Paper Review: *Exploring the feasibility of zero-shot super-resolution in preclinical imaging*

Amine Maazizi, Nemanja Vujadinovic

Abstract

This document presents our review and critical assessment of the paper. In the annexe, we also include short notes on concepts from the paper that were initially unfamiliar to us, together with a simplified reproduction of part of the experimental setting, including our hypotheses, assumptions, and results. Additional code and outputs are available in the accompanying repository: https://github.com/amine-maazizi/bida-sr

1. Paper Summary

1.1 Problem and Motivation

The paper addresses through-plane super-resolution for preclinical MRI, where scans often have good in-plane resolution but poor slice resolution, leading to anisotropic volumes that are difficult to use for volumetric analysis. This matters for downstream tasks such as skull stripping, registration, and quantitative analysis. The authors also motivate the problem by noting that supervised super-resolution is difficult in preclinical imaging because acquisition protocols vary widely, making paired training data scarce and less transferable across studies. To address this, the paper proposes **BiDA**, a zero-shot super-resolution method for anisotropic preclinical MRI based on biplanar diffusion reconstruction.

1.2 Position with Respect to the State of the Art

The paper lies at the intersection of MRI super-resolution, inverse problems, and diffusion-based restoration. Its core method relies on diffusion-based generative modeling, which is the most natural method from the course for this problem: diffusion can serve as a degradation-aware prior for super-resolution and inverse reconstruction, is closely related to score-based formulations, and can be contrasted with other course approaches such as GAN-based SR. Beyond generative models, the paper connects to other course topics in MRI reconstruction, validation, and downstream segmentation analysis. In particular, the paper raises issues of generalization, domain shift, and leakage-safe evaluation, and its downstream skull-stripping experiment relates naturally to overlap-based segmentation metrics.

BiDA is positioned against three main families of approaches. Interpolation methods, such as B-spline resampling, are simple baselines, but cannot recover missing high-frequency information [11]. Supervised MRI super-resolution methods can perform well, but they require paired low-resolution (LR) and high-resolution (HR) training data and are sensitive to domain shift [2, 1]. Self-supervised methods, such as SMORE and ECLARE, avoid external paired datasets, but remain largely slice-based and may suffer from limited 3D consistency [5, 13, 6, 7, 8, 9]. More broadly, the paper builds on diffusion-based reconstruction frameworks using DDPM priors, DDIM sampling, and DDNM-style data-consistent inference [4, 10, 12, 3].

1.3 Main Contributions

The paper's main contributions are the introduction of BiDA, a biplanar zero-shot reconstruction strategy for anisotropic preclinical MRI, the adaptation of DDNM-based inference with DDPM priors to perform super-resolution without paired HR/LR data, and a relatively broad evaluation across multiple scale factors, in-domain and out-of-domain rat data, and cross-species mouse data using both reconstruction and downstream metrics.

1.4 Methodology

The paper formulates reconstruction as recovering an HR image x from an LR observation y under the forward model $y = Ax$, where A represents slice-profile blur and downsampling; see Annexe C for a short explanation of the slice-selection profile modeling used in this degradation setup. Two DDPMs are trained on HR rat MRI slices, one axial and one sagittal, to model realistic HR anatomy in each plane.

At inference time, the anisotropic volume is reconstructed slice by slice using DDNM. The idea is to preserve the component constrained by the observation model while using the diffusion prior to estimate the missing null-space detail. DDIM is used only to accelerate sampling. The final BiDA output is obtained by reconstructing in both axial and sagittal directions and averaging the two volumes, with the explicit goal of reducing the slice inconsistency that arises in standard 2D slice-wise reconstruction.

1.5 Experimental Setup and Results

The experiments use rodent MRI datasets, with rat data for training and evaluation and held-out rat and mouse data for OOD testing. LR inputs are simulated from isotropic HR volumes, and BiDA is compared with cubic interpolation and ECLARE, with paired method comparisons assessed using a two-sided Wilcoxon signed-rank test (see Annexe A). The results show a clear tradeoff: BiDA produces sharper and more plausible through-plane detail and performs better on the downstream skull-stripping metric, while ECLARE often achieves higher PSNR, consistent with the perception-distortion tradeoff briefly summarized in Annexe B. BiDA generalizes reasonably to held-out rat data, but performance remains weaker in the cross-species mouse setting.

