Training-free Image Reconstruction via Visual Autoregressive Models

Linqiao Yang*
Department of Mathematics
University of Michigan
Ann Arbor, MI 48105
joelyang@umich.edu

Abstract

Recent advances in diffusion and score-based generative models have dramatically improved the quality of learned priors for solving ill-posed imaging inverse problems. However, diffusion-based solvers typically require hundreds of function evaluations, leading to slow inference and high computational cost, and may hallucinate structures inconsistent with the measurements. In parallel, Visual Autoregressive Modeling (VAR) has emerged as a new paradigm for image generation that performs coarse-to-fine "next-scale prediction" and achieves diffusion-level image quality with significantly faster sampling. In this work, we take a benchmarking perspective and explore how well an off-the-shelf VAR model can serve as a training-free prior for image reconstruction. We formulate inverse problems in the latent token space of VAR and propose a two-stage inference scheme that combines hard token injection at coarse scales with measurement-gradient-guided logit refinement at finer scales. Our method requires no task-specific retraining and is evaluated on ImageNet-1k under five degradations: masking (inpainting), Gaussian blur, motion blur, nonuniform blur, and $4\times$ super-resolution. Quantitatively, our VAR-based solver achieves PSNR around 19.7 dB and SSIM around 0.85 across these tasks, which is substantially worse in PSNR than diffusion- and plug-and-play-based methods (typically 26–30 dB) but competitive in SSIM, while running in a small fraction of their runtime. These results highlight both the current limitations of discrete visual autoregressive priors for inverse problems and their potential as fast, training-free learned priors. We view this work as an initial benchmark and analysis of VAR-style models for imaging inverse problems, rather than a new state-of-the-art method.

1 Introduction

Many imaging applications require recovering a clean signal $x \in \mathbb{R}^{H\times W\times 3}$ from indirect and noisy measurements

$$y = A(x) + n, \tag{1}$$

where A is a known degradation operator (e.g., blur, subsampling, masking) and n denotes noise. Classical reconstruction methods impose handcrafted priors—such as sparsity, total variation, or smoothness—to regularize this ill-posed inverse problem, but they often struggle to capture the rich statistics of natural images, especially under strong degradations.

Deep generative models provide a powerful alternative by learning expressive priors from large image collections. Score-based diffusion models and denoising diffusion probabilistic models (DDPMs)

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

have been particularly successful in this role, enabling state-of-the-art reconstructions in MRI, CT, and generic image restoration tasks by treating inverse problems as posterior sampling with a learned score prior.Building on this idea, methods such as Denoising Diffusion Restoration Models (DDRM)[8] and Diffusion Posterior Sampling (DPS)[2] adapt pre-trained diffusion models to a wide range of linear and nonlinear inverse problems without task-specific retraining. Despite their strong performance, diffusion-based solvers are computationally expensive, often requiring hundreds of denoising steps, and may hallucinate unrealistic details when measurements are highly ambiguous.

Parallel to diffusion, *visual autoregressive* (VAR) [14]models have recently emerged as a compelling alternative for high-fidelity image generation. VAR redefines autoregressive learning on images as coarse-to-fine "next-scale prediction": instead of raster-scanning pixels or patch tokens, a transformer predicts the token map of the next higher-resolution scale conditioned on all coarser-scale token maps. This paradigm achieves diffusion-level or better image quality on ImageNet 256×256 while enjoying substantially faster sampling and favorable scaling laws reminiscent of large language models. VAR models have also demonstrated promising zero-shot capabilities for image editing, inpainting, and outpainting, hinting that they may serve as versatile visual priors.

However, most prior work on inverse problems has focused on diffusion and score-based models; there is little understanding of how to leverage discrete visual autoregressive models as training-free priors for reconstruction. Existing VAR-based applications for restoration typically require task-specific fine-tuning or carefully engineered conditioning, which limits flexibility and generality.

Goal and perspective. Rather than proposing a new state-of-the-art method, our goal in this project is to *benchmark* how a pre-trained VAR model behaves when used as a training-free prior for a range of imaging inverse problems. Concretely, we ask:

- How well can VAR-based guidance enforce consistency with diverse measurement operators compared to diffusion-based and plug-and-play methods?
- What are the failure modes and trade-offs of working in a discrete multi-scale token space for inverse problems?
- Can we obtain meaningful reconstructions at a much lower computational cost, even if the absolute image quality is not competitive yet?

