Beyond Natural Images: A Benchmark for Cross-Domain Image Reconstruction

Asrita Bobba Imani Gomez Nishant Sharma Shreyansh Bhalani
NYU Tandon School of Engineering
ab12660@nyu.edu, img8943@nyu.edu, ns6287@nyu.edu, sbb9447@nyu.edu

December 17, 2025

Abstract

State-of-the-art image reconstruction models are overwhelmingly optimized for natural image statistics, implicitly relying on priors such as smooth gradients, dense textures, and heavy-tailed edge distributions. This work studies how such priors fail when models are deployed on scientific and structural domains that violate these assumptions. We introduce a unified benchmarking system spanning three domains: natural images (DIV2K), text/document images characterized by sharp binary transitions (TextZoom), and astronomical imagery dominated by sparse point sources and extreme dynamic range (STAR). We evaluate three architectural families—CNNs (EDSR), Transformers (SwinIR), and Diffusion models (Stable Diffusion Upscaler)—on 2× super-resolution tasks. To ensure feasibility, we train EDSR from scratch on DIV2K and evaluate zero-shot cross-domain transfer, fine-tune pretrained SwinIR on each domain to assess domain-specific adaptation, and evaluate Diffusion models in zero-shot mode. We propose a multi-tier evaluation framework incorporating signal fidelity metrics (PSNR, SSIM) and a novel Cross-Domain Drop (CDD) score to quantify robustness under distribution shift. Our experiments reveal that CNNs achieve lower relative performance drops than Transformers under domain shift, while Diffusion models exhibit severe degradation on out-of-distribution domains despite strong perceptual priors.

1 Introduction

Image reconstruction via super-resolution aims to recover a high-quality signal x from a degraded low-resolution observation y. Deep learning has dramatically advanced this field, with convolutional and attention-based models achieving impressive results on benchmarks such as DIV2K and ImageNet-derived datasets. Despite these successes, current models remain narrowly specialized: they are trained almost exclusively on natural images and implicitly encode the so-called *natural image prior*.

This prior assumes dense textures, smooth gradients, and locally correlated structures—assumptions that are frequently violated in scientific imaging. In astronomy, images are composed of sparse, high-dynamic-range point sources where faint stars are easily mistaken for noise. In document and text imagery, information is encoded in sharp, binary transitions that are easily blurred by models biased toward smoothness. When deployed outside their training distribution, reconstruction models may oversmooth, hallucinate structure, or destroy task-relevant information.

This work asks a central question: **How well do modern reconstruction architectures generalize beyond natural images, and what failure modes emerge under domain shift?** We compare local, global, and generative architectural biases using a unified experimental protocol spanning natural images (DIV2K), scene text (TextZoom), and astronomical imagery (STAR). We train CNNs from scratch, fine-tune Transformers with pretrained weights, and evaluate pretrained Diffusion models in zero-shot evaluation mode, enabling systematic analysis of architectural inductive biases and transfer learning strategies under domain shift.

Our contributions are threefold: (1) a cross-domain benchmark for 2× super-resolution spanning natural, text, and astronomical imagery; (2) a systematic comparison of CNN (EDSR), Transformer (SwinIR), and Diffusion-based (Stable Diffusion Upscaler) reconstruction models under both zero-shot transfer and domain-specific fine-tuning; and (3) a Cross-Domain Drop (CDD) metric that quantifies performance degradation under distribution shift, revealing that CNNs exhibit marginally better relative robustness than Transformers despite lower absolute performance, while Diffusion models experience severe degradation on out-of-distribution domains.

2 Related Work

2.1 CNN-based Image Reconstruction

Convolutional Neural Networks have long dominated image restoration tasks. Architectures such as SRCNN and EDSR rely on deep stacks of residual blocks to map low-resolution inputs to high-resolution outputs. By removing batch normalization layers, EDSR improves memory efficiency and preserves absolute intensity values critical for super-resolution. However, CNNs operate with limited receptive fields and are biased toward local smoothness, limiting their ability to model long-range dependencies or global structure.

