ALIGN & INVERT: SOLVING INVERSE PROBLEMS WITH DIFFUSION AND FLOW-BASED MODELS VIA REPRESENTATIONAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Enforcing the alignment of internal representations in diffusion or flow-based models with those from pretrained self-supervised models during training has recently been shown to provide a powerful inductive bias, significantly improving both convergence speed and the quality of generated images. In this paper, we move beyond training and instead focus on inverse problems, where pretrained diffusion or flow-based models are used as *priors*. We propose applying *representational alignment* of diffusion/flow-based models with a *pretrained self-supervised visual encoder*, such as DINOV2, for guidance at inference time. Although the ground truth signal is unavailable in inverse problems, we show that representational alignment based on approximations of the ground truth can yield significant gains, steering the reverse process of diffusion/flow-based models toward higher-quality reconstructed images. We provide theoretical insights by uncovering a connection between representational alignment and perceptual metrics. Under mild assumptions, we further show that our approach can improve the perception–distortion trade-off frontier, i.e., enhance perceptual quality with negligible degradation in distortion. Finally, we demonstrate its versatility by integrating it into multiple state-of-the-art solvers. Extensive experiments on super-resolution, box inpainting, Gaussian deblurring, and motion deblurring show that our proposed method consistently enhances reconstruction quality.

1 INTRODUCTION

Pretrained diffusion and flow-based models, (Sohl-Dickstein et al., 2015; Ho et al., 2020; Esser et al., 2024; Lipman et al., 2023), have recently been at the heart of methods focusing on addressing various types of inverse problems. The crux of these approaches is to perform diffusion sampling, incorporating measurement information during the reverse process, with the goal to reconstruct the clean image at the final step, Patel et al. (2024); Chung et al. (2023). Undoubtedly, diffusion and flow-based models have pushed the boundaries in addressing a wealth of inverse problems, Daras et al. (2024). However, they still struggle to produce highly detailed images particularly in cases of severe degradation and in challenging scenes, such as natural images with rich textures and complex backgrounds. These issues are further exacerbated when using latent diffusion models (Rombach et al., 2022), where the inclusion of the encoder-decoder architecture introduces additional challenges during the reverse process due to the nonlinearities introduced. To mitigate these limitations, additional inductive biases are needed during inference, Rout et al. (2023).

Recent studies have shown that diffusion models learn semantic features in their hidden states, (Li et al., 2023). The more expressive these features are the better the diffusion model performs of on the generative task, (Xiang et al., 2023). Building on this insight, the seminal work of Yu et al. (2024) introduced a regularizer that aligns the internal representations of the diffusion model with those of a pretrained visual encoder, (Oquab et al., 2023). This framework, termed *representational alignment* (REPA), was shown to act as a strong semantic constraint leading to higher-fidelity generations and significantly faster convergence.

The success of REPA has attracted considerable interest in this direction, Wang et al. (2025b); Tian et al. (2025); Yao et al. (2025); Leng et al. (2025); Wang et al. (2025a). In particular, (Wang et al.,

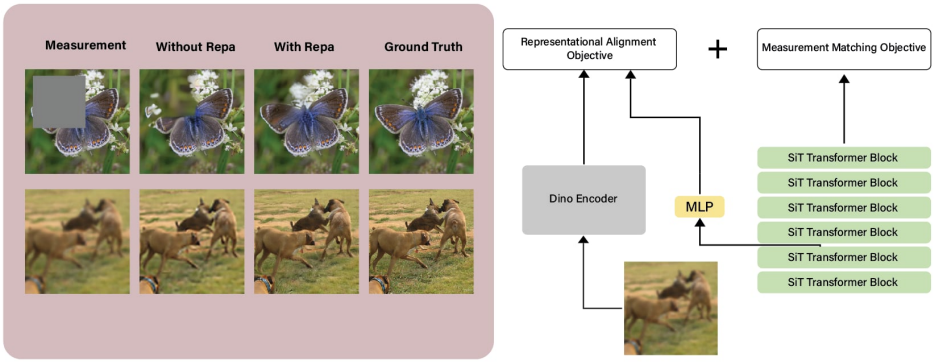


Figure 1: Overview of our proposed framework. Left: Box inpainting (top row) and Gaussian deblurring (bottom row) results, where adding REPA (3rd column) improves perceptual quality of reconstructed images over the non-REPA approach (2nd column). Right: Alignment mechanism between diffusion model features and pretrained DINO embeddings, combined with measurement matching. An approximate $\hat{x}_0$ is given as input to DINO encoder.

2025b) observed that while REPA enhances early training dynamics, its benefits diminish over time and can even degrade performance in later iterations. To address this issue, they introduced an early-stopping strategy which mitigated the late-stage degradation and further improved overall convergence speed. Other works have explored applying alignment losses directly to VAE latent spaces Yao et al. (2025); Xu et al. (2025), finding that diffusion models trained on these regularized spaces converge faster. Despite these advances, most efforts focus on improving training efficiency, whereas relatively little attention has been paid to leveraging representation alignment during inference, particularly for solving inverse problems. In this paper, we aim to address the following question:

Can we apply representational alignment ideas to benefit existing algorithms for solving inverse problems using pretrained diffusion models?

Contributions. Our contributions can be summarized as follows:

- We propose a new approach for solving inverse problems with latent diffusion and flow-based models enforcing alignment between internal diffusion features and DINOv2 representations. As shown, even without ground-truth images, alignment with approximate reconstructions remains meaningful due to the robustness of the DINOv2 encoder (see Fig. 1).
- We provide theoretical insights linking our REPA approach to perceptual metrics, such as maximum mean discrepancy, and show it improves the frontier of the perception–distortion trade-off.
- We demonstrate versatility by integrating representational alignment into existing state-of-the-art inverse algorithms.
- Empirical results highlight improvements over prior work on super-resolution, Gaussian and motion deblurring, and box inpainting.

2 Related Work

2.1 Diffusion & Flow-based Models

Generative models based on denoising are a class of deep generative models that aim to learn a data distribution $p_0(x)$ from a finite set of samples, enabling the generation of new, realistic data, (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Liu et al., 2023). These models define a) a forward (noising) process, which gradually interpolates between the data distribution and a reference distribution $p_1(x)$ (typically a Gaussian), and b) a reverse (denoising) process, which is trained to invert this corruption and iteratively reconstruct data from noise. Following the formulation of stochastic interpolants Albergo et al. (2023), this process can be expressed as

$$x_t = \alpha_t x^* + \sigma_t \varepsilon, \qquad x^* \sim p_0(x),\ \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

where α_t is a decreasing function of t and σ_t is an increasing function of t, with boundary conditions $\alpha_0 = \sigma_1 = 1$ and $\alpha_1 = \sigma_0 = 0$. Throughout, we use the notation $\dot{\alpha}_t := \frac{\mathrm{d}\alpha_t}{\mathrm{d}t}$ and $\dot{\sigma}_t := \frac{\mathrm{d}\sigma_t}{\mathrm{d}t}$ to denote time derivatives. Lipman et al. (2023) showed that there exists a probability flow ordinary differential equation whose solution follows the same time marginals as the process in equation 1:

$$\dot{x}_t = v(x_t, t), \tag{2}$$

where $v(x_t, t)$ denotes the true drift field that transports $p_1(x)$ to $p_0(x)$. To enable sampling from the data distribution, one has to train a neural network $v_\theta(x_t, t)$ to approximate $v(x_t, t)$. This is done by regressing the network output toward the optimal vector field using *conditional flow matching* Lipman et al. (2023); Hyvärinen & Dayan (2005), minimizing

$$\mathcal{L}_{\text{velocity}}(\theta) := \mathbb{E}_{x^*,\varepsilon,t}\big[\|v_\theta(x_t, t) - \dot{\alpha}_t x^* - \dot{\sigma}_t \varepsilon\|^2\big], \tag{3}$$

Beyond the ODE view, there also exists a reverse *stochastic differential equation* (SDE) with the same time marginals:

$$\mathrm{d}x_t = v(x_t, t)\,\mathrm{d}t - \tfrac{1}{2} g^2(t)\, s(x_t, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{w}_t, \tag{4}$$

where $s(x_t, t) = \nabla_{x_t} \log p_t(x_t)$ is the score, $g(t)$ is the diffusion coefficient, and $\bar{w}_t$ is a standard Wiener process running backward in time. Notably, the score can be expressed directly in terms of the velocity field:

$$s(x_t, t) = \sigma_t^{-1} \cdot \frac{\alpha_t v(x_t, t) - \dot{\alpha}_t x_t}{\dot{\alpha}_t \sigma_t - \alpha_t \dot{\sigma}_t}. \tag{5}$$

This identity shows that the score parameterization used in diffusion models and the velocity parameterization used in flow-based models are mathematically equivalent. In practice, diffusion models correspond to a stochastic discretization of the probability flow ODE, while flow-based models correspond to its deterministic counterpart. Both share the same time marginals but differ in the choice of sampler—SDE-based for diffusion, ODE-based for flow

2.2 Diffusion Models with Representation Guidance

Diffusion models learn rich, discriminative features in their hidden states, which are crucial for their generative performance Xiang et al. (2023). Nevertheless, their learned representations still fall behind on downstream tasks when compared with state-of-the-art self-supervised visual encoders, Assran et al. (2023); Siméoni et al. (2025). This representation gap has been identified as a key bottleneck for improving diffusion model training, which has spurred recent efforts to improve their latent space representations Yu et al. (2024). To address this, a growing line of work introduces representation guidance, which conditions diffusion models on external representations from pretrained self-supervised encoders Li et al. (2024). These approaches fall into two categories: (i) conditioning the diffusion model on external representations, either learned, Li et al. (2024), or retrieved from a database of samples, Blattmann et al. (2022); Hu et al. (2023); Sheynin et al. (2022); and (ii) explicitly aligning the model's internal representations with those of large pretrained encoders, thereby encouraging more semantically meaningful feature spaces, Tian et al. (2025); Hu et al. (2023); Leng et al. (2025).