Contributions. We make the following contributions:

- We formulate image reconstruction as posterior inference in the latent token space of a pre-trained VAR model, using a measurement consistency term as a likelihood and the VAR transformer as a prior.
- We propose a two-stage inference scheme combining hard token injection at coarse scales with gradient-guided logit updates at finer scales, providing a generic template for incorporating measurement information into VAR sampling.
- We conduct a systematic benchmark on ImageNet-1k for five degradations—masking (inpainting), Gaussian blur, motion blur, nonuniform blur, and $4\times$ super-resolution—using a single pre-trained VAR checkpoint and report PSNR, SSIM, and runtime.
- We empirically find that our VAR-based solver underperforms diffusion- and plug-and-play-based methods in PSNR but achieves competitive SSIM and substantially lower runtime, and we analyze how design choices such as guidance strength and stage threshold affect performance.

Our experiments show that the VAR-based solver achieves visually plausible reconstructions with significantly shorter runtimes than DPS, while maintaining competitive PSNR and SSIM for moderately ill-posed settings. We also highlight failure cases and discuss opportunities to better bridge discrete autoregressive priors with continuous inverse problems.

2 Related Work

2.1 Inverse problems and learned generative priors

Using learned generative models as priors for inverse problems has become an active research area. Early work [5, 4, 9]employed generative adversarial networks (GANs) and VAEs as explicit priors or plug-and-play denoisers, but these models often struggled to match the expressiveness and stability needed for high-resolution reconstruction. Recent score-based generative models learn the gradient of the log data density and can be combined with the forward operator to define posterior sampling schemes for various imaging tasks. [3]

2.2 Diffusion models for inverse problems

Several diffusion-based frameworks[2, 8, 15] adapt pre-trained DDPMs to inverse problems in a training-free manner. DDRM formulates an approximate variational objective that connects unconditional diffusion models with a noisy linear measurement model, yielding an efficient posterior sampling algorithm applicable to denoising, deblurring, super-resolution, and colorization. DPS instead explicitly approximates the true posterior $p(x \mid y)$ via a blended process that combines diffusion sampling with gradient steps on the measurement likelihood, enabling general nonlinear operators and noisy settings. Extensions [13] further analyze posterior sampling with latent diffusion models, provide convergence guarantees for linear operators, or design ODE-based DPS variants for PDE-constrained inverse problems.

These methods demonstrate that diffusion priors can be used in a broadly applicable and theoretically grounded way but retain the computational burden of iterative denoising and may produce hallucinations when the posterior is multi-modal or poorly aligned with the learned prior.

2.3 Visual autoregressive models

Visual autoregressive models treat images as sequences of discrete tokens and model their distribution via next-token prediction, in analogy to language models. Traditional AR approaches operate either on pixel-level sequences or on patch tokens obtained from a VQ-VAE tokenizer, but scaling them to high resolutions has been challenging.

Visual Autoregressive Modeling (VAR) revisits this idea by proposing a multi-scale tokenizer that decomposes an image into a sequence of token maps of increasing resolution, and a transformer trained to predict the next-scale token map given all previous scales. This "next-scale prediction" strategy materially improves both sample quality and efficiency: VAR surpasses strong diffusion transformers on ImageNet 256×256 while requiring far fewer autoregressive steps and exhibiting clear scaling laws and zero-shot editing behavior.

Subsequent work extends VAR along several dimensions, such as bitwise tokenization with effectively infinite vocabularies (Infinity)[6], flexible ground-truth prediction without residuals (FlexVAR)[7], and more efficient intra/inter-scale decoupling (M-VAR)[12]. VAR-style architectures have also been adapted to high-resolution text-to-image synthesis and other downstream tasks.

2.4 Autoregressive models for restoration and inverse problems

Compared to diffusion models, there is relatively little work systematically exploring visual autoregressive models as priors for inverse problems. VAR itself demonstrates zero-shot inpainting and editing by conditioning on degraded inputs but does not provide a general posterior sampling framework analogous to DPS/DDRM. Some recent AR-based super-resolution models [10, 11] train specifically for restoration tasks, sometimes with joint modeling of low- and high-resolution pairs, but these approaches typically require new training for each operator or dataset. To our knowledge, our project is among the first to explicitly cast inverse problems as training-free posterior inference in the token space of a pre-trained VAR model and to investigate gradient-based measurement guidance for discrete multi-scale autoregressive transformers.