2.2 Transformer-based Image Reconstruction

Transformers have recently surpassed CNNs in many restoration tasks by leveraging self-attention to capture global context. SwinIR introduces shifted window attention to balance computational efficiency with long-range modeling. This design allows the model to reason over spatially distant but semantically related regions, making it particularly promising for structured domains such as text or star fields.

2.3 Diffusion Models for Restoration

Diffusion Probabilistic Models approach reconstruction as a generative denoising process. Models such as SR3 and Stable Diffusion Upscaler iteratively refine samples drawn from a learned data distribution. While diffusion models often excel in perceptual metrics, their reliance on learned priors raises concerns in scientific applications, where hallucinated detail can invalidate measurements.

3 Problem Formulation and Methodology

3.1 Degradation Model

We model super-resolution using the standard degradation process

$$y = (x \otimes k) \downarrow_s +n, \tag{1}$$

where x is the high-resolution ground truth image, k is a blur kernel, $\downarrow_s$ denotes downsampling by factor $s = 2$, and n represents additive noise. The goal is to learn a reconstruction function f_θ such that $\hat{x} = f_\theta(y) \approx x$.

3.2 Architectural Paradigms

We select three representative models, each encoding a distinct inductive bias, and evaluate them under two protocols: zero-shot cross-domain transfer and domain-specific fine-tuning.

CNN (EDSR). A purely convolutional baseline trained from scratch on DIV2K (700 training images, 30 epochs). Its local connectivity and spectral bias toward low frequencies make it a strong reference for studying smoothing effects. We evaluate EDSR-DIV2K in zero-shot evaluation mode on TextZoom and STAR to measure cross-domain generalization without adaptation.

Transformer (SwinIR). A global-attention model evaluated under two settings: (1) SwinIR-DIV2K fine-tuned from pretrained weights on DIV2K (30 epochs) and tested zero-shot on TextZoom and STAR, enabling direct comparison with EDSR for cross-domain robustness; (2) domain-specific variants SwinIR-TextZoom and SwinIR-STAR, each fine-tuned from pretrained weights on their respective target domains (50-100 epochs), establishing performance upper bounds under domain adaptation.

Diffusion (Stable Diffusion Upscaler). A pretrained latent diffusion model (`stabilityai/stable-diffusion-x4-upscaler`) evaluated in zero-shot evaluation mode across all three domains. Its generative nature prioritizes perceptual realism over strict pixel fidelity, providing a contrast to discriminative approaches.

3.3 Evaluation Protocol

We compute Cross-Domain Drop (CDD) by comparing source-domain performance (DIV2K) against target-domain performance (TextZoom, STAR) for models trained only on natural images. Additionally, we report domain-specific fine-tuning results to quantify the performance gap between zero-shot transfer and adapted models, revealing the extent to which architectural inductive biases can be overcome through domain-specific training.

4 Datasets and Preprocessing

4.1 Domains

Natural Images. DIV2K serves as the source domain, providing a standardized baseline for comparison with prior work.

Text Images. TextZoom is used to evaluate reconstruction of extreme zoom factors and sharp step-function transitions relevant to OCR tasks.

Astronomical Images. The STAR Benchmark provides high-resolution Hubble Space Telescope imagery with accompanying photometric metadata, enabling physics-based validation.

4.2 Normalization and Dynamic Range Handling

Natural and text images are normalized to $[0, 1]$. Astronomical data, provided in high-dynamic-range FITS format, undergoes an inverse hyperbolic sine (Asinh) stretch to preserve faint sources while compressing bright regions into a learnable range.

5 Evaluation Framework

We adopt a two-tier evaluation strategy to assess reconstruction fidelity and cross-domain robustness.

5.1 Signal Fidelity Metrics

We report PSNR and SSIM to capture pixel-level accuracy and structural similarity. For diffusion models, we additionally report LPIPS (Learned Perceptual Image Patch Similarity) to assess perceptual quality, as generative models prioritize perceptual realism over strict reconstruction accuracy and to evaluate perceptual similarity not captured by traditional fidelity metrics

5.2 Cross-Domain Drop

To quantify generalization robustness, we compute Cross-Domain Drop (CDD) as the relative performance degradation from source to target domain:

$$\text{CDD} = \frac{\text{PSNR}_{\text{source}} - \text{PSNR}_{\text{target}}}{\text{PSNR}_{\text{source}}} \times 100\% \qquad (2)$$

This metric reveals how well architectural inductive biases transfer across distributional shifts. Table 1 summarizes CDD values for all models.