Our work builds on this second family of approaches, which have shown promise for improving both sample quality and training efficiency by encouraging semantically meaningful representations to emerge inside the diffusion model. In particular, Yu et al. (2024) introduced Representation Alignment (REPA), a method that enforces feature-level alignment in a transformer-based diffusion model, (Ma et al., 2024). REPA maximizes the patch-wise similarity between the DINO representation of the target image x, i.e., $f_{\text{DINO}}^{[n]}(x) \in \mathbb{R}^{D_1}$, where $n = 1, 2, \ldots, N$ corresponds to patch index, and the intermediate hidden states $h_t^{[n]} \in \mathbb{R}^{D_2}$ of the diffusion model via maximizing the following term:

$$\textsc{Repa}\left(x, \text{DiffDecoder}(h_t)\right) = \mathbb{E}_{x,t,\epsilon}\left[\frac{1}{N}\sum_{n=1}^{N} \cos\left(f_{\text{DINO}}^{[n]}(x), g_\phi(h_t^{[n]})\right)\right], \tag{6}$$

where $g_\phi : \mathbb{R}^{D_2} \to \mathbb{R}^{D_1}$ is a learnable multi-layer perceptron (MLP) with parameters ϕ that projects hidden states into the same embedding space as $f(x)$, and D_1, D_2 are the corresponding embedding dimensions used. By jointly optimizing $\textsc{Repa}\left(x, \text{DiffDecoder}(h_t)\right)$ with the standard diffusion loss, the model learns internal representations that are semantically richer and better aligned with high-level visual features, ultimately yielding higher-quality image generation and faster convergence during training.

2.3 Solving Inverse Problems With Diffusion- and Flow-based Models

The goal of inverse problems is to recover the clean signal x_0 from noisy or degraded measurements y. In the flow- and diffusion-based framework, this is accomplished by replacing the unconditional score function in equation 4 with the conditional score $\nabla_{x_t} \log p(x_t \mid y)$. Applying Bayes' rule, the reverse-time SDE becomes

$$dx_t = \Big(f(t)x_t - g^2(t)\big[\nabla \log p_t(x_t) + \nabla \log p_t(y \mid x_t)\big]\Big)dt + g(t)d\bar{w}_t, \tag{7}$$

The term $\nabla \log p_t(x_t)$ is the unconditional score and can be computed using a pretrained diffusion model. Intuitively, this term guides x_t toward the learned data distribution. The term $\nabla_{x_t} \log p(y \mid x_t)$ enforces consistency with the measurements. However, computing $p(y \mid x_t)$ is generally intractable, as it requires marginalizing over all possible clean signals x_0

As a result, practical algorithms rely on approximations of $\nabla_{x_t} \log p(y \mid x_t)$. A recent survey by Daras et al. (2024) provides a comprehensive taxonomy of these methods.One of the most prominent gradient-based method is ***Diffusion Posterior Sampling (DPS)*** method, (Chung et al., 2023), which uses the following approximation,

$$p(y \mid x_t) = \int p(y|x_0)p(x_0 \mid x_t)\, dx_0 \approx p\left(y \mid \hat{x}_0 = \mathbb{E}[x_0 \mid x_t]\right). \tag{8}$$

In many practical scenarios, performing the diffusion process directly in pixel space is computationally prohibitive. Latent diffusion models (LDMs) address this limitation by operating in a compressed latent space. Let $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}^d$ be an encoder mapping an image x_0 to its latent representation $z_0 = \mathcal{E}(x_0)$, and let $\mathcal{D}$ denote the corresponding decoder such that $\hat{x}_0 = \mathcal{D}(z_0) \approx x_0$. Building on this formulation, recent work has proposed latent-space counterparts of pixel-based inverse problem solvers. For example, Diffusion Posterior Sampling (DPS) has been formulated in latent space by redefining the measurement likelihood as

$$p(y \mid z_t) \approx p\big(y \,\big|\, \hat{x}_0 = \mathcal{D}\big(\mathbb{E}[z_0 \mid z_t]\big)\big). \tag{9}$$

This straightforward formulation—commonly referred to as Latent DPS has been observed to introduce artifacts, largely due to the nonlinearity of the decoder. To mitigate this effect, Rout et al. (2023) proposed a regularization term encouraging the latent variables to remain close to a fixed point of the encoder–decoder mapping, while Song et al. (2024) enforced data consistency during optimization followed by a projection of the latent variables back onto the noisy manifold.

3 Proposed Approach

Given a noisy observation $y \in \mathbb{R}^m$ of an unknown signal $x_0 \in \mathbb{R}^n$, our goal is to sample from $p_\theta(x_0 \mid y)$ using a pretrained diffusion- or flow-based model.

The main challenge that we should address in applying representational alignment to inverse problems is to find what is the best input to the DINO, given the fact that the ground truth signal is not available. Specifically, we need access to an approximate $\hat{x}_0$ of the ground truth x_0 that gives a *proxy representation* $c_{\text{proxy}} \in \mathbb{R}^{P \times D}$ such that $c_{\text{proxy}} \equiv f_{\text{DINO}}(\hat{x}_0) \approx f_{\text{DINO}}(x_0)$).

Then, we can sample using a *regularized objective* that augments the standard posterior with a representation term:

$$\mathcal{J}_t(z_t \mid y, c_{\text{proxy}}) = \underbrace{\log p(y \mid z_t)}_{\text{likelihood}} + \underbrace{\log p(z_t)}_{\text{known unconditional dist. of } z_t} + \lambda \underbrace{\log p(c_{\text{proxy}} \mid z_t)}_{\text{additional REPA term}}, \tag{10}$$

where $\lambda > 0$ controls the strength of the additional guidance term. In eq. (10), the first term is the likelihood, which enforces measurements matching, the second term is the unconditional distribution over z_t that our flow or diffusion model has already learned, and the last term encourages semantic alignment with c_{proxy} in DINO feature space.

Representation Alignment term. Let us express the flow- or diffusion-based autoencoder $u_\theta = G_1 \circ (G_2(z_t, t))$, and denote $h_t^{[n]} = G_2^{[n]}(z_t, t)$ the n-th patch feature extracted from the diffusion model's hidden state h_t at timestep t. Following the strategy of REPA, we approximate the representation likelihood as a sum of cosine similarities between patch embeddings:

$$\log p(c_{\text{proxy}} \mid z_t) \;\approx\; \sum_{n=1}^{N} \cos\big(c_{\text{proxy}}^{[n]},\; g_\phi(G_2^{[n]}(z_t, t))\big), \tag{11}$$

where g_ϕ is a learnable projection head. This encourages z_t to follow trajectories whose intermediate representations remain aligned with c_{proxy} in the DINO feature space.

On the selection of c_{proxy}. To address the fact that representational alignment,Yu et al. (2024), is infeasible for inverse problems where ground-truth signals are not unavailable at inference time, we can approximate the ground-truth representation with $f_{\text{DINO}}(y)$, i.e., $\hat{x}_0 = y$, for problems where y is in the same space as the sought image, or $f_{\text{DINO}}(\mathbb{E}[z_0|z_t])$ (i.e., $\hat{x}_0 = \mathbb{E}[z_0|z_t]$). In both cases, we leverage the robustness of pretrained DINOv2 features. Interestingly, as advocated in the experimental section, DINOv2 encoder is robust to common corruptions encountered in super-resolution, deblurring, and inpainting due to the self-supervised way that follows during the training process. Appendix A.6 provides an ablation study quantifying the reliability of this approximation across tasks.

Our full algorithm in its general form for a velocity-based sampler is presented in Table 1. We instantiate this framework with two state-of-the-art latent diffusion methods, one gradient-based and one projection-based, with implementation details deferred to Appendix A.4 and A.5.