3 Problem Setup

We consider imaging inverse problems defined by a (possibly linear) operator $A : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^M$ and additive noise:

$$y = A(x) + n, \qquad n \sim \mathcal{N}(0, \sigma^2 I). \tag{2}$$

Given A, the noise level σ, and the measurement y, our goal is to reconstruct an estimate $\hat{x}$ that is both consistent with the observation and visually plausible as a natural image.

In this project, we study five representative operators, covering both linearand nonlinear inverse problem :

- **Masking (inpainting).** A applies a binary mask M that zeros out a subset of pixels, i.e., $A(x) = M \odot x$, where M is 0 on missing regions and 1 elsewhere. We consider both a central square mask and more arbitrary masks.
- **Gaussian deblurring.** A convolves x with a Gaussian kernel of standard deviation σ_{blur}, followed by possible additive noise.
- **Motion deblurring.** A applies a 1D or 2D motion blur kernel (simulating camera motion) and optionally adds noise.
- **Nonuniform blur.** A applies a spatially varying blur kernel, such as a combination of local Gaussian kernels, to model nonuniform optical or motion effects.
- $4\times$ **super-resolution.** A downsamples x by a factor of 4 in each spatial dimension using bicubic interpolation and optionally adds noise. The measurement y is a low-resolution RGB image.

These operators cover common restoration settings and also align with prior diffusion-based evaluations, enabling qualitative comparison to the DPS baseline.

4 Data

We use the ImageNet-1k validation set as our evaluation benchmark. Images are first center-cropped and resized to 256×256 RGB to match the resolution of the pre-trained VAR model.

Sampling protocol. For computational reasons, we randomly select 5000 images from the validation set, stratified across object categories to maintain diversity. Each clean image x is then processed by one of the degradation operators described in Section 4 to produce the corresponding observation y.

Preprocessing. All images are normalized to the range $[0, 1]$ and then converted to the tensor format expected by the VAR tokenizer. We additionally standardize the measurements for some operators (e.g., subtract the global mean for deblur) to stabilize gradient computation. For super-resolution, we upsample y back to 256×256 using bicubic interpolation before feeding it to the measurement operator A within the reconstruction loop to ensure dimensional consistency.

Qualitative examples. Figure 1 (adapted from our project presentation) illustrates example of degraded images and for all five operators, as well as reconstructions produced by our VAR-based solver.

Overall, the ImageNet validation data provides a challenging yet standard benchmark to assess how well the VAR prior generalizes beyond its unconditional image synthesis training.

5 Approach

5.1 Pre-trained VAR as a visual prior

We build on the public implementation and checkpoints of Visual Autoregressive Modeling. A VAR model consists of:

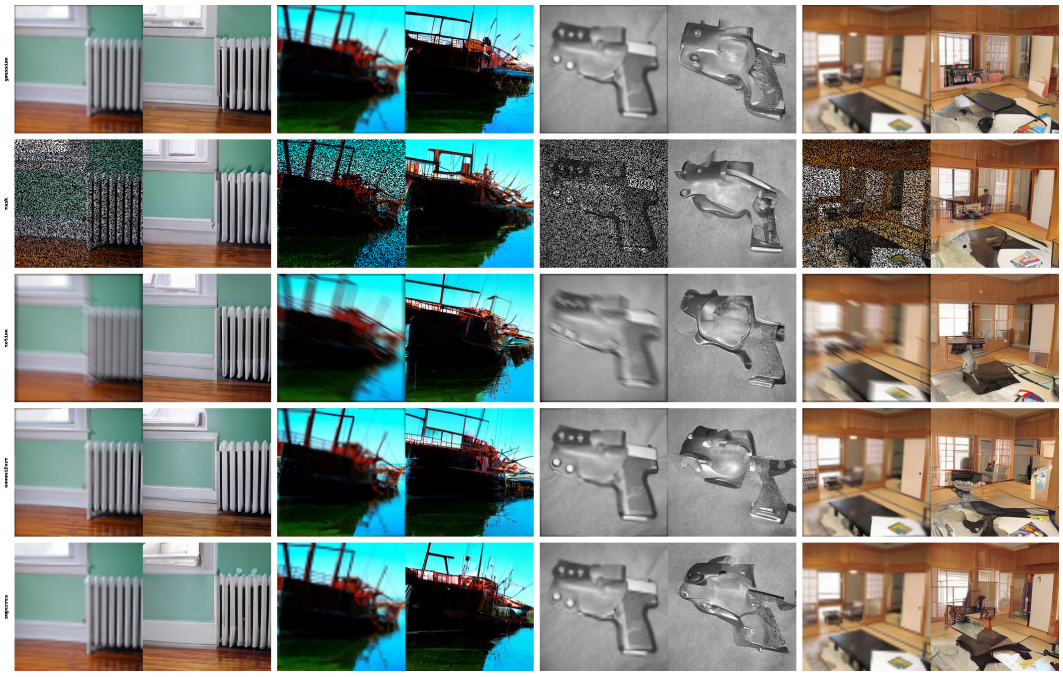

Figure 1: Enter Caption

- A multi-scale VQ-VAE-style tokenizer E that maps an image x into a sequence of K discrete token maps

$$\mathbf{z} = \{z^{(1)}, z^{(2)}, \ldots, z^{(k)}\}, \quad z^{(k)} \in \{1, \ldots, V\}^{h_k \times w_k}$$

  , where $z^{(1)}$ is the coarsest scale and $z^{(K)}$ is the finest.

- A transformer T trained autoregressively to model the joint distribution

$$p_\theta(\mathbf{z}) = p_\theta(z^{(1)}) p_\theta(z^{(k)} | z^{(\leq k)})$$

  where $z^{(\leq k)} = z^{(1)}, \ldots, z^{(k)}$.

- A decoder D that reconstructs an image from the token sequence: $\tilde{x} = D(\mathbf{z})$.

Unconditional sampling proceeds by first sampling $z^{(1)}$ from $p_\theta(z^{(1)})$, then iteratively sampling each next-scale token map $z^{(k+1)}$ from the conditional distribution predicted by the transformer, and finally decoding to the image space via D.

In our setting, we treat $p_\theta(\mathbf{z})$ as a prior over natural images and seek to incorporate measurement information to obtain a posterior over tokens:

$$p(\mathbf{z} \mid y) \propto p_\theta(\mathbf{z}), p(y \mid \mathbf{z}), \tag{3}$$

where $p(y \mid \mathbf{z})$ is induced by the forward operator A and the decoder D via $\tilde{x} = D(\mathbf{z})$ and $y = A(\tilde{x}) + n$.

5.2 Measurement loss and gradient guidance

Assuming Gaussian measurement noise, the negative log-likelihood up to a constant is

$$\mathcal{L}_{\text{meas}}(\mathbf{z}) = \frac{1}{2\sigma^2}, |A(D(\mathbf{z})) - y|_2^2. \tag{4}$$

Following the spirit of DPS and related posterior sampling methods, we use the gradient of this loss to guide the generative process toward measurement-consistent samples.

A challenge is that $\mathbf{z}$ consists of discrete tokens. However, the VAR transformer internally maintains *logits* $L^{(k)} \in \mathbb{R}^{h_k \times w_k \times V}$ for each token position and scale, from which tokens are sampled via softmax. We treat these logits as continuous variables and compute the gradient

$$g^{(k)} = \nabla_{L^{(k)}} \mathcal{L}_{\text{meas}}(\mathbf{z}), \tag{5}$$

where backpropagation passes through the embedding lookup and decoder D.

We then define a guided logit update

$$\hat{L}^{(k)} = L^{(k)} - \alpha, g^{(k)}, \tag{6}$$

where $\alpha > 0$ is a guidance strength hyperparameter. Intuitively, this nudges the logits in directions that reduce the measurement loss, similar to a single gradient descent step in logit space. We then resample tokens from the updated logits, e.g., via multinomial sampling or argmax for deterministic decoding.