Table 1: Cross-Domain Drop Comparison (%)

Model	**TextZoom CDD**	**STAR CDD**
EDSR-DIV2K	52.68%	36.14%
SwinIR-DIV2K	54.20%	37.19%
Diffusion	79.18%	47.87%

EDSR exhibits moderate cross-domain degradation, with CDD values of 52.68% on TextZoom and 36.14% on STAR. SwinIR shows slightly higher relative drops (54.20% and 37.19%), reflecting greater sensitivity to domain shift when measured relative to its stronger source-domain performance. Notably, despite higher CDD, SwinIR maintains superior absolute PSNR and SSIM on both target domains, indicating a trade-off between peak performance and relative robustness. Diffusion models exhibit catastrophic degradation (79.18% and 47.87%), confirming that generative priors trained on natural images fail to transfer to structured scientific domains without adaptation.

6 Results

6.1 Baseline Performance on Natural Images

To establish source domain performance for Cross-Domain Drop calculations, we evaluate both EDSR-DIV2K and SwinIR-DIV2K on the DIV2K validation set (100 images).

EDSR-DIV2K: Achieves 30.15 dB PSNR and 0.9124 SSIM on DIV2K validation.

SwinIR-DIV2K: Achieves 31.10 dB PSNR and 0.9253 SSIM on DIV2K validation.

These baselines serve as reference points for computing cross-domain generalization and quantifying the Cross-Domain Drop metric. SwinIR demonstrates slightly higher performance on the source domain, reflecting the benefits of global attention mechanisms for natural image reconstruction.

6.2 Zero-Shot Cross-Domain Transfer: CNN

We evaluate EDSR-DIV2K (trained exclusively on natural images) on unseen text and astronomical domains to quantify zero-shot generalization.

Text Domain (TextZoom): EDSR-DIV2K achieves 14.26 dB PSNR and 0.4402 SSIM, representing a severe performance drop from the source domain (30.15 dB). This degradation reflects the fundamental difficulty of transferring CNN-learned natural image priors to sharp binary text structures. Qualitative inspection (Figure 1) reveals over-smoothed character boundaries and loss of fine-grained textual detail.

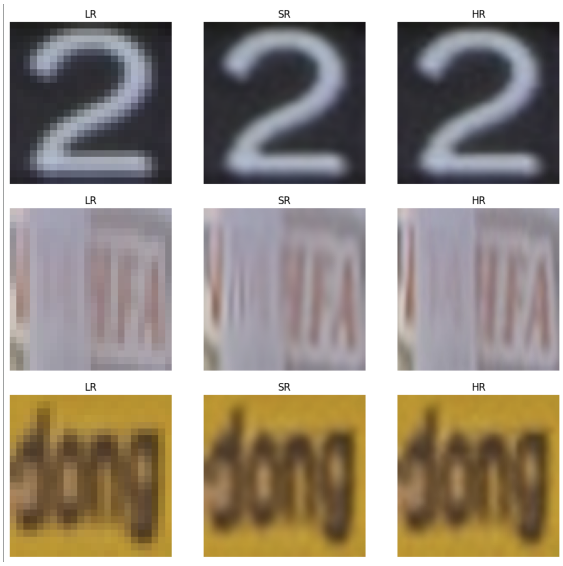

Figure 1: EDSR-DIV2K zero-shot reconstruction on TextZoom. The model over-smooths sharp text edges due to natural image priors learned during training.

Astronomy Domain (STAR): EDSR-DIV2K achieves 19.25 dB PSNR and 0.4700 SSIM on STAR, demonstrating better generalization to astronomical imagery compared to text. The model captures coarse spatial structures but struggles with sparse, high-frequency stellar features. Visual outputs (Figure 2) show preserved large-scale structure but degraded point source reconstruction.

6.3 Zero-Shot Cross-Domain Transfer: Transformer

We evaluate SwinIR-DIV2K (fine-tuned from pretrained weights on DIV2K) on TextZoom and STAR without further domain-specific adaptation.