Algorithm 1 RePA Inverse Algorithm

Require: Flow model $u_\theta = G_1 \circ G_2$, measurement y, learning rate η, regularizer strength λ, proxy representation c_{proxy}

initialize $z_T \sim \mathcal{N}(0, I)$
for $t \in \{T, \dots, 1\}$ **do**
  $v \leftarrow u_\theta(z_t, t)$
  $\hat{z}_0 \leftarrow \mathbb{E}[z_0 \mid z_t]$
  $\Delta t \leftarrow 1/T$
  $z_{t-1} \leftarrow z_t + \Delta t \cdot v$
  $z_{t-1} \leftarrow z_{t-1} + \eta \nabla_{z_t} \log p(y \mid z_t)$
  $z_{t-1} \leftarrow z_{t-1} + \lambda \nabla_{z_t} \sum_{n=1}^{N} \cos\big(f_{\text{DINO}}^{[n]}(\hat{x}_0),\; g_\phi(G_2(z_t, t)^{[n]})\big)$
end for
return z_0

4 IMPROVING THE PERCEPTION-DISTORTION TRADE-OFF USING REPRESENTATIONAL ALIGNMENT

Recent works, Blau & Michaeli (2018); Ohayon et al. (2023), have brought to light and theoretically formalize the so-called *perception-distortion trade-off* that inverse image restoration algorithms have to address. Namely, it has been provably shown that as the distortion (measured, for example, by MSE, SSIM, PSNR) of a restored image decreases, its perceptual quality tends to deteriorate. In this section, we theoretically show that aligning the representation alignment of latent representations of diffusion- and flow-based models with those of pre-trained foundation models, such as DINOv2, can be viewed as a an additional perceptual quality optimization step, which provably improves the frontier of perception-distortion tradeoff.

We first provide the mathematical formulation of the perception-distortion trade-off.

Definition 1 (Blau & Michaeli (2018))**.** *Let p_x denote the distribution of ground thruth images and $p_{\hat{x}|y}$, the one of restored images given y. The perception-distortion trade-off of a restoration task can be formulated as follows:*

$$P(\hat{x}) = \min_{p_{\hat{x}|y}} d(p_x, p_{\hat{x}|y}) \quad \textit{subject to} \quad \mathbb{E}[\Delta(x, \hat{x})] \leq \epsilon, \tag{12}$$

where $d(p_x, p_{\hat{x}|y})$ is a divergence measure between the distributions of ground thruth images p_x, and the restored one $p_{\hat{x}|y}$, and $\Delta(x, \hat{x})$ is a distortion measure.

Given the above definition, we next establish a connection between the representational alignment term based on DINOv2 model, and the perceptual quality of the restored images.

Representational Alignment using DINOv2 features as a Perceptual Quality Metric. We first focus on the measurement matching and representation alignment steps of our proposed Algorithm 1. Specifically, we can see these updates as gradient descent steps on the following objective function,

$$L(z) = \underbrace{\|y - \mathcal{A}(\mathcal{D}(z)\|_2^2}_{\text{Distortion measure}} - \lambda \underbrace{\sum_{n=1}^{N} \cos\left((f_{\text{DINO}}^{[n]}(\bar{x}), g_\phi(z^{[n]})\right)}_{\text{REPA}(\bar{x}, \mathcal{D}(z))}, \tag{13}$$

where $\bar{x}$ corresponds to an approximation of the unknown ground truth image.

From (13), it is clear that our approach seeks to minimize the distortion term (the first term of the objective, corresponding to measurement matching) while simultaneously maximizing the regularization term, $\text{REPA}(\bar{x}, \mathcal{D}(z))$.

Next, we provide Lemma 1, which helps us to establish a connection between our regularization term and a distributional divergence measure. This links this term with the perceptual quality of the image and hence the perception-distortion problem of Definition 1.

First, we proceed with the following definition.

Definition 2 ((Misalignment between DINO and diffusion encoder representation)). *Let an image x and N pathes $\{x^{[n]}\}_{n=1}^N$. We define the misalignment between* DINO *and diffusion encoder representations of image x as follows,*

$$\text{MISREPA}(x) = \sum_{n=1}^{N} \|f_{\text{DINO}}^{[n]}(x) - g_\phi\left(\text{DIFFENC}^{[n]}(x)\right)\|_2^2, \tag{14}$$

where $\text{DIFFENC}^{[n]}(x) = z^{[n]}$, is the diffusion encoder.

Lemma 1 (REPA as a divergence measure). *Let $\text{REPA}(x, \hat{x}) = \sum_{n=1}^{N} \cos\left(f_{\text{DINO}}^{[n]}(x), g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\right)$. Assume that $\{x^{[n]}\}_{n=1}^N$ and $\{\hat{x}^{[n]}\}_{n=1}^N$ are N i.i.d samples of conditional probability distributions of patches of distributions $\bar{p}_{x^{patch}|x}, \bar{p}_{\hat{x}^{patch}|\hat{x}}$. Under Assumption 1, it holds,*

$$\max_{\hat{x}} \text{REPA}(x, \hat{x}) \leq \min_{\hat{x}} \widehat{\text{MMD}}_{\text{DINO}}(\{x^{[n]}\}_{n=1}^N, \{\hat{x}^{[n]}\}_{n=1}^N) + \text{MISREPA}(\hat{x}), \tag{15}$$

where $\widehat{\text{MMD}}_{\text{DINO}}(\{x^{[n]}\}_{n=1}^N, \{\hat{x}^{[n]}\}_{n=1}^N) = \|\frac{1}{N}\sum_{n=1}^N f_{\text{DINO}}(x^{[n]}|x^{[1]}, x^{[2]}, \ldots, x^{[N]}) - \frac{1}{N}\sum_{n=1}^N f_{\text{DINO}}(\hat{x}^{[n]}|\hat{x}^{[1]}, \hat{x}^{[2]}, \ldots, \hat{x}^{[N]})\|_2^2$ is an empirical maximum mean discrepancy measure[1].

From Lemma 1, we draw a connection between the representational alignment term and maximum mean discrepancy (MMD), Arbel et al. (2019), which is can serve as perceptual metric defined over the image patches. Interestingly, this connection provides for the first time, some insights into the efficiency of representational alignment beyond the scope of inverse problems.

Note that the connection between kernel mean embeddings of patches and a distributional distance—such as the empirical estimate of $\widehat{\text{MMD}}$ in Lemma 1—can also be established using encoders other than DINO that operate on patches. However, unlike alternative feature mappings, DINO is a foundation model whose representations have been empirically shown to maximize alignment with human perceptual judgments, Fu et al. (2023).

Remark 1. *Note that the upper bound of the representational alignment term in (17), can become tighter if the misalignment between* DINO *and diffusion encoder representations is 0.*

[1]Note that $f_{\text{DINO}}(x^{[n]}|x^{[1]}, x^{[2]}, \ldots, x^{[N]}) \equiv f_{\text{DINO}}^{[n]}(x)$. Here, we use this new notation to show connection with condition kernel mean embeddings, Song et al. (2013) and MMD.

Assumption 1. DINO *and diffusion encoder representations are aligned i.e., for any* $x \sim p_x$*, it holds* $\text{MISREPA}(x) = 0$.

The above assumption can be satisfied since DINO and diffusion encoder representations have gone through an alignment process during the training process, Yu et al. (2024).

Remark 2. *As described above, in inverse problems we do not know the ground truth signals, hence alignment is taking place on approximations of it. This source of additional error, if considered, would contribute to an additional term in the RHS of (17), thus forming a less tight bound for* $\text{REPA}(x, \hat{x})$.

Theorem 1. *(Informal) Let the Perception-Distortion formulation given in equation 12 and assume a distortion measure* $\Delta(x, \hat{x})$*. Let* $x^* \in \arg\min_{\hat{x}} \Delta(x, \hat{x})$ *and* $x^\star \in \arg\min_{\hat{x}} \Delta(x, \hat{x}) + \lambda P(\hat{x})$ *By incorporating* REPA *regularization to inverse algorithms will improve the perception-distortion frontier, i.e.,*

$$\begin{aligned} \Delta(x, x^\star) &\leq \Delta(x, x^*) + \epsilon \\ P(x^\star) &\leq P(x^*) - \Omega(\sqrt{\epsilon}) + \mathcal{O}(\lambda^2), \end{aligned} \tag{16}$$

The proof of this Theorem is provided at the appendix. The theorem demonstrates that incorporating a perceptual quality metric via the representational alignment term, and choosing a sufficiently small λ, enhances perceptual quality with negligible impact on distortion, thereby improving the trade-off frontier.