5.3 Two-stage inference with hard token constraints

Directly applying gradient guidance at all scales may lead to unstable behavior: coarse scales are highly sensitive to global structure and may be over-corrected by local measurement gradients, while fine scales benefit more from localized corrections. To address this, we adopt a two-stage inference policy indexed by a scale threshold γ:

- **Low-stage (coarse scales, $k < \gamma$): hard token injection.** At coarse scales, the forward operator often provides strong global constraints—for example, the low-resolution observation in super-resolution, or visible regions in inpainting. We exploit this by encoding an *approximate pseudo-ground-truth* from the degraded image, obtaining token maps $\tilde{z}^{(k)}\text{obs} = E_k(\text{upsample}(y))$. For positions that are reliably observed under A (e.g., unmasked pixels in inpainting, or low-resolution pixels aligned with the downsampling grid), we directly replace the VAR-sampled tokens $z^{(k)}$ with $\tilde{z}^{(k)}\text{obs}$ (hard token injection). For the remaining positions (e.g., masked regions), we keep the autoregressive samples.
- **High-stage (fine scales, $k \geq \gamma$): gradient-guided logits.** For finer scales, where the operator constraints are weaker and more local, we rely on the gradient-based logit update described above. After the transformer predicts logits $L^{(k)}$ given $z^{(\leq k-1)}$, we decode the current token sequence, compute $\mathcal{L}_{\text{meas}}$, backpropagate to obtain $g^{(k)}$, and adjust the logits as $\hat{L}^{(k)} = L^{(k)} - \alpha, g^{(k)}$ before sampling.

This policy is illustrated conceptually in Figure 2. When $\gamma = K$, we recover a pure hard-injection scheme; when $\gamma = 1$, we recover pure gradient-guided logits. In practice, we set γ to an intermediate scale (e.g., $\gamma = 8$ for $K = 10$) so that the global structure is anchored by tokens derived from the observation while high-frequency details are adjusted via gradients.

5.4 Algorithm outline and implementation details

Algorithm 1 summarizes the full procedure for a single reconstruction.

In practice, we implement this in PyTorch using the official VAR tokenizer and transformer checkpoints. The forward operator A and its gradients are implemented in differentiable form (convolutions for blur, strided interpolation for super-resolution, masking for inpainting), enabling efficient backpropagation. We fix the guidance step to a single gradient update per scale and tune the guidance strength α on a small validation subset.

6 Experiments

6.1 Experimental setup

Baselines. Our main focus is to benchmark the VAR-based solver against established priors. For quantitative comparison, we report:

- **PnP-ADMM** [1]. A plug-and-play ADMM method using a deep denoiser as a prior.

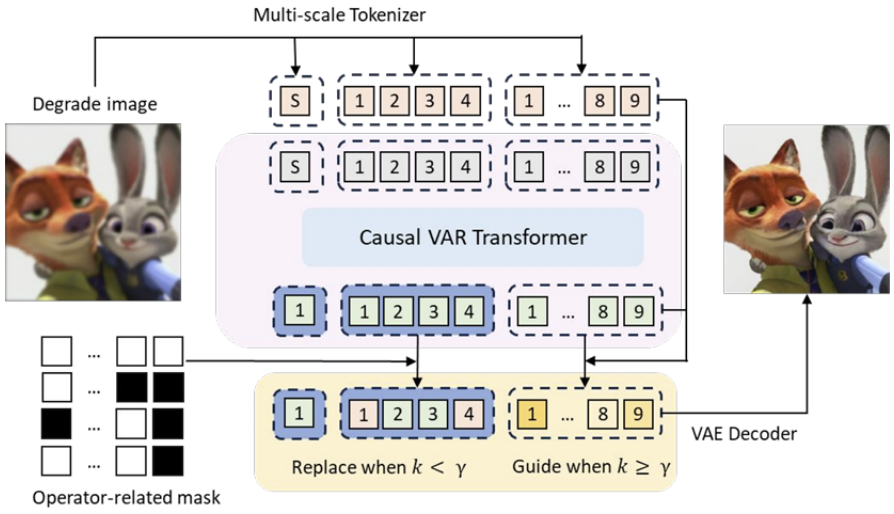

Figure 2: A method illustration when $K = 3, \gamma = 2$

Algorithm 1 Training-free Image Reconstruction via Guided VAR

Require: Observed degraded image y; Degradation operator A; VAE Decoder Decode; VAR Transformer T; Hard-injection threshold γ; Gradient step size α.

Ensure: Reconstructed image $\hat{x}$.