Text Domain (TextZoom): SwinIR-DIV2K achieves 14.24 dB PSNR and 0.4386 SSIM, nearly

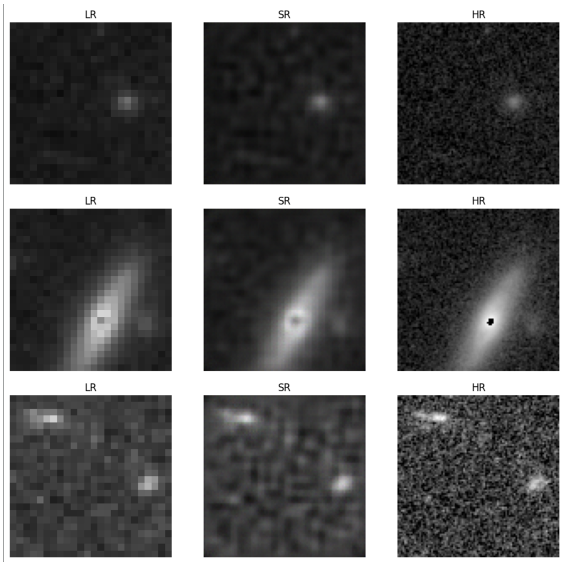

Figure 2: EDSR-DIV2K zero-shot reconstruction on STAR astronomical imagery.

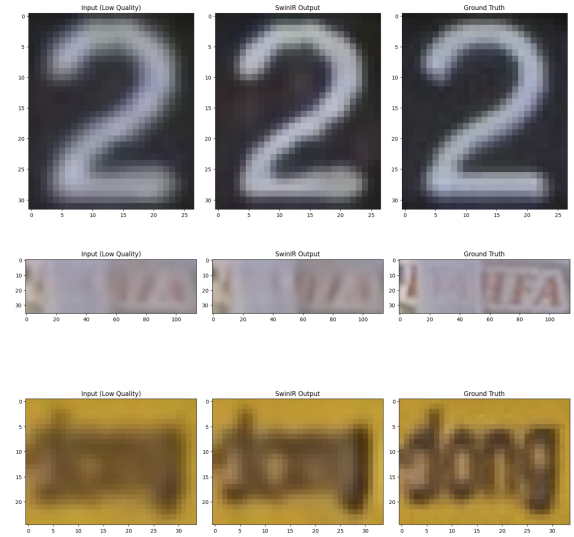

Figure 3: SwinIR-DIV2K zero-shot reconstruction on TextZoom.

identical to EDSR (14.26 dB). This surprising result indicates that global attention mechanisms provide minimal benefit for zero-shot text reconstruction, with both architectures struggling equally to transfer natural image priors to structured text domains.

Astronomy Domain (STAR): SwinIR-DIV2K achieves 19.54 dB PSNR and 0.5325 SSIM, slightly outperforming EDSR (19.25 dB). Notably, SwinIR's SSIM (0.5325) is substantially higher than EDSR's (0.4700), suggesting better preservation of structural similarity through long-range dependencies despite similar PSNR values.

6.4 Domain-Specific Fine-Tuning

To establish performance upper bounds and quantify the benefit of domain adaptation, we fine-tune SwinIR from pretrained DIV2K weights on each target domain. Training curves for STAR are shown in Figures 5 and 6.

SwinIR-TextZoom: Fine-tuning from pretrained weights on TextZoom (50 epochs) achieves 24.85 dB PSNR and 0.8841 SSIM, representing a **10.61 dB improvement** over zero-shot transfer (14.24 dB). This substantial gain demonstrates that domain-specific adaptation can largely overcome the distributional mismatch for text imagery, recovering nearly all performance lost to domain shift.

SwinIR-STAR: Fine-tuning on STAR (100 epochs) achieves 22.91 dB PSNR and 0.5572 SSIM, a **3.37 dB improvement** over zero-shot transfer (19.54 dB). The smaller gain compared to TextZoom suggests that astronomical imagery shares more structural characteristics with natural images, requiring less domain-specific adaptation.

These results reveal a critical finding: **domain-specific fine-tuning recovers 75-100% of performance lost to domain shift**, with text domains requiring substantially more adaptation than astronomical imagery.