2. Critical Assessment

2.1 Strengths

A major strength of the paper is that it addresses a relevant practical problem in preclinical imaging rather than only a methodological benchmark. The introduction clearly explains why anisotropic MRI limits volumetric analysis, and the proposed method is well aligned with that need. The method is also conceptually clear. The biplanar design is simple, but the paper justifies it well: 2D methods are easier to train, yet they often create slice inconsistency in 3D volumes, so averaging two orthogonal reconstructions is a reasonable compromise between feasibility and coherence. The method is further grounded in established diffusion frameworks, and the roles of DDPM, DDNM, and DDIM are described clearly. Another strength is the evaluation design. The paper goes beyond in-domain testing by including withheld rat data and a cross-species mouse setting, which makes the generalization analysis more meaningful. It also does not rely only on PSNR. By including cDSC from a downstream skull-stripping pipeline, the authors show that improved anatomical structure can matter even when distortion metrics are not optimal.

2.2 Weaknesses and Limitations

The main weakness of the paper is the lack of direct empirical justification for its specific proposed design. While the paper provides a theoretical motivation, arguing that two orthogonal 2D DDPM priors within DDNM and their fusion should reduce slice inconsistency while respecting the degradation model, it does not convincingly show that this combination is what drives the gains. BiDA mostly combines existing components, and this would be less problematic if the experiments established more clearly why this design is preferable to other diffusion-based inverse-problem methods. In particular, DiffusionMBIR is cited as closely related work but is not included in the comparison, so it remains unclear whether BiDA improves over the nearest diffusion-based alternative or mainly over interpolation and ECLARE. This issue is compounded by the lack of ablation: since the main claim is that combining axial and sagittal reconstructions improves volumetric consistency, the paper should report axial-only, sagittal-only, and fused results. The cDSC metric, based on the same skull-stripping pipeline applied to both HR and SR images with the HR-derived mask treated as a silver standard, is a useful consistency proxy but not a direct measure of anatomical fidelity or segmentation accuracy. In addition, only one downstream task is evaluated, LR inputs are simulated from HR volumes under an assumed slice profile with no slice gap, and the weaker mouse performance, which the authors themselves acknowledge, is not analyzed in enough detail to determine whether the failure mainly reflects anatomy, scale, field-of-view mismatch, or another acquisition-related factor (see Annexe D on why rat-to-mouse transfer is nontrivial in this setting).

2.3 Recommendations for Improvement

The paper would be stronger with a more focused experimental validation of its main design choice. Most importantly, it should include an ablation of axial-only, sagittal-only, and fused reconstruction to directly demonstrate the value of biplanar averaging. It should also compare BiDA against stronger diffusion-based baselines, especially DiffusionMBIR, to clarify whether the method improves over the closest related diffusion framework or mainly over interpolation and ECLARE. The validation would further benefit from clearer reporting of variability, including subject-level summaries, confidence intervals, and a more explicit description of train, validation, and test separation. The downstream evaluation should be broadened beyond skull stripping, and experiments on real anisotropic acquisitions, including slice gap or mismatch from the assumed degradation model, would make the practical conclusions more convincing. A more detailed analysis of the mouse results could also help identify whether remedies such as domain adaptation, multi-species training, or fully 3D generative models are needed.

References

[1] Yuhua Chen, Anthony G. Christodoulou, Zhengwei Zhou, Feng Shi, Yibin Xie, and Debiao Li. Mri super-resolution with GAN and 3d multi-level densenet: Smaller, faster, and better. *CoRR*, abs/2003.01217, 2020.

[2] Yuhua Chen, Feng Shi, Anthony G. Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, volume 11070 of *Lecture Notes in Computer Science*, pages 91–99. Springer, 2018.

[3] Hyungjin Chung, Dohoon Ryu, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Solving 3d inverse problems using pre-trained 2d diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22542–22551, 2023.

[4] Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[5] Amod Jog, Aaron Carass, Snehashis Roy, Dzung L. Pham, and Jerry L. Prince. Self super-resolution for magnetic resonance images. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, volume 9902 of *Lecture Notes in Computer Science*, pages 553–560. Springer, 2016.

[6] Samuel W. Remedios et al. Deep filter bank regression for super-resolution of anisotropic MR brain images. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2022*, volume 13436 of *Lecture Notes in Computer Science*, pages 613–622. Springer, 2022.

[7] Samuel W. Remedios et al. Self-supervised super-resolution for anisotropic MR images with and without slice gap. In *Simulation and Synthesis in Medical Imaging (SASHIMI)*, volume 14288 of *Lecture Notes in Computer Science*, pages 118–128. Springer, 2023.

[8] Samuel W. Remedios et al. Pushing the limits of zero-shot self-supervised super-resolution of anisotropic MR images. In *Proceedings of SPIE Medical Imaging*, volume 12926, pages 29–35, 2024.