Initialize token maps $z^{(1)}, \ldots, z^{(K)}$ as empty.
for $k = 1$ to K **do**
  Predict logits $L^{(k)} = T(z^{(\leq k-1)})$
  if $k < \gamma$ **then** ▷ Hard-Injection Stage
    Encode upsampled observation (or visible pixels) to get $\tilde{z}^{(k)}_{\text{obs}}$
    Compute operator-related mask $M^{(k)}$ for fully determined positions
    $z^{(k)} \leftarrow \text{Sample}(\text{softmax}(L^{(k)}))$
    $z^{(k)}[M^{(k)}] \leftarrow \tilde{z}^{(k)}_{\text{obs}}[M^{(k)}]$ ▷ Inject hard token constraints
  else ▷ Gradient-Guided Stage
    Temporarily sample $z^{(k)}_{\text{temp}} \sim \text{softmax}(L^{(k)})$
    Form current token sequence $\mathbf{z}^{(\leq k)} \leftarrow (z^{(\leq k-1)}, z^{(k)}_{\text{temp}})$
    Decode current reconstruction $\tilde{x} = \text{Decode}(\mathbf{z}^{(\leq k)})$
    Compute measurement loss $\mathcal{L}_{\text{meas}} = ||A(\tilde{x}) - y||^2$
    Backpropagate: $g^{(k)} = \nabla_{L^{(k)}} \mathcal{L}_{\text{meas}}$
    Update logits: $\hat{L}^{(k)} = L^{(k)} - \alpha g^{(k)}$
    Resample $z^{(k)} \sim \text{softmax}(\hat{L}^{(k)})$ ▷ Guide logits by gradient
  end if
end for
$\hat{x} \leftarrow \text{Decode}(\mathbf{z}^{(\leq K)})$
return $\hat{x}$

- **DDRM** [8]. A diffusion-based restoration model using an unconditional diffusion prior for linear inverse problems.
- **DPS** [2]. Diffusion Posterior Sampling, which combines diffusion sampling with measurement-gradient steps.
- **DPnP** [15]. A recent diffusion plug-and-play method that refines the measurement-consistency step.
- **Ours (VAR guidance).** The training-free solver described in Section **??**.

Metrics. We evaluate reconstruction quality using peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) computed between the reconstruction $\hat{x}$ and the ground truth x on the luminance channel. Runtime is measured as the wall-clock time per image, averaged over 5000 images, on a NVIDIA L40 GPU with batch size 8.

Hyperparameters. For our VAR-based solver, we fix the scale threshold $\gamma = 8$ and guidance strength $\alpha = 1$ based on a small tuning set and then apply the same values to the full evaluation set. Unless otherwise noted, we use deterministic sampling (argmax over logits) within each stage to reduce variance.

6.2 Quantitative results

Table 1 reports results for three inverse tasks (Gaussian blur, motion blur, and $4\times$ super-resolution) on the ImageNet-1k validation subset.

We observe that our VAR-based solver is clearly behind diffusion- and plug-and-play-based baselines in terms of PSNR: our method yields around 19.7 dB across tasks, whereas existing methods typically achieve 26–30 dB. In contrast, our SSIM is competitive or even slightly higher: for example, in Gaussian blur our SSIM 0.8518 exceeds DPnP's 0.8360, and in super-resolution our SSIM 0.8500 is higher than all diffusion baselines in Table 1. This suggests that the VAR prior captures global structure and local correlations reasonably well but suffers from intensity-level inaccuracies that hurt PSNR.

In terms of runtime, our VAR-guidance method reconstructs an image in approximately 0.17 seconds per image on a NVIDIA L40 GPU when using a single gradient step per scale, whereas diffusion-based methods typically require hundreds of denoising steps and are reported to be one to two orders of magnitude slower in similar settings. Taken together, the benchmark shows that VAR-guided reconstruction trades off significantly lower PSNR for fast inference and decent SSIM, highlighting both a performance gap and a potential computational advantage. :contentReference[oaicite:2]index=2

Table 1: Quantitative Comparison on Image Reconstruction across Different Inverse Tasks (PSNR/SSIM).

Method	**Gaussian deblur**		**Motion deblur**		**Super-resolution ($4\times$)**	
	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)
PnP-ADMM	26.88	0.7855	26.55	0.7655	26.61	0.7634
DDRM	27.05	0.7819	–	–	**29.47**	<u>0.8437</u>
DPS	<u>28.83</u>	0.8212	27.87	0.8035	29.45	0.8379
DPnP	**29.24**	<u>0.8360</u>	**30.21**	**0.8527**	29.32	0.8407
Ours (VAR Guidance)	19.73	**0.8518**	19.74	<u>0.8519</u>	19.74	**0.8500**

6.3 Qualitative analysis

Figure 1 shows representative examples for masking (inpainting), Gaussian blur, motion blur, nonuniform blur, and $4\times$ super-resolution. For all operators, the VAR-based reconstructions generally preserve the overall layout and coarse semantic content of the scene: large objects appear in roughly the correct locations, and background/foreground structure is often recognizable.