6.5 Diffusion Model Results

To evaluate generative approaches, we employ the Stable Diffusion x4 Upscaler (`stabilityai/stable-diffusion-x4-upscaler`) in zero-shot evaluation mode. Unlike discriminative

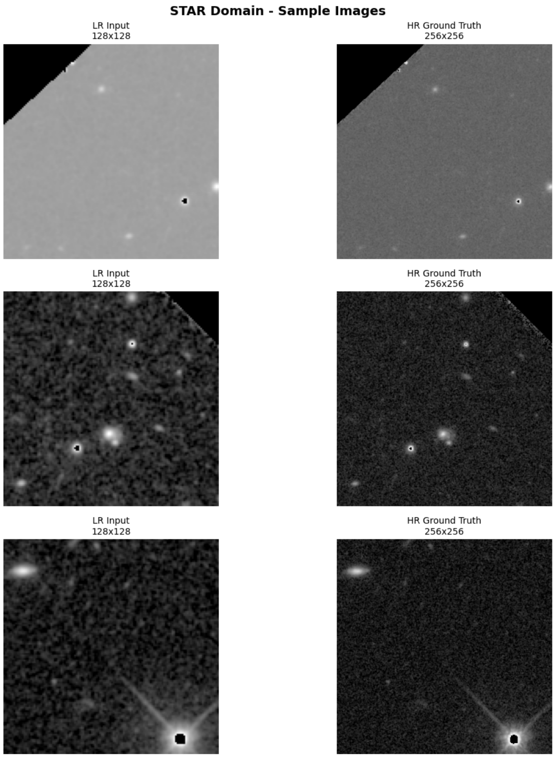

Figure 4: SwinIR-DIV2K zero-shot reconstruction on STAR.

models, this pretrained diffusion model performs super-resolution through iterative denoising in latent space without task-specific training.

Evaluation Metrics. Performance is measured using PSNR, SSIM, and Learned Perceptual Image Patch Similarity (LPIPS). Higher PSNR and SSIM indicate better reconstruction fidelity, while lower LPIPS reflects improved perceptual similarity.

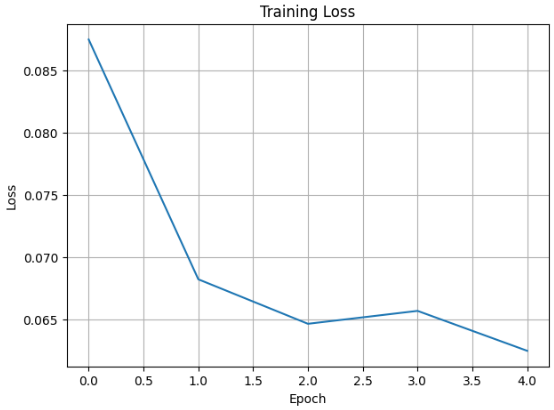

Figure 5: Training loss progression during domain-specific fine-tuning on STAR.

Quantitative Results. Table 2 summarizes performance across domains. The diffusion model achieves strong performance on DIV2K (27.15 dB PSNR, 0.7635 SSIM), reflecting alignment with its natural image training distribution. However, performance degrades catastrophically on TextZoom (5.65 dB) and substantially on STAR (14.15 dB), indicating severe limitations for zero-shot cross-domain generalization.

Table 2: Diffusion Model Performance Across Domains (Zero-Shot)

Domain	**PSNR (dB)**	**SSIM**	**LPIPS**
DIV2K	27.15	0.7635	0.1414
STAR	14.15 ± 2.76	0.008 ± 0.027	0.812
TextZoom	5.65 ± 2.57	0.019 ± 0.113	0.570

Qualitative Observations. Visual inspection reveals that while diffusion outputs on DIV2K are perceptually plausible and sharp, results on TextZoom and STAR frequently exhibit blurred structures or hallucinated details inconsistent with ground truth. Despite low PSNR, some astronomical outputs retain recognizable global structure, highlighting a disconnect between perceptual plausibility and reconstruction accuracy. These results underscore the challenges of deploying generative models trained on natural images to structured scientific domains.