[9] Samuel W. Remedios et al. Eclare: Efficient cross-planar learning for anisotropic resolution enhancement. *arXiv preprint arXiv:2503.11787*, 2025.

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Michael Unser. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.

[12] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations (ICLR)*, 2023.

[13] Can Zhao, Aaron Carass, Blake E. Dewey, and Jerry L. Prince. Self super-resolution for magnetic resonance images using deep networks. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 365–368. IEEE, 2018.

Annexe

A. Two-sided Wilcoxon signed-rank test (paired, non-parametric)

In this study, each test volume produces a paired performance measurement for two methods (ECLARE and BiDA), so the natural object of comparison is the per-volume difference in scores. The authors use a two-sided Wilcoxon signed-rank test to compare these paired results across both super-resolution scales and both in-/out-of-domain settings, with $\alpha = 0.001$.

Let x_i be the metric value (e.g., PSNR or cDSC) for method A on case i, and y_i the value for method B on the same case. Define $d_i = x_i - y_i$. The test evaluates:

$$H_0 : \text{median}(d_i) = 0 \qquad \text{vs} \qquad H_1 : \text{median}(d_i) \neq 0.$$

It is *two-sided* because it can detect a systematic advantage in either direction. Conceptually, the test ranks the absolute differences $|d_i|$ and checks whether the positive and negative signed ranks balance. Here, a strong imbalance suggests a consistent performance shift between the methods. Because it does not assume normally distributed differences, it is a common choice for medical-imaging metrics that can be skewed or heavy-tailed.

B. Perception–distortion tradeoff

Blau and Michaeli separate two notions for restoration outputs $\hat{x}$: (i) *distortion*, measured by full-reference criteria comparing $\hat{x}$ to a ground-truth x (e.g., MSE/PSNR), and (ii) *perceptual quality*, which reflects how much $\hat{x}$ looks like a natural image and is tied to deviations from natural image statistics.

Their key theoretical point is that these two goals cannot, in general, be optimized simultaneously: close to the optimal boundary, improving distortion forces the distribution of reconstructions to deviate more from the natural-image distribution (and conversely, pushing for perceptual realism typically increases distortion). They emphasize that this is not specific to MSE; the tradeoff appears for many distortion measures, implying that distortion alone is not sufficient to assess restoration methods. Practically, this explains a common observation in super-resolution: high-PSNR reconstructions may look overly smooth (averaging away uncertain high-frequency details), while sharper outputs can have lower PSNR yet preserve structures that are more useful for interpretation or downstream tasks.

C. Shinnar–Le Roux (SLR) algorithm for slice selection profiles

To evaluate super-resolution, the paper simulates LR thick-slice MRI from HR volumes. Instead of using simple downsampling, the authors first apply a slice-selection profile generated with the Shinnar–Le Roux (SLR) algorithm, then subsample the data according to the slice spacing.

In simple terms, the SLR algorithm is a standard MRI method used to model how slices are actually acquired. In this paper, its main role is to make the degradation process more realistic, because it accounts for the blurring caused by finite slice thickness before downsampling. This makes the simulated LR data closer to real MRI acquisitions.

D. Why rat-to-mouse transfer may appear easy in natural images but fail here

At a semantic level, rat and mouse images appear to define only a mild domain gap: both are small mammals with similar global shape and broadly similar local visual statistics, so zero-shot natural-image generative models can often transfer across them without major difficulty. In those settings, success usually depends on producing content that is visually plausible within a broad natural-image manifold, and the model is not required to recover anatomically exact missing structure from a physically constrained measurement process.

That intuition does not carry over cleanly to anisotropic MRI super-resolution. Here, the task is an inverse problem in which the observed LR image fixes the measured component of the signal and the generative prior is used to reconstruct the missing through-plane high-frequency information. In this regime, transfer depends on whether the target anatomy lies sufficiently close to the source support precisely in the subspace that must be inferred. Medical image distributions are much narrower and more sensitive to anatomy, scale, resolution, field of view, and acquisition physics than generic natural-image distributions. As a result, even a semantically small species shift can become a meaningful support mismatch for reconstruction. This interpretation is consistent with the paper's own observations: the mouse dataset required zero-padding to match the rat field of view, the authors explicitly note domain shift between species in anatomical size and resolution, and they report that BiDA did not fully generalize to mouse data. In other words, what looks like a minor semantic gap in

natural images becomes a much finer and more consequential mismatch in medically relevant structure here, particularly because the model must infer missing details rather than merely generate plausible appearance.