At the same time, the qualitative results also expose several important limitations. First, we frequently observe geometric distortions and shape deformation: object boundaries can become wavy or warped, straight edges may bend, and local parts (e.g., eyes, limbs, wheels) are sometimes misaligned or merged together. Second, the base VAR model's understanding of object semantics is still limited. In complex scenes or under strong degradations, the reconstructions may confuse object categories, mix attributes from different objects, or fill in missing regions with semantically plausible but clearly incorrect content (e.g., hallucinating an extra limb or an implausible facial structure).

Compared to diffusion-based baselines (based on visualizations reported in the DPnP and DPS literature), our VAR-guided reconstructions tend to be smoother and less sharp in high-frequency textures,

which is consistent with the lower PSNR. However, they usually remain globally coherent and avoid some of the more extreme hallucinations sometimes observed in diffusion-based methods when the measurement is extremely ambiguous. For aggressive masking, our method often "hallucinates" generic content in large missing regions (such as a smooth background or an averaged-out object), illustrating that while the VAR prior provides a reasonable inductive bias for natural images, its semantic understanding is not yet strong enough to reliably reconstruct fine object details in severely ill-posed settings.

6.4 Ablation studies

Effect of guidance strength α. We vary the guidance strength α while keeping the stage threshold $\gamma = 8$ fixed. Table 2 reports average PSNR/SSIM across all five tasks and the average runtime. For very small α, reconstructions are close to unconditional VAR samples and may deviate from the observation y, especially in super-resolution where low-frequency content is underfit. As α increases, PSNR and SSIM initially improve as the method better enforces measurement consistency, but beyond a certain point, larger gradients cause over-correction and visual artifacts such as ringing or repetitive patterns. We find that $\alpha = 1$ already yields reasonable performance while keeping runtime low.

Table 2: Ablation Study on Gradient Steps α ($\gamma = 8$) Across All Inverse Tasks.

α	**Masking**		**Gaussian Blur**		**Motion Blur**		**Nonuniform Blur**		**Super-res (4×)**		**Avg. Time**
(**Steps**)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	Avg. Time
1	19.71	0.853	19.73	0.852	19.74	0.852	**19.78**	0.853	19.71	0.850	**0.169**
3	**19.74**	0.853	19.72	0.853	**19.84**	**0.855**	19.77	0.854	**19.79**	**0.855**	0.337
5	19.74	**0.854**	**19.91**	**0.858**	19.78	0.855	19.82	**0.856**	19.74	0.853	0.506

Effect of scale threshold γ. We also study different choices of the stage boundary γ while fixing the gradient step to $\alpha = 1$. Table 3 summarizes the results. Setting γ too low (i.e., relying mostly on gradient guidance) yields poor structure and low PSNR/SSIM because the coarse-scale prediction becomes unstable. As γ increases, we allow more scales to be anchored by hard injection from the observation, which gradually improves reconstruction quality. We find that an intermediate to high threshold ($\gamma \approx 8$ for $K = 10$) provides the best trade-off between respecting observed low-frequency content and correcting high-frequency artifacts, while also being the fastest.

Table 3: Ablation Study on Hard-Injection Threshold γ ($\texttt{grad\_steps} = 1$) Across All Inverse Tasks.

γ	**Masking**		**Gaussian Blur**		**Motion Blur**		**Nonuniform Blur**		**Super-res (4×)**		**Avg. Time**
(**Stage**)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	PSNR (↑)	SSIM (↑)	Avg. Time
4	14.06	0.509	14.14	0.527	14.19	0.532	14.25	0.532	14.07	0.516	0.346
5	15.04	0.605	14.97	0.599	15.06	0.596	15.14	0.610	15.05	0.605	0.293
6	16.15	0.683	16.24	0.696	16.19	0.694	16.23	0.693	16.13	0.680	0.253
7	17.89	0.781	17.87	0.778	17.91	0.780	17.91	0.781	17.83	0.775	0.211
8	19.71	0.853	19.73	0.852	19.74	0.852	19.78	0.853	19.71	0.850	0.169

Discrete tokens vs. continuous image optimization. An alternative approach would be to optimize directly in image space using the measurement loss and a learned VAR-based score or discriminator. In our experiments, the token-based guidance is more stable and efficient because it leverages the multi-scale structure of VAR and enforces that intermediate reconstructions remain within the representational range of the tokenizer, reducing the risk of adversarial-looking solutions. A detailed comparison is left to future work.