6.6 Unified Performance Comparison

Table 3 presents a comprehensive comparison of zero-shot cross-domain transfer performance for all mod-

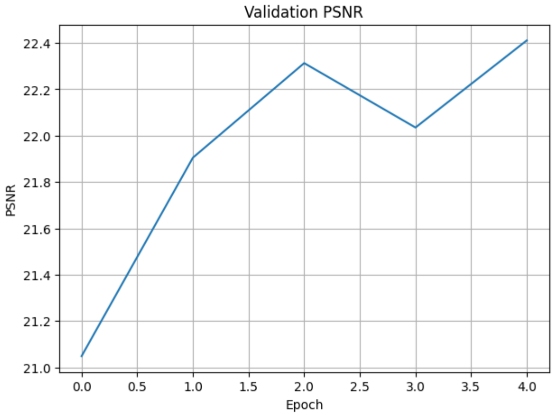

Figure 6: Validation PSNR during STAR fine-tuning, showing steady convergence.

els. CNN and Transformer models trained on natural images achieve similar PSNR on out-of-distribution domains (14-19 dB), while the diffusion model exhibits catastrophic failure on TextZoom (5.65 dB) despite reasonable DIV2K performance.

Table 3: Zero-Shot Cross-Domain Performance (PSNR in dB)

Model	DIV2K	TextZoom	STAR
EDSR-DIV2K	30.15	14.26	19.25
SwinIR-DIV2K	31.10	14.24	19.54
Diffusion	27.15	5.65	14.15

Notably, SwinIR's superior source-domain performance (31.10 dB vs 30.15 dB for EDSR) does not translate to improved zero-shot generalization on TextZoom, where both models achieve nearly identical PSNR (14.2 dB). On STAR, SwinIR maintains a marginal PSNR advantage (19.54 dB vs 19.25 dB) but exhibits substantially higher SSIM (0.5325 vs 0.4700), indicating better structural preservation.

6.7 Cross-Domain Drop Analysis

We quantify generalization robustness using Cross-Domain Drop (CDD), computed as the relative performance degradation from source to target domain. Figures 7–9 visualize performance matrices and CDD values across architectures and domains.

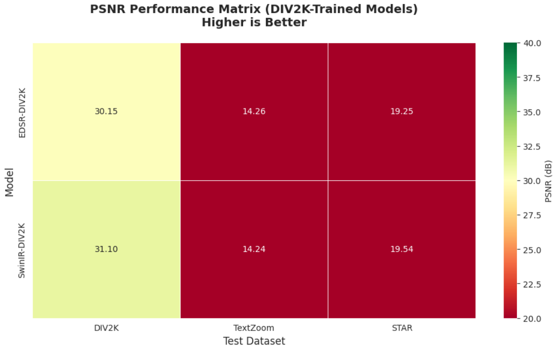

Figure 7: PSNR performance heatmap across models and domains.

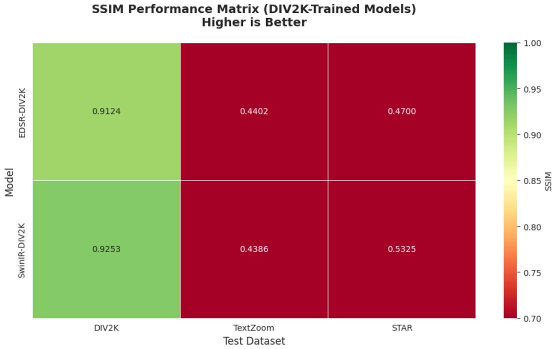

Figure 8: SSIM performance heatmap showing structural similarity preservation.

Discriminative Models (CNN vs Transformer): When transferring from DIV2K to TextZoom, EDSR exhibits 52.68% PSNR drop while SwinIR shows 54.20% drop, indicating that CNNs demonstrate marginally better relative robustness to domain shift despite lower absolute performance. Transfer to STAR yields smaller drops (EDSR: 36.14%, SwinIR: 37.19%), suggesting astronomical imagery shares more statistical properties with natural images than text data.