5 EXPERIMENTAL RESULTS

In this section, we present experimental results that demonstrate the effectiveness of our alignment regularizer. We evaluated reconstruction quality using four widely adopted metrics: PSNR (peak signal-to-noise ratio), SSIM (structure similarity index) (Wang et al., 2004), LPIPS (learned perceptual image patch similarity) (Zhang et al., 2018), and FID (Fréchet Inception Distance) (Heusel et al., 2017). All metrics are averaged over 100 images from the ImageNet 256×256 validation set (Deng et al., 2009).

We consider four inverse problems: super-resolution, box inpainting, Gaussian deblurring, and motion deblurring. For super-resolution, images are downsampled by a factor of 4. For box inpainting, we mask out a 128×128 square region. For Gaussian deblurring, we apply a Gaussian kernel of size 61×61 with standard deviation 3.0. Motion blur is applied using randomly generated kernels with intensity 0.5. In all experiments we also add Gaussian noise with standard deviation 0.01.

5.1 EFFECTIVENESS OF THE REPA REGULARIZER

We first evaluate the effect of REPA when applied to two latent-space inverse problem solvers: Latent DPS and ReSample. We give a full overview of the algorithms and how we incorporate the regularizer in each case in the Appendix. For both methods, we use the latent diffusion model trained with representational alignment from Yu et al. (2024), which also provides a pretrained MLP that maps the internal representations of the diffusion model to the DINO space. Table 1 reports the quantitative results. Across all four inverse problems, incorporating REPA yields consistent improvements in perceptual metrics, with notable reductions in LPIPS and FID for both Latent DPS and ReSample. Figure 11 shows qualitative examples for the tasks of box inpainting and Gaussian debluring. Reconstructions obtained with REPA are noticeably sharper and visually closer to the ground truth compared to those from the unregularized methods.

Effect of REPA with Varying Discretization Steps. Figure 5 illustrates how LPIPS varies with the number of discretization steps across three inverse problems. Incorporating the REPA regularizer consistently reduces perceptual error compared to Latent DPS alone. More importantly, REPA enables comparable or better perceptual quality with substantially fewer sampling steps: for example, in super-resolution and Gaussian deblurring, REPA achieves the same LPIPS with roughly $4\times$ fewer steps, while in motion deblurring it achieves the same level with about $2\times$ fewer steps. These results highlight the efficiency gains provided by our method, reducing computational cost without sacrificing the reconstruction quality of the baseline method.

Table 1: Performance comparison of Latent DPS and ReSample (with and without the proposed Repa regularizer) on $4\times$ super-resolution, box inpainting, Gaussian deblurring, and motion deblurring tasks (ImageNet validation). Gaussian noise with $\sigma = 0.01$ was added in all cases.

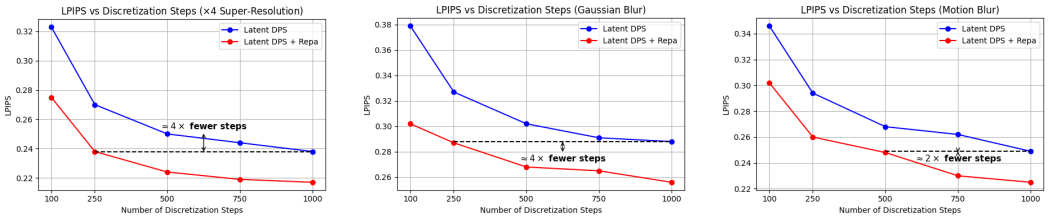
Task	Method	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑
$4\times$ Super-resolution	Latent DPS	0.238	123.82	26.88	0.732
	Latent DPS + REPA	0.217	86.88	26.82	0.731
	Resample	0.208	97.02	26.60	0.732
	Resample + REPA	**0.197**	**74.70**	26.41	0.724
Box Inpainting	Latent DPS	0.151	116.53	20.53	0.827
	Latent DPS + REPA	**0.139**	**88.69**	20.45	0.824
Gaussian Deblurring	Latent DPS	0.288	152.96	25.76	0.648
	Latent DPS + REPA	0.256	102.99	25.66	0.669
	Resample	0.259	115.47	26.19	0.699
	Resample + REPA	**0.223**	**102.59**	25.99	0.695
Motion Deblurring	Latent DPS	0.249	129.08	27.19	0.738
	Latent DPS + REPA	0.225	90.23	27.01	0.735
	Resample	0.210	94.95	27.27	0.736
	Resample + REPA	**0.192**	**75.11**	27.18	0.738

(a) Box inpainting (b) Gaussian deblurring

Figure 2: Qualitative comparison of Latent DPS with and without the proposed REPA regularizer on two inverse problems: (a) box inpainting and (b) Gaussian deblurring.

Table 2: Performance comparison of state of the art algorithms on $4\times$ super-resolution, box inpainting, Gaussian deblurring, and motion deblurring tasks (ImageNet validation). Gaussian noise with $\sigma = 0.01$ was added in all cases.

Task	Method	LPIPS ↓	FID ↓	PSNR ↑	SSIM ↑
$4\times$ Super-resolution	DPS	0.244	92.13	24.42	0.681
	Latent DAPS	0.252	120.57	27.13	0.744
	Latent DPS + REPA	0.217	86.88	26.82	0.731
	Resample + REPA	**0.197**	**74.70**	25.18	0.680
Box Inpainting	DPS	0.191	97.95	19.11	0.769
	Latent DPS + REPA	0.139	88.69	20.45	0.824
Gaussian Deblurring	DPS	0.366	157.73	19.55	0.461
	Latent DAPS	0.300	163.36	26.02	0.692
	Latent DPS + REPA	0.256	102.99	25.66	0.669
	Resample + REPA	**0.223**	**102.59**	25.12	0.665
Motion Deblurring	DPS	0.242	89.06	24.17	0.678
	Latent DAPS	0.274	135.18	27.01	0.734
	Latent DPS + REPA	0.225	90.23	27.01	0.735
	Resample + REPA	**0.192**	**75.11**	27.18	0.738



(a) Super Resolution (b) Gaussian Debluring (c) Motion Debluring

Figure 3: Comparison of LPIPS as a function of discretization steps for Latent DPS and Latent DPS + REPA. Using our regularizer, comparable performance is achieved with substantially fewer number of steps.

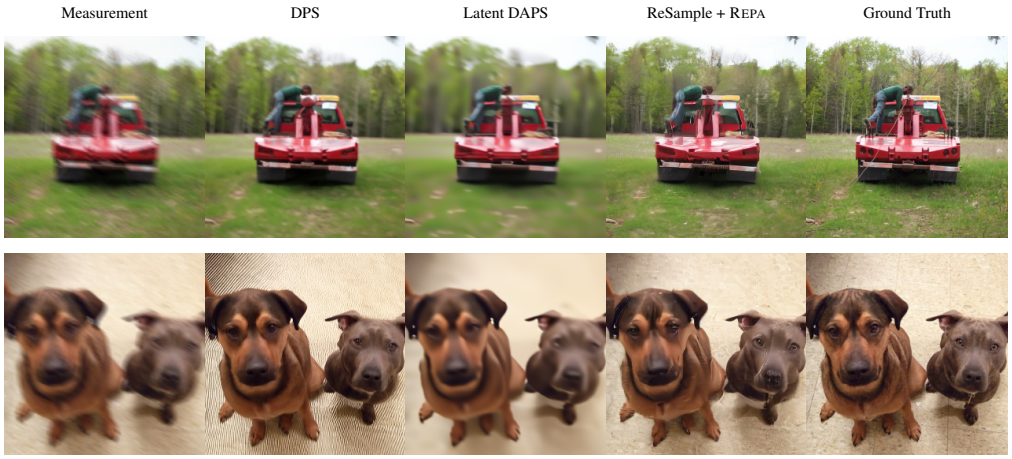


Figure 4: Qualitative comparison for motion deblurring. Each column shows the measurement, DPS, Latent DAPS, ReSample + REPA, and the ground-truth image.

5.2 COMPARISON WITH OTHER SOTA ALGORITHMS

After establishing that the REPA regularizer improves reconstruction quality, we compare our approach with other state-of-the-art algorithms. For pixel-space methods, we consider DPS (Chung et al., 2023), which performs gradient-based sampling and is evaluated using the pretrained diffusion model from Dhariwal & Nichol (2021). For latent diffusion approaches, we include Latent DAPS. Since our method relies on a model trained on ImageNet, we evaluated Latent DAPS with the standard conditional latent diffusion model trained on ImageNet from Rombach et al. (2022), to ensure a fair comparison. All aforementioned algorithms are run with their default configurations. Table 2 presents the quantitative results, where the inclusion of our regularizer consistently achieves state-of-the-art performance. Figure 11 provides qualitative examples for motion deblurring, further highlighting the advantage of our method. In particular, DPS often introduces visible artifacts, whereas Latent DAPS fails to fully correct blur and produces reconstructions lacking sharpness.