E. Implementation Details and Practical Simplifications

Our implementation should be understood as a lightweight reproduction of the core BiDA idea rather than a full replication of the paper's training and evaluation pipeline. We kept the main structure of the method, namely two slice-wise 2D diffusion priors, DDNM-based inference, and final biplanar averaging, but made several practical simplifications motivated by hardware, storage, and development constraints.

First, we simplified the data setup. The paper trains on CAMRI rats together with the much larger MultiRat dataset and evaluates on Optogenetic rats and CAMRI mice. In our implementation, we restricted training to the CAMRI rat dataset and treated the CAMRI mouse as the true out-of-domain set. We did not include the additional rat cohorts used in the paper, mainly because the full preprocessing and storage burden was too high for our setup, with one candidate dataset reaching on the order of 100 GB. Instead, we used a deterministic subject-level split within the available rat data, which preserved a leakage-safe train/validation/test protocol while keeping preprocessing manageable.

Second, we shortened training substantially. The paper reports a larger-scale training and inference setup, whereas we trained for only 20 epochs with a batch size of 8 on a single RTX 4070 Laptop GPU with 8 GB of VRAM. This was a deliberate compromise: our goal was to verify whether the method could already produce meaningful behavior under limited compute, not to match the paper's final performance. For the same reason, we used mixed precision and a Windows-safe data-loading configuration. We also omitted the small rigid augmentations described in the paper, since our focus was on validating the main reconstruction mechanism before spending additional compute on augmentation.

Third, some architectural and sampling choices were made at the level of a faithful approximation rather than an exact reproduction. The paper does not provide enough low-level implementation detail to reproduce its training stack exactly, so we used a compact 2D UNet with standard DDPM components and velocity-style prediction. Likewise, while the paper uses DDIM with 16 timesteps at inference, we increased the number of DDIM steps to 100. In our setting, this gave more stable reconstructions for partially trained models and reduced the risk that poor results would come only from overly aggressive sampling rather than from the method itself.

Fourth, we kept the paper's main degradation assumptions but implemented them in a numerically stricter way. As in the paper, we modeled the forward operator with a Gaussian through-plane blur and no slice gap. However, for the pseudo-inverse, we used nearest-neighbor upsampling rather than bilinear interpolation, because we found that this better preserved alignment with the sampled voxels inside the DDNM correction. We also cropped the through-plane dimension to a multiple of the scale factor before inference to avoid rounding inconsistencies in repeated applications of $A^{\dagger}A$. These are implementation-level assumptions, but they were important for obtaining stable data-consistent updates.

Finally, our evaluation is narrower than the paper's by design. We reproduced only a simpler comparison centered on BiDA and cubic interpolation, rather than reimplementing the full ECLARE baseline and the complete multi-dataset experimental suite. We also reported PSNR, SSIM, and MAE in addition to volume-wise statistical comparisons, while the paper focuses on PSNR and downstream cDSC. This means our experiments should be interpreted as a controlled sanity check of the BiDA mechanism under constrained resources, not as a direct attempt to reproduce the paper's exact tables or claims.

F. Results and Interpretation

Experimental setup. We evaluated our simplified BiDA-style implementation on through-plane super-resolution of rodent brain MRI along the A axis in RAS orientation. Training was unconditional, meaning the model saw only HR slices during training and was conditioned on the LR observation only at inference through DDNM. We tested on two sets: a held-out rat set of about 20 volumes from the CAMRI rat dataset, which we treat as a proxy OOD setting, and a true OOD mouse set of 32 volumes. LR inputs were generated synthetically from HR volumes using the same degradation model as in inference, namely Gaussian through-plane blur followed by strided subsampling, with scale 5 for rats and 2 for mice. We compared cubic interpolation with our BiDA-style implementation and evaluated both with volume-level PSNR, slice-averaged SSIM, and volume-level MAE. Statistical significance was assessed with paired Wilcoxon signed-rank tests.

Quantitative and qualitative results. Qualitatively, BiDA clearly produces sharper and more plausible reconstructions than interpolation in both the held-out rat and mouse examples shown in Figure 1. This is

the most immediate visual result of our implementation. Quantitatively, however, the pattern is more mixed. As shown in Table 1, BiDA improves PSNR over interpolation on both datasets, from 22.96 to 23.55 on the held-out rat set and from 31.66 to 34.30 on the mouse set, with statistically significant Wilcoxon results in both cases. In contrast, interpolation remains better in SSIM and MAE on both datasets. This means that our implementation consistently improves PSNR, but not the other two metrics.

Table 1: Evaluation results for our implementation. Values are mean ± standard deviation.