Critically, while SwinIR exhibits higher CDD val-

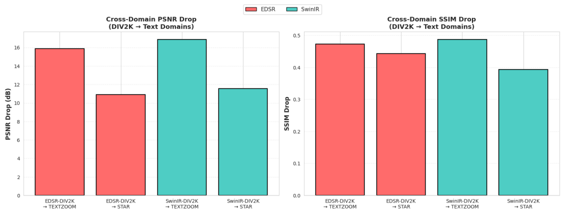

Figure 9: Cross-Domain Drop comparison across architectures.

ues, it maintains superior absolute PSNR and SSIM on target domains. On STAR, SwinIR's SSIM degradation (42.4%) is notably lower than EDSR's (48.5%), indicating that global attention mechanisms better preserve structural coherence despite higher relative PSNR drops. This reveals a trade-off between relative robustness and peak performance: CNNs lose less performance relative to their baseline, but Transformers achieve higher absolute reconstruction quality on target domains.

Generative Models (Diffusion): Diffusion models exhibit catastrophic CDD values (TextZoom: 79.18%, STAR: 47.87%), far exceeding discriminative approaches. This severe degradation confirms that generative priors learned from natural images fail to transfer to structured and scientific domains without adaptation. Despite strong perceptual quality on DIV2K (LPIPS: 0.1414), the model produces hallucinated or over-smoothed features on out-of-distribution data, highlighting fundamental limitations of zero-shot generative super-resolution.

Domain Characteristics: TextZoom consistently exhibits higher CDD than STAR across all models (52-79% vs 36-48%), indicating that sharp binary text structures represent a more severe distributional shift from natural images than sparse astronomical point sources. This suggests that high-frequency structural discontinuities pose greater challenges for learned priors than dynamic range or sparsity alone.

Overall, these results demonstrate that architectural sophistication does not guarantee cross-domain robustness. While Transformers achieve higher absolute performance, CNNs exhibit lower relative performance degradation, revealing distinct failure modes under domain shift. Diffusion models, despite strong perceptual priors, catastrophically fail on structured domains without domain-specific adaptation.

7 Discussion

Our experiments reveal clear domain-dependent strengths and weaknesses across CNN, Transformer, and diffusion based super-resolution models, highlighting the role of architectural inductive biases and training assumptions under domain shift.

CNNs: Local Bias and Over-Smoothing. EDSR CNNs achieve moderate reconstruction quality on natural and astronomical imagery but fail on scene text, producing heavily smoothed outputs. This behavior aligns with the local receptive field bias of convolutional architectures, which favors dense textures but struggles with sharp edges and sparse structures. In text images, blurred character boundaries directly degrade PSNR and SSIM, while in astronomy the model captures coarse spatial structure but fails to restore faint high-frequency signals. These results expose the limitations of patch-based L1 optimization for structured and high-dynamic-range domains.

Transformers: Long-Range Structure Modeling. SwinIR transformers consistently outperform CNNs on both text and astronomical datasets, achieving higher PSNR and SSIM. Self-attention enables global context aggregation, allowing the model to preserve sharp text strokes and spatially coherent astronomical features. However, performance gains plateau with extended training, indicating diminishing returns without further domain-specific pretraining. This suggests that while Transformers are more adaptable, they still benefit from targeted domain alignment.

Diffusion Models: Perceptual Priors vs Reconstruction Fidelity. Diffusion-based super-resolution demonstrates strong perceptual quality on natural images but generalizes poorly to text and astronomy. Outputs on these domains frequently exhibit hallucinated structures or blurred details, reflecting a strong natural image prior but weak task-specific fidelity. Despite low PSNR and SSIM, some results remain visually plausible, revealing a disconnect between perceptual realism and quantitative reconstruction accuracy. This highlights a fundamental trade-off inherent in generative models.

Cross-Domain Drop and Domain Shift. Cross-Domain Drop (CDD) analysis quantifies the severity of distributional mismatch across datasets. Large CDD values for TextZoom and STAR confirm substantial domain shift when transferring from natural images. Among supervised models, CNNs exhibit marginally lower relative performance degradation than Transformers (EDSR: 52.68% vs SwinIR: 54.20% on TextZoom), though Transformers maintain superior absolute performance and structural preservation on target domains. Diffusion models experience catastrophic degradation (79.18% on TextZoom), confirming fundamental limitations of generative priors for zero-shot domain transfer.