6 CONCLUSIONS

In this work, we introduced representational alignment as a tool for solving inverse problems with diffusion- and flow-based models. Specifically, we demonstrated that aligning model representations with those of a pretrained DINO encoder can effectively guide the reverse process toward perceptually higher-quality reconstructions. Furthermore, we established a theoretical connection between this phenomenon and the perception–distortion trade-off by deriving an empirical estimate of maximum mean discrepancy of distribution of patches, and the representational alignment term. Our work, paves the way for the development new approaches for solving inverse problems moving beyond standard priors.

7 Reproducibility Statement

We are committed to release our code and implementation details upon acceptance to the conference to ensure reproducibility of our results.

References

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32, 2019.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6228–6237, 2018.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 12606–12633, 2024.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Vincent Tao Hu, David W Zhang, Yuki M Asano, Gertjan J Burghouts, and Cees GM Snoek. Self-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18413–18422, 2023.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. *arXiv preprint arXiv:2504.10483*, 2025.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023.

Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37: 125441–125468, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024.

Guy Ohayon, Tomer Michaeli, and Michael Elad. The perception-robustness tradeoff in deterministic image restoration. *arXiv preprint arXiv:2311.09253*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Maitreya Patel, Song Wen, Dimitris N Metaxas, and Yezhou Yang. Steering rectified flow models in the vector field for controlled image generation. *arXiv preprint arXiv:2412.00100*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *The Twelfth International Conference on Learning Representations*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-repa: Aligning diffusion u-nets to vits. *arXiv preprint arXiv:2503.18414*, 2025.

Chenyu Wang, Cai Zhou, Sharut Gupta, Zongyu Lin, Stefanie Jegelka, Stephen Bates, and Tommi Jaakkola. Learning diffusion models with flexible representation guidance. *arXiv preprint arXiv:2507.08980*, 2025a.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Ziqiao Wang, Wangbo Zhao, Yuhao Zhou, Zekai Li, Zhiyuan Liang, Mingjia Shi, Xuanlei Zhao, Pengfei Zhou, Kaipeng Zhang, Zhangyang Wang, et al. Repa works until it doesn't: Early-stopped, holistic alignment supercharges diffusion training. *arXiv preprint arXiv:2505.16792*, 2025b.

Weilai Xiang, Hongyu Yang, Di Huang, and Yunhong Wang. Denoising diffusion autoencoders are unified self-supervised learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Wanghan Xu, Xiaoyu Yue, Zidong Wang, Yao Teng, Wenlong Zhang, Xihui Liu, Luping Zhou, Wanli Ouyang, and Lei Bai. Exploring representation-aligned latent space for better generation. *arXiv preprint arXiv:2502.00359*, 2025.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15703–15712, 2025.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A APPENDIX

A.1 THE USE OF LARGE LANGUAGE MODELS (LLMS)

In our work we employed LLMs specifically ChatGPt as an auxiliary tool to accelerate code writing tasks such as data visualization. All algorithms, theoretical developments, and experimental designs were conceived and implemented by the authors.

A.2 PROOFS

Lemma 2 (REPA as a divergence measure)**.** *Let* $\text{REPA}(x, \hat{x}) = \sum_{n=1}^{N} \cos\left(f_{\text{DINO}}^{[n]}(x), g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\right)$. *Assume that* $\{x^{[n]}\}_{n=1}^N$ *and* $\{\hat{x}^{[n]}\}_{n=1}^N$ *are* N *i.i.d samples of conditional probability distributions of patches of distributions* $\bar{p}_{x^{patch}|x}, \bar{p}_{\hat{x}^{patch}|\hat{x}}$. *Under Assumption 1, it holds,*

$$\max_{\hat{x}} \text{REPA}(x, \hat{x}) \leq \min_{\hat{x}} \widehat{\text{MMD}}_{\text{DINO}}(\{x^{[n]}\}_{n=1}^N, \{\hat{x}^{[n]}\}_{n=1}^N) + \text{MISREPA}(\hat{x}), \tag{17}$$

where $\widehat{\text{MMD}}_{\text{DINO}}(\{x^{[n]}\}_{n=1}^N, \{\hat{x}^{[n]}\}_{n=1}^N) = \|\frac{1}{N}\sum_{n=1}^N f_{\text{DINO}}(x^{[n]}|x^{[1]}, x^{[2]}, \ldots, x^{[N]}) - \frac{1}{N}\sum_{n=1}^N f_{\text{DINO}}(\hat{x}^{[n]}|\hat{x}^{[1]}, \hat{x}^{[2]}, \ldots, \hat{x}^{[N]})\|_2^2$ *is an empirical maximum mean discrepancy measure*[2]*.*

Proof. We have,

$$\text{REPA}(x, \hat{x}) = \sum_{n=1}^{N} \cos\left(f_{\text{DINO}}^{[n]}(x), g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\right), \tag{18}$$

where,

$$\cos\left((f_{\text{DINO}}^{[n]}(x), g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\right) = \frac{\langle f^{[n]}(x), g_\phi(\text{DIFFENC}(\hat{x}^{[n]})\rangle}{\|f^{[n]}(x)\|_2 \|g_\phi(\text{DIFFENC}(\hat{x}^{[n]})\|_2}. \tag{19}$$

We assume DINO features and latent representations of diffusion/flow-based models are normalized, and get,

$$\begin{aligned} \cos\left((f_{\text{DINO}}^{[n]}(x), g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\right) &= 2 - \|f^{[n]}(x) - g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\|_2^2 \rightarrow \\ \sum_{n=1}^{N} \cos\left((f_{\text{DINO}}^{[n]}(x), g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\right) &= 2N - \sum_{n=1}^{N} \|f^{[n]}(x) - g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\|_2^2 \\ &\leq 2N - \|\sum_{n=1}^{N} f^{[n]}(x) - \sum_{n=1}^{N} g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\|_2^2 \end{aligned} \tag{20}$$

We use the empirical maximum mean discrepancy (MMD) on DINO embeddings, and assume that inputs of DINO encoder $\{x^{[n]}|x^{[1]}, x^{[2]}, \ldots, x^{[N]}\}_{n=1}^N, \{\hat{x}^{[n]}|\hat{x}^{[1]}, \hat{x}^{[2]}, \ldots, \hat{x}^{[N]}\}_{n=1}^N$ are i.i.d, samples of distributions $\bar{p}_{x|x_0}, \bar{p}_{\hat{x}|\hat{x}_0}$, respectively,

$$\widehat{\text{MMD}}_{\text{DINO}}(\{x^{[n]}\}_{n=1}^N, \{\hat{x}^{[n]}\}_{n=1}^N) = \|\frac{1}{N}\sum_{n=1}^N f_{\text{DINO}}(x^{[n]}|x^{[1]}, x^{[2]}, \ldots, x^{[N]}) - \frac{1}{N}\sum_{n=1}^N f_{\text{DINO}}(\hat{x}^{[n]}|\hat{x}^{[1]}, \hat{x}^{[2]}, \ldots, \hat{x}^{[N]})\|_2^2 \tag{21}$$

[2]Note that $f_{\text{DINO}}(x^{[n]}|x^{[1]}, x^{[2]}, \ldots, x^{[N]}) \equiv f_{\text{DINO}}^{[n]}(x)$. Here, we use this new notation to show connection with condition kernel mean embeddings, Song et al. (2013) and MMD.

Note that,

$$\|\frac{1}{N}\sum_{n=1}^{N} f_{\text{DINO}}(x^{[n]}) - \frac{1}{N}\sum_{n=1}^{N} f_{\text{DINO}}(\hat{x}^{[n]})\|_2^2 \leq 2\Big(\|\frac{1}{N}\sum_{n=1}^{N} f_{\text{DINO}}(x^{[n]}) - \frac{1}{N}\sum_{n=1}^{N} g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\|_2^2 + \|\|\frac{1}{N}\sum_{n=1}^{N} f_{\text{DINO}}(\hat{x}^{[n]}) - \frac{1}{N}\sum_{n=1}^{N} g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\|_2^2\Big)$$

$$\leq \frac{2}{N}\underbrace{\sum_{n=1}^{N}\|f_{\text{DINO}}(x^{[n]}) - g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\|_2^2}_{①} + \frac{2}{N}\underbrace{\sum_{n=1}^{N}\|f_{\text{DINO}}(\hat{x}^{[n]}) - g_\phi(\text{DIFFENC}(\hat{x}^{[n]}))\|_2^2}_{②} \tag{22}$$

We have ① is the $\text{REPA}(x, \hat{x})$ term (see equation 11), and term ② expresses the misalignment of DINO and diffusion encoders (DIFFENC) and is 0 unders Assumption 1. □

A.3 PROOF OF THEOREM 4.2

Given an x image (possibly corrupted version of a clean one), we assume x^* is the optimal estimate that minimizes the distortion measure i.e.,

$$x^* \in \arg\min_{\hat{x}} \Delta(x, \hat{x}) \tag{23}$$

We also assume that our approach solve the following optimization problem,

$$x^*(\lambda) \in \arg\min_{\hat{x}} \Delta(x, \hat{x}) + \lambda P(\hat{x}) \tag{24}$$

We also assume that there exists a λ such that $P(x^*(\lambda)) \leq P(x^*(\lambda))$, i.e., the perceptual metric can be fruther improved when solving the regularized problem given in (24).