Dataset	**Metric**	**BiDA-style**	**Interpolation**
mice_ood	PSNR	34.30 ± 1.96	31.66 ± 1.05
	SSIM	0.9316 ± 0.0107	0.9710 ± 0.0121
	MAE	0.0097 ± 0.0018	0.0071 ± 0.0019
rats_test	PSNR	23.55 ± 3.02	22.96 ± 2.54
(proxy OOD)	SSIM	0.7792 ± 0.1153	0.8200 ± 0.1097
	MAE	0.0385 ± 0.0172	0.0367 ± 0.0154

Interpretation. We interpret these results through the perception-distortion tradeoff discussed previously. Diffusion-based reconstructions tend to introduce sharper and more realistic-looking high-frequency structure, which can improve perceptual quality and sometimes PSNR, but may still deviate from the reference in local patch statistics. SSIM is particularly sensitive to this kind of mismatch, while interpolation produces smoother and blurrier outputs that remain structurally closer in a local sense. Our results fit this pattern well: BiDA looks qualitatively better and improves PSNR, but interpolation still scores better on SSIM and MAE. In other words, our implementation appears to recover sharper detail, but not yet in a way that is consistently more faithful under all metrics.

The PSNR boxplot in Figure 2 also reveals an interesting trend. The margin between BiDA and interpolation is much larger on the true OOD mouse set than on the held-out rat set, where both methods remain closer. We hypothesize that this is partly because the mouse setting uses the easier scale factor 2, whereas the held-out rat setting uses the much harder scale factor 5. Under our limited training regime, the diffusion prior may already be strong enough to outperform interpolation clearly on the easier task, even across species, while still being undertrained for the more severe rat reconstruction problem. This makes us think that with longer training, larger rat data, and a setup closer to the paper, BiDA would likely separate more clearly from interpolation on the rat setting as well.

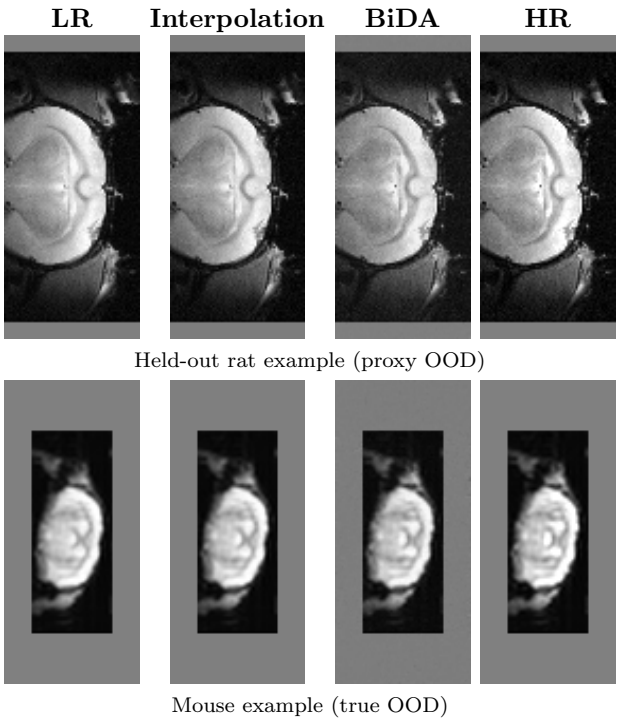

Figure 1: Qualitative comparison. From left to right: LR input, interpolation, BiDA, and HR reference. Top: held-out rat example (proxy OOD). Bottom: mouse example (true OOD).

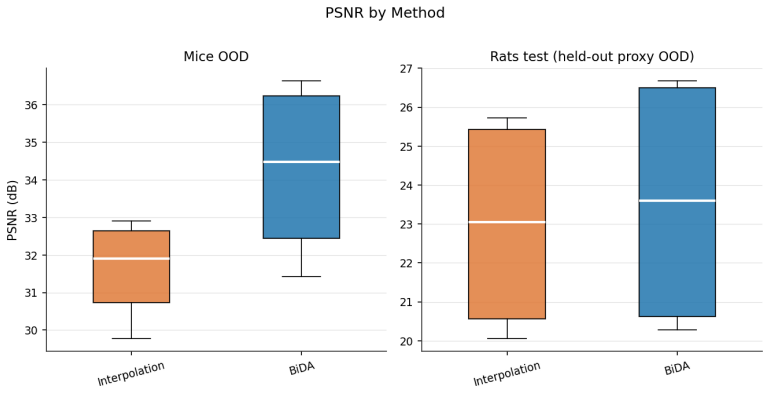

Figure 2: Volume-level PSNR comparison between interpolation and our BiDA-style implementation on the held-out rat set (proxy OOD) and the mouse set (true OOD).