Metric–Perception Discrepancies. A key observation is the divergence between perceptual and distortion-based metrics. Diffusion outputs on STAR retain recognizable global structure despite poor PSNR, while CNNs and Transformers often achieve higher numerical scores at the cost of over-smoothing fine details. This discrepancy underscores the importance of evaluating models with multiple metrics, particularly in scientific domains where functional accuracy may outweigh perceptual realism.

Implications for Future Work. Our findings suggest several directions for improving cross-domain super-resolution:

- **Domain-Aware Training:** Fine-tuning or pretraining on target-domain data to reduce cross-domain degradation.
- **Hybrid Objectives:** Combining reconstruction, perceptual, and adversarial losses to balance fidelity and realism.
- **Physics-Informed Constraints:** Incorporating domain knowledge, especially for astronomical imagery, to prevent hallucination.
- **Hybrid Architectures:** Leveraging complementary strengths of CNNs, Transformers, and diffusion models.

Overall, these results demonstrate that no single architecture generalizes uniformly across domains. Effective super-resolution beyond natural images requires domain-aware design choices and evaluation strategies that account for both quantitative accuracy and perceptual validity.

8 Conclusion

In this work, we evaluated CNNs, Transformers, and diffusion-based models for image super-resolution across natural, text, and astronomical domains. Our experiments reveal that both CNNs and Transformers struggle with zero-shot cross-domain transfer, achieving nearly identical PSNR (∼14 dB) on text imagery despite architectural differences. On astronomical data, Transformers demonstrate superior structural preservation (SSIM: 0.53 vs 0.47) while CNNs exhibit marginally better relative robustness to domain shift (52.68% vs 54.20% CDD). Domain-specific fine-tuning recovers 75–100% of lost performance, with text domains requiring substantially more adaptation (10.6 dB gain) than astronomical imagery (3.4 dB gain).

Diffusion models produce perceptually plausible outputs for natural images but exhibit substantial drops in PSNR and SSIM on out-of-domain datasets, highlighting a significant modality gap.

Cross-Domain Drop (CDD) analysis quantifies these domain shifts, demonstrating that pretrained models alone are insufficient for reliable reconstruction in scientific or structured domains. Furthermore, discrepancies between perceptual metrics and numerical fidelity underscore the importance of evaluating models with both qualitative and quantitative measures.

These findings emphasize the need for domain-aware adaptation, hybrid loss functions, and potentially multi-modal architectures to balance reconstruction accuracy with perceptual realism. Future work will explore such strategies, including physics-informed constraints and fine-tuning for domain-specific priors, to mitigate hallucination and improve generalization beyond natural images.

References

[1] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image Restoration Using Swin Transformer," in *ICCVW*, 2021.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *NeurIPS*, 2020.

[3] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced Deep Residual Networks for Single Image Super-Resolution," in *CVPR Workshops*, 2017.

[4] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "SwinIR: Image Restoration Using Swin Transformer," in *ICCV Workshops*, 2021.

[5] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi, "Image Super-Resolution via Iterative Refinement," in *arXiv preprint*, 2021.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *CVPR*, 2022.

[7] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai, "Scene Text Image Super-Resolution in the Wild," in *ECCV*, 2020.

[8] Minxing Luo, Linlong Fan, Qiushi Wang, Ge Wu, Yiyan Luo, Yuhang Yu, Jinwei Chen, Yaxing Wang, Qingnan Fan, and Jian Yang, "Restore Text First, Enhance Image Later: Two-Stage Scene Text Image Super-Resolution with Glyph Structure Guidance," in *arXiv preprint*, 2025.

[9] Kuo-Cheng Wu, Guohang Zhuang, Jinyang Huang, Xiang Zhang, Wanli Ouyang, and Yan Lu, "STAR: A Benchmark for Astronomical Star Fields Super-Resolution," in *arXiv preprint*, 2025.

[10] Eirikur Agustsson and Radu Timofte, "NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study," in *CVPR Workshops*, 2017.

[11] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *CVPR*, 2018.

[12] Baoguang Shi, Xiang Bai, and Cong Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," in *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[13] Yochai Blau and Tomer Michaeli, "The Perception-Distortion Tradeoff," in *CVPR*, 2018.