Next we make the following assumptions on the distortion measure.

Assumption 2. *The distortion measure $\Delta(x, \hat{x})$ and the perception metric $P(x)$ are L_Δ- and L_P-smooth, repsectively, around x^* i.e., $x \in \{x : \|x - x^*\|_2 \leq \epsilon\}$, it holds*

$$\begin{aligned} \Delta(x, \hat{x}) &\leq \Delta(x, x^*) + \frac{L_D}{2}\|\hat{x} - x^*\|_2^2 \\ P(\hat{x}) &\leq P(x^*) + \nabla P(x^*)(\hat{x} - x^*) + \frac{L_P}{2}\|\hat{x} - x^*\|_2^2 \end{aligned} \tag{25}$$

Assumption 3. *The distortion measure $\Delta(x, \hat{x})$ is μ_D-strongly convex, around x^* i.e., $\bar{x} \in \{\bar{x} : \|\bar{x} - x^*\|_2 \leq \epsilon\}$, it holds*

$$\nabla^2 \Delta(x, \hat{x}) \succeq \mu I \tag{26}$$

Theorem 2. *Let the Perception-Distortion formulation given in equation 12 and assume a distortion measure $\Delta(x, \hat{x})$ that satisfies Assumption 3. By incorporating* REPA *regularization to inverse algorithms will improve the perception-distortion frontier i.e., $\exists \lambda > 0$, such that*

$$\begin{aligned} \Delta(x, x(\lambda)) &\leq \Delta(x, x^*) + \epsilon \\ P(x(\lambda) &\leq P(x^*) - \sqrt{\epsilon} + \mathcal{O}(\lambda^2), \end{aligned} \tag{27}$$

Proof. Since $x^*(\lambda)$ is a minimizer of 24, we have,

$$\nabla \Delta(x, x^*(\lambda)) + \lambda \nabla P(x^*(\lambda)) = 0 \tag{28}$$

From Taylor expansion on $\nabla \Delta(x, x^*(\lambda))$ and $\nabla P(x^*(\lambda))$, we get,

$$\begin{aligned} \nabla \Delta(x, x^\star(\lambda)) = \nabla^2 \Delta(x, x^*(\lambda))(x^*(\lambda) - x^*) + \gamma_D, \\ \nabla P(x^\star(\lambda)) = \nabla P(x^*) + \gamma_P \end{aligned} \tag{29}$$

, where $\|\gamma_D\| \leq \mathcal{O}(\|x^*(\lambda) - x^*\|_2^2)$, $\|\gamma_P\| \leq \mathcal{O}(\|x^*(\lambda) - x^*\|_2^2)$.

Since the Hessian $\nabla^2\Delta(x, x^*(\lambda))$ is invertible (due to strong convexity), and by plugging (29) into (28), we can get

$$\begin{aligned} x^*(\lambda) - x^* = -\lambda\nabla^{-2}\Delta(x, x^*)\nabla P(x^*) - \nabla^{-2}\Delta(x, x^*)\gamma_D - \lambda\nabla^{-2}\Delta(x, x^*)\gamma_P \rightarrow \\ \|x^*(\lambda) - x^*\|_2 \leq \lambda\|\nabla^{-2}\Delta(x, x^*)\nabla P(x^*)\|_2 + \mathcal{O}(\lambda) \end{aligned} \tag{30}$$

Since $\Delta(x, \hat{x})$ is L_D-smooth, we have,

$$\begin{aligned} \Delta(x, x^\star(\lambda)) \leq \Delta(x, x^*) + \frac{L_D}{2}\|x^*(\lambda) - x^*\|_2^2 \rightarrow \\ \Delta(x, x^\star(\lambda)) \leq \Delta(x, x^*) + \frac{\lambda^2 L_D}{2}\|\nabla^{-2}\Delta(x, x^*)\nabla P(x^*)\|_2^2 + \mathcal{O}(\lambda^2) \end{aligned} \tag{31}$$

From Taylor expansion and smoothness of $P(\hat{x}$, we get,

$$\begin{aligned} P(x^\star(\lambda)) \leq P(x^*) + \nabla^\top P(x^*)(x^\star\lambda) - x^*) + \frac{L_P}{2}\|x^\star(\lambda) - x^*\|_2^2 \rightarrow \\ P(x^\star(\lambda)) \leq P(x^*) - \lambda\frac{L_P}{2}\nabla^\top P(x^*)\nabla^{-2}\Delta(x, x^*)\nabla P(x^*) + \mathcal{O}(\lambda^2) \end{aligned} \tag{32}$$

Let us now assume that $\Delta(x, x^\star(\lambda)) - \Delta(x, x^*) \leq \epsilon$, which could be attained when $\lambda = \sqrt{\frac{2\epsilon}{L_D\|\nabla^{-2}\Delta(x,x^*)\nabla P(x^*)\|_2^2}}$. Then that implies,

$$P(x^\star(\lambda)) - P(x^*) \leq -k\sqrt{\epsilon} + \mathcal{O}(\lambda^2), \tag{33}$$

where $k = \frac{\sqrt{2}\frac{L_P}{2}\nabla^\top P(x^*)\nabla^{-2}\Delta(x,x^*)\nabla P(x^*)}{\sqrt{L_D\|\nabla^{-2}\Delta(x,x^*)\nabla P(x^*)\|_2^2}}$

□

A.4 Details About Latent DPS + Repa Implementation

We implement Latent DPS + Repa following the procedure in Algorithm 2. A key design choice is the learning rate schedule η_t, for which we consider two adaptive parameterizations:

$$\eta(t) = \begin{cases} \dfrac{\kappa}{\|y - A\mathcal{D}(\mathbb{E}[z_0 \mid z_t])\|_2} & \text{(Adaptive inverse norm)} \\ \dfrac{\kappa}{\max\left(\frac{t}{1-t}, 1\right)} & \text{(Adaptive SNR-based learning rate)} \end{cases}$$

where κ is a tunable scaling factor. Table 3 compares these schedules for $4\times$ super-resolution on ImageNet, showing that the SNR-based strategy consistently achieves better perceptual performance (lower LPIPS) and slightly higher PSNR. For each schedule, κ and λ are tuned on a small validation set. We use the SNR-based learning rate for all main experiments unless otherwise stated.

$4\times$ **Super-resolution (ImageNet)**

(a) Latent DPS

LR Type	κ	PSNR ↑	LPIPS ↓
Inverse norm	7.75	26.09	0.274
SNR-based	2.25	**26.82**	**0.238**

(b) Latent DPS + Repa

LR Type	λ	κ	PSNR ↑	LPIPS ↓
Inv norm	0.005	5	25.90	0.235
SNR-based	0.01	2	**26.82**	**0.217**

Table 3: Comparison of learning rate strategies for $4\times$ super-resolution on ImageNet.

Algorithm 2 Latent DPS + REPA Algorithm

Require: flow model $u_\theta = G_1 \circ G_2$, measurement y, pretrained encoder f
1: initialize $x_T \sim \mathcal{N}(0, I)$
2: **for** $t \in \{T, \dots, 0\}$ **do**
3: $v \leftarrow u_\theta(z_t, t)$
4: $\hat{z}_0 \leftarrow \mathbb{E}[z_0 \mid z_t]$
5: $\Delta t \leftarrow 1/T$
6: $z_{t-1} \leftarrow z_t + \Delta t \cdot v$
7: $z_{t-1} \leftarrow z_{t-1} - \eta \nabla_{z_t} \|y - \mathcal{A}(\mathcal{D}(\hat{z}_0))\|_2^2$
8: $z_{t-1} \leftarrow z_{t-1} + \lambda \nabla_{z_t} \sum_{n=1}^{N} \cos((f_{\text{DINO}}^{[n]}(y), g_\phi(G_2(z_t, t)^{[n]}))$
9: **end for**
10: **return** x_0

Algorithm 3 Latent DPS + REPA Algorithm

Require: flow model $u_\theta = G_1 \circ G_2$, measurement y, pretrained encoder f, resample steps C, parameter γ
1: initialize $z_T \sim \mathcal{N}(0, I)$
2: **for** $t \in \{T, \dots, 0\}$ **do**
3: $v \leftarrow u_\theta(z_t, t)$
4: $\hat{z}_0 \leftarrow \mathbb{E}[z_0 \mid z_t]$
5: $\Delta t \leftarrow 1/T$
6: $z_{t-1} \leftarrow z_t + \Delta t \cdot v$
7: **if** $t \in C$ **then**
8: $\tilde{z}_0(y) \leftarrow \arg\min_z \frac{1}{2}\|y - \mathcal{A}(\mathcal{D}(z))\|_2^2$ ▷ initialize at $\hat{z}_0$
9: $z_{t-1} \leftarrow \text{STOCHASTICRESAMPLE}(\tilde{z}_0(y),\, z_{t-1},\, \gamma)$
10: **end if**
11: $z_{t-1} \leftarrow z_{t-1} - \eta \nabla_{z_t} \|y - \mathcal{A}(\mathcal{D}(\hat{z}_0))\|_2^2$
12: $z_{t-1} \leftarrow z_{t-1} + \lambda \nabla_{z_t} \sum_{n=1}^{N} \cos\big((f_{\text{Dino}}^{[n]}(y),\, g_\phi(G_2(z_t, t)^{[n]})\big)$
13: **end for**
14: $x_0 \leftarrow \mathcal{D}(z_0)$
15: **return** x_0

A.5 DETAILS ABOUT RESAMPLE + REPA IMPLEMENTATION

In Table 3, we present an adaptation of the algorithm proposed by Song et al. (2024) to the flow-based setting, augmented with the REPA regularizer as described in the methodology section. We note that the original ReSample algorithm employs a three-stage procedure for enforcing data consistency: (i) gradient steps in latent space, similar to Latent DPS; (ii) pixel-space optimization, which is computationally efficient and captures high-level semantics but often leads to blurrier reconstructions; and (iii) latent space optimization as outlined to line7 of Algorithm3. In contrast, our adaptation omits the pixel-space stage, as our focus is on maximizing perceptual quality. We found that this modification together with the inclusion of the REPA reguralizer yields sharper and more visually convincing results, while still benefiting from the refinement effect of the final latent-space consistency updates.

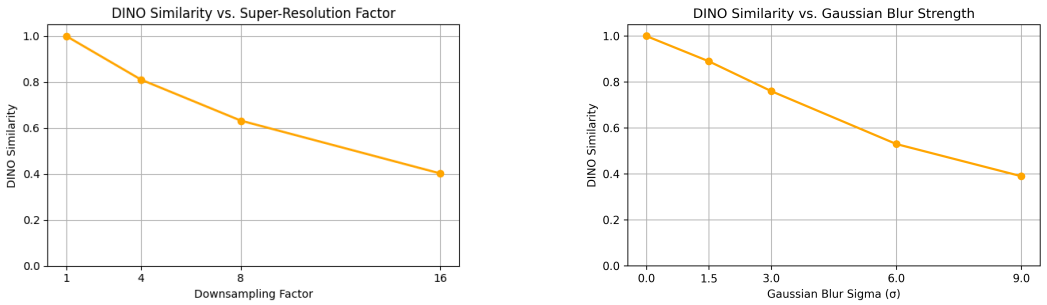
Importantly, the resulting algorithm remains consistent with the central principle of ReSample—that initialization strongly impacts the effectiveness of data-consistency updates. In our case, REPA guides the initialization toward reconstructions with stronger perceptual quality, which the subsequent consistency updates then refine further, improving fidelity and sharpness.

A.6 ROBUSTNESS OF DINO REPRESENTATIONS TO CORRUPTIONS

To evaluate the robustness of DINO representations under various corruptions, we conduct experiments on a fixed set of 100 images. For each image, we compute the average patch similarity between its ground truth representation and that of its corrupted version. We assess this similarity

across different corruption types—specifically super-resolution and Gaussian deblurring—at varying levels of severity. We report the average similarity across the selected subset of 100 images.

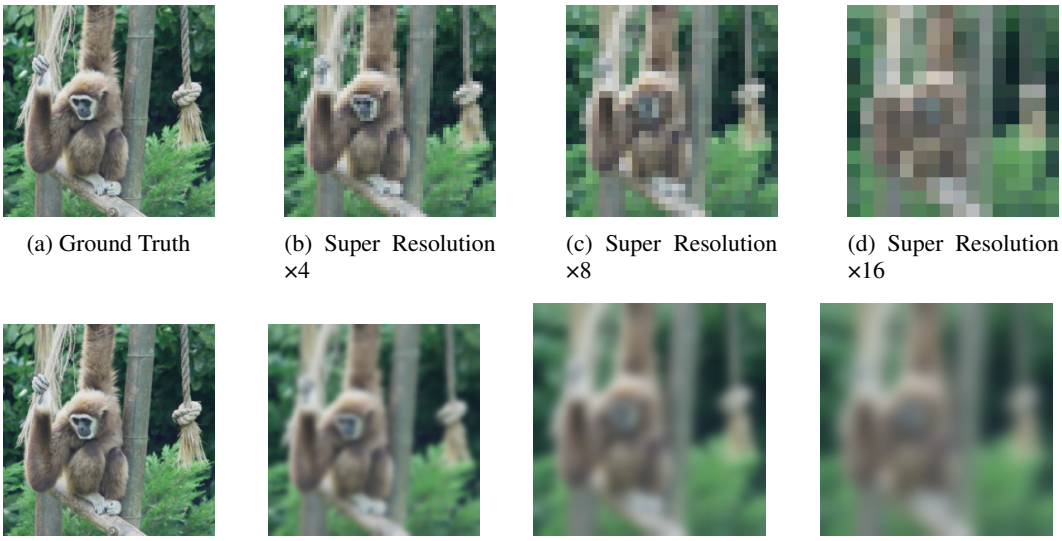
Figure 5 presents how DINO representation similarity changes as the corruption severity increases for the two tasks. Figure 6 further visualizes how the pixel-space appearance of a single image changes across corruption levels. Interestingly, we observe that DINO representations remain significantly more robust to both super-resolution and Gaussian deblurring. Despite substantial visual degradation in pixel space, DINO maintains a strong representational alignment with the original image.



(a) Super-Resolution

(b) Gaussian Deblurring

Figure 5: Similarity of representations under increasing levels of corruption.



(a) Ground Truth

(b) Super Resolution ×4

(c) Super Resolution ×8

(d) Super Resolution ×16

(e) Ground Truth

(f) Gaussian deblur $\sigma = 3$

(g) Gaussian deblur $\sigma = 6$

(h) Gaussian deblur $\sigma = 9$

Figure 6: Visual comparison of a corrupted image used in the similarity experiments for different corruption types and severity levels.

A.7 Additional Qualitative Results

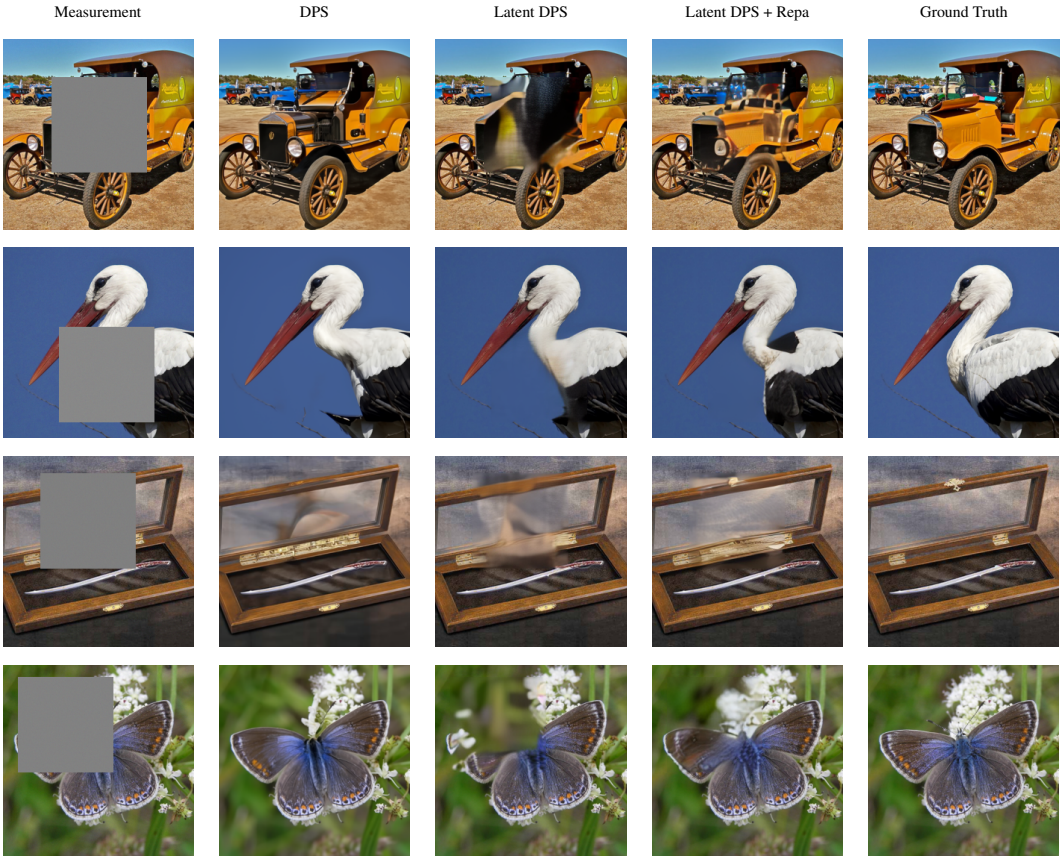


Figure 7: Qualitative comparison for Boc Inpainting. Each row shows: Measurement, pixel-space method DPS, latent method DPS, its Repa-enhanced version, and the Ground Truth.

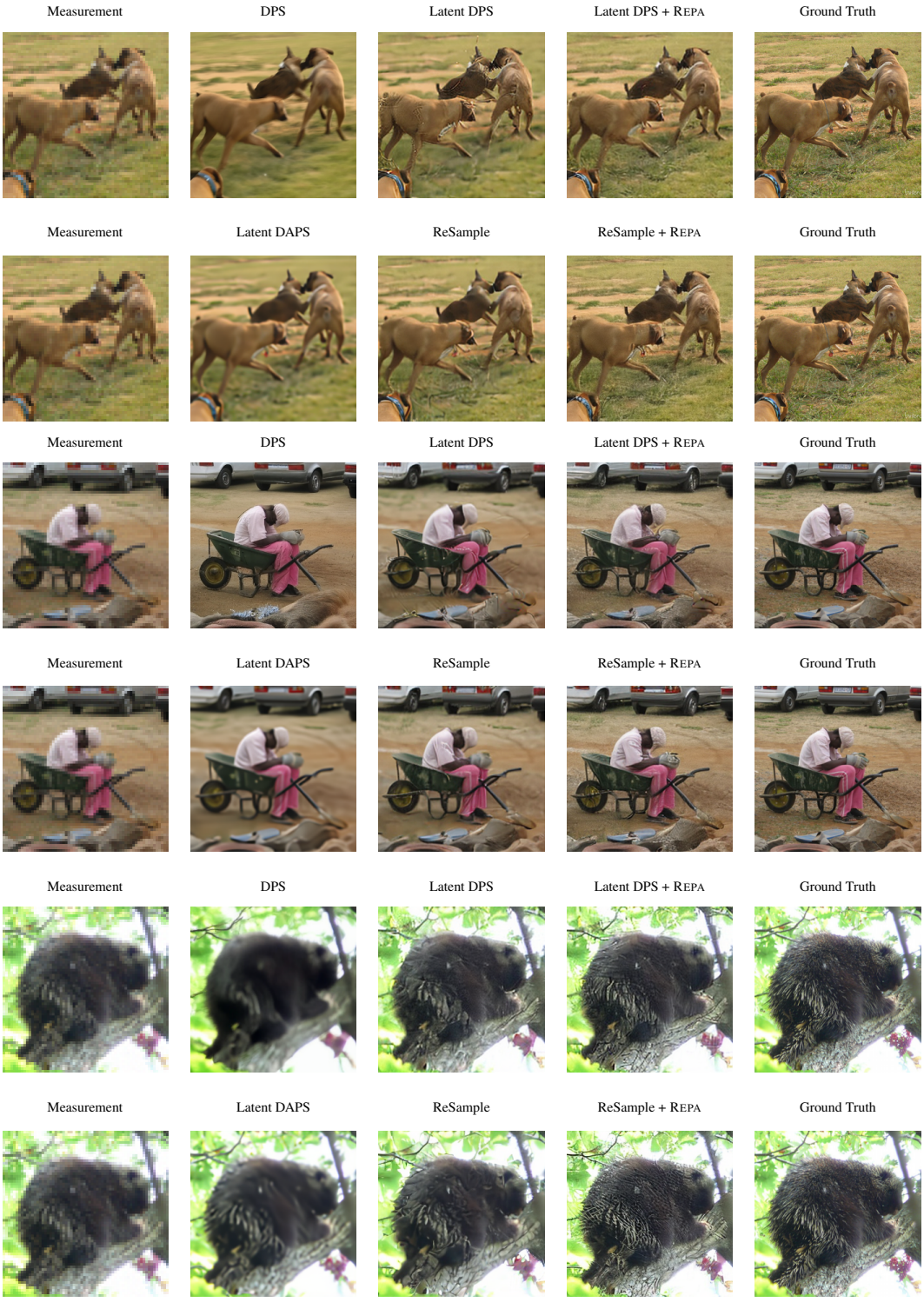


Figure 8: Qualitative comparison for $4\times$ super-resolution. Each row shows: Measurement, a baseline method (DPS or DAPS), a latent solver (Latent DPS or ReSample), its REPA-enhanced version, and the Ground Truth.

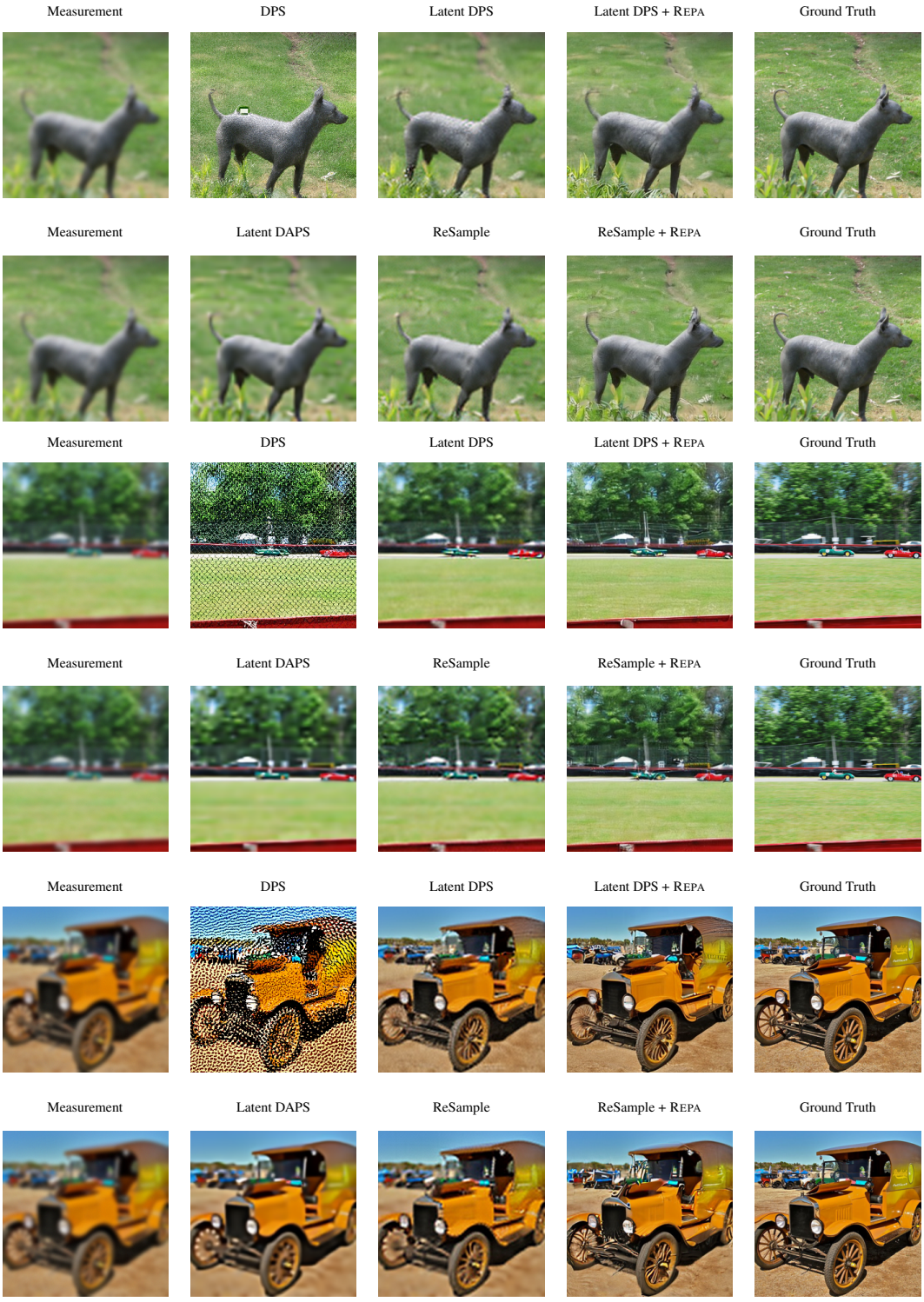


Figure 9: Qualitative comparison for Gaussian Deblurring. Each row shows: Measurement, a baseline method (DPS or DAPS), a latent solver (Latent DPS or ReSample), its REPA-enhanced version, and the Ground Truth.