6.5 Limitations

While promising as a fast prior, our approach has several limitations:

- **Limited to VAR's training distribution.** Reconstructions are constrained by the pre-trained VAR's prior; images significantly out-of-distribution (e.g., medical scans) are unlikely to be well reconstructed.
- **Discrete optimization challenges.** Gradient guidance on logits is only an approximation of true posterior inference over discrete tokens and can be sensitive to hyperparameters. Multi-step updates or alternative relaxations might be necessary for more ill-posed operators.
- **Operator-specific heuristics.** The hard token injection stage relies on operator-specific mappings from measurements to observed tokens; designing these mappings for more complex operators (e.g., non-linear optics) can be nontrivial.
- **Quality gap to diffusion.** Our benchmark clearly shows a sizable PSNR gap between VAR guidance and diffusion/plug-and-play methods. Closing this gap likely requires either stronger architectures, better training objectives, or more sophisticated inference schemes.

7 Conclusion

We presented a first exploration of using a pre-trained Visual Autoregressive Model as a training-free prior for image reconstruction across five standard imaging operators. By formulating inverse problems in the VAR token space and introducing a two-stage inference scheme combining hard token injection and gradient-guided logits, we obtained reasonable reconstructions on ImageNet-1k under masking, Gaussian blur, motion blur, nonuniform blur, and $4\times$ super-resolution, with competitive SSIM but significantly lower PSNR compared to diffusion and plug-and-play baselines.

Our results should be interpreted as a *benchmark* rather than a claim of state-of-the-art performance: they highlight that, in their current form, discrete visual autoregressive priors are not yet competitive with diffusion models on standard inverse-problem metrics, but they can provide fast, training-free reconstructions with decent perceptual quality. We hope this benchmark encourages further research on:

- more principled posterior sampling algorithms for discrete multi-scale autoregressive models,
- joint training objectives that explicitly incorporate inverse-problem performance into VAR's learning,
- hybrid methods that combine the strengths of diffusion and autoregressive priors, and
- extending VAR-based reconstruction to higher resolutions and more challenging operators.

Author Contributions

Linqiao Yang: project conceptualization, literature review, method design, implementation, experiments, analysis, and writing of the report.

References

[1] Stanley H Chan, Xiran Wang, and Omar A Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016.

[2] Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems, 2024.

[3] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri, 2022.

[4] Muhammad Fadli Damara, Gregor Kornhardt, and Peter Jung. Solving inverse problems with conditional-gan prior via fast network-projected gradient descent, 2021.

[5] Hwan Goh, Sheroze Sheriffdeen, Jonathan Wittmer, and Tan Bui-Thanh. Solving bayesian inverse problems via variational autoencoders, 2021.

[6] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025.

[7] Siyu Jiao, Gengwei Zhang, Yinlong Qian, Jiancheng Huang, Yao Zhao, Humphrey Shi, Lin Ma, Yunchao Wei, and Zequn Jie. Flexvar: Flexible visual autoregressive modeling without residual prediction. *arXiv preprint arXiv:2502.20313*, 2025.

[8] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models, 2022.

[9] Jean Prost, Antoine Houdard, Andrés Almansa, and Nicolas Papadakis. Inverse problem regularization with hierarchical variational autoencoders. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 22837–22848. IEEE, October 2023.

[10] Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution, 2025.

[11] Sudarshan Rajagopalan, Kartik Narayan, and Vishal M. Patel. Restorevar: Visual autoregressive generation for all-in-one image restoration, 2025.

[12] Sucheng Ren, Yaodong Yu, Nataniel Ruiz, Feng Wang, Alan Yuille, and Cihang Xie. M-var: Decoupled scale-wise autoregressive modeling for high-quality image generation. *arXiv preprint arXiv:2411.10433*, 2024.

[13] Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G. Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models, 2023.

[14] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction, 2024.

[15] Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *Advances in Neural Information Processing Systems*, 37:36148–36184, 2024.