. Using a shared TinyUSFM feature space, the framework provides a modality-aligned alternative to pixel-fidelity metrics, natural-image perceptual metrics, and generic quality control baselines for ultrasound reconstruction and generative imaging. Future work will explore extending the framework to other ultrasound foundation models, higher-PSNR quality regimes, dynamic ultrasound, and diverse medical imaging modalities. The proposed metrics inherit the coverage and potential biases of the TinyUSFM backbone, and their reliability may be affected by rare pathologies, unseen devices, or under-represented acquisition settings. Overall, we believe this work provides the ultrasound community with a concrete answer to the question clinicians have always implicitly asked: Does this reconstruction show what I need to see?

Acknowledgements This work was supported in part by the National Key R&D Program of China under Grant 2024YFF0507300 and Grant 2024YFF0507303, in part by the National Natural Science Foundation of China under Grant 62531004, in part by the Fudan University Medicine-Engineering Collaboration Program under Grant yg-2025-general-07, and in part by the Shanghai Municipal Education Commission Program on AI for Transforming Academic Research Paradigms (24RGZNB05).

Disclosure of Interests The authors have no competing interests to declare.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863
2. Baad, M., Lu, Z.F., Reiser, I., Paushter, D.: Clinical significance of us artifacts. RadioGraphics **37**(5), 1408–1423 (2017). https://doi.org/10.1148/rg.2017160175
3. Bamber, J.C., Daft, C.: Adaptive filtering for reduction of speckle in ultrasonic pulse-echo images. Ultrasonics **24**(1), 41–44 (1986). https://doi.org/10.1016/0041-624X(86)90072-7
4. Cai, P., Yang, T., Xie, Q., Liu, P., Li, P.: A lightweight hybrid model for the automatic recognition of uterine fibroid ultrasound images based on deep learning. Journal of Clinical Ultrasound **52**(6), 753–762 (2024). https://doi.org/10.1002/jcu.23703
5. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015), https://arxiv.org/abs/1508.06576

6. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS) (2017), https://arxiv.org/abs/1706.08500
7. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., Guo, Y.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024). https://doi.org/10.1016/j.media.2024.103202
8. Kwon, J., Jiao, J., Self, A., Noble, J.A., Papageorghiou, A.: A kernel density estimation based quality metric for quality assessment of obstetric ultrasound video. In: Trustworthy Machine Learning for Healthcare (TML4H 2023). Lecture Notes in Computer Science, vol. 13932, pp. 134–146. Springer (2023). https://doi.org/10.1007/978-3-031-39539-0_12
9. Leclerc, S., Smistad, E., Pedrosa, J., Östvik, A., Grenier, T., Espinosa, F., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019). https://doi.org/10.1109/TMI.2019.2900516
10. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) (2021), https://arxiv.org/abs/2108.10257
11. Ma, C., Jiao, J., Liang, S., Fu, J., Wang, Q., Li, Z., Wang, Y., Guo, Y.: Tinyusfm: Towards compact and efficient ultrasound foundation models. arXiv preprint arXiv:2510.19239 (2025). https://doi.org/10.48550/arXiv.2510.19239, https://arxiv.org/abs/2510.19239
12. Marzola, F., van Alfen, N., Doorduin, J., Meiburger, K.M.: Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. Computers in Biology and Medicine **135**, 104623 (2021). https://doi.org/10.1016/j.compbiomed.2021.104623
13. Meiburger, K.M., Marzola, F., Zahnd, G., Faita, F., Loizou, C.P., Lainé, N., et al.: Carotid ultrasound boundary study (cubs): Technical considerations on an open multi-center analysis of computerized measurement systems for intima-media thickness measurement on common carotid artery longitudinal b-mode ultrasound scans. Computers in Biology and Medicine **144**, 105333 (2022). https://doi.org/10.1016/j.compbiomed.2022.105333
14. Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing **21**(12), 4695–4708 (2012). https://doi.org/10.1109/TIP.2012.2214050
15. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters **20**(3), 209–212 (2013), https://live.ece.utexas.edu/research/Quality/niqe_spl.pdf
16. Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., Romero, E.: An open access thyroid ultrasound image database. In: Medical Imaging 2015: Computer-Aided Diagnosis. Proceedings of SPIE, vol. 9287 (2015). https://doi.org/10.1117/12.2073532
17. Raina, D., Ntentia, D., Chandrashekhara, S.H., Voyles, R., Saha, S.K.: Expert-agnostic ultrasound image quality assessment using deep variational clustering. arXiv preprint arXiv:2307.02462 (2023), https://arxiv.org/abs/2307.02462
18. Robinson, D.E., Knight, P.C.: Interpolation scan conversion in pulse-echo ultrasound. Ultrasonic Imaging **4**(4), 297–310 (1982). https://doi.org/10.1177/016173468200400401

19. Singla, R., Ringstrom, C., Hu, G., Lessoway, V., Reid, J., Nguan, C., Rohling, R.: The open kidney ultrasound data set. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 (2023). https://doi.org/10.1007/978-3-031-44521-7_15
20. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
21. Wu, L., Cheng, J.Z., Li, S., Lei, B., Wang, T., Ni, D.: Fuiqa: Fetal ultrasound image quality assessment with deep convolutional networks. IEEE Transactions on Cybernetics **47**(5), 1336–1349 (2017). https://doi.org/10.1109/TCYB.2017.2671898
22. Xu, Y., Zheng, B., Liu, X., Wu, T., Ju, J., Wang, S., Lian, Y., Zhang, H., Liang, T., Sang, Y., Jiang, R., Wang, G., Ren, J., Chen, T.: Improving artificial intelligence pipeline for liver malignancy diagnosis using ultrasound images and video frames. Briefings in Bioinformatics **24**(1), bbac569 (2023). https://doi.org/10.1093/bib/bbac569
23. Zhang, B., Liu, H., Luo, H., Li, K.: Automatic quality assessment for 2d fetal sonographic standard plane based on multitask learning. Medicine **100**(4), e24427 (2021)
24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
25. Zhao, Q., Lyu, S., Bai, W., Cai, L., Liu, B., Cheng, G., Wu, M., Sang, X., Yang, M., Chen, L.: Mmotu: A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. arXiv preprint arXiv:2207.06799 (2022), https://arxiv.org/abs/2207.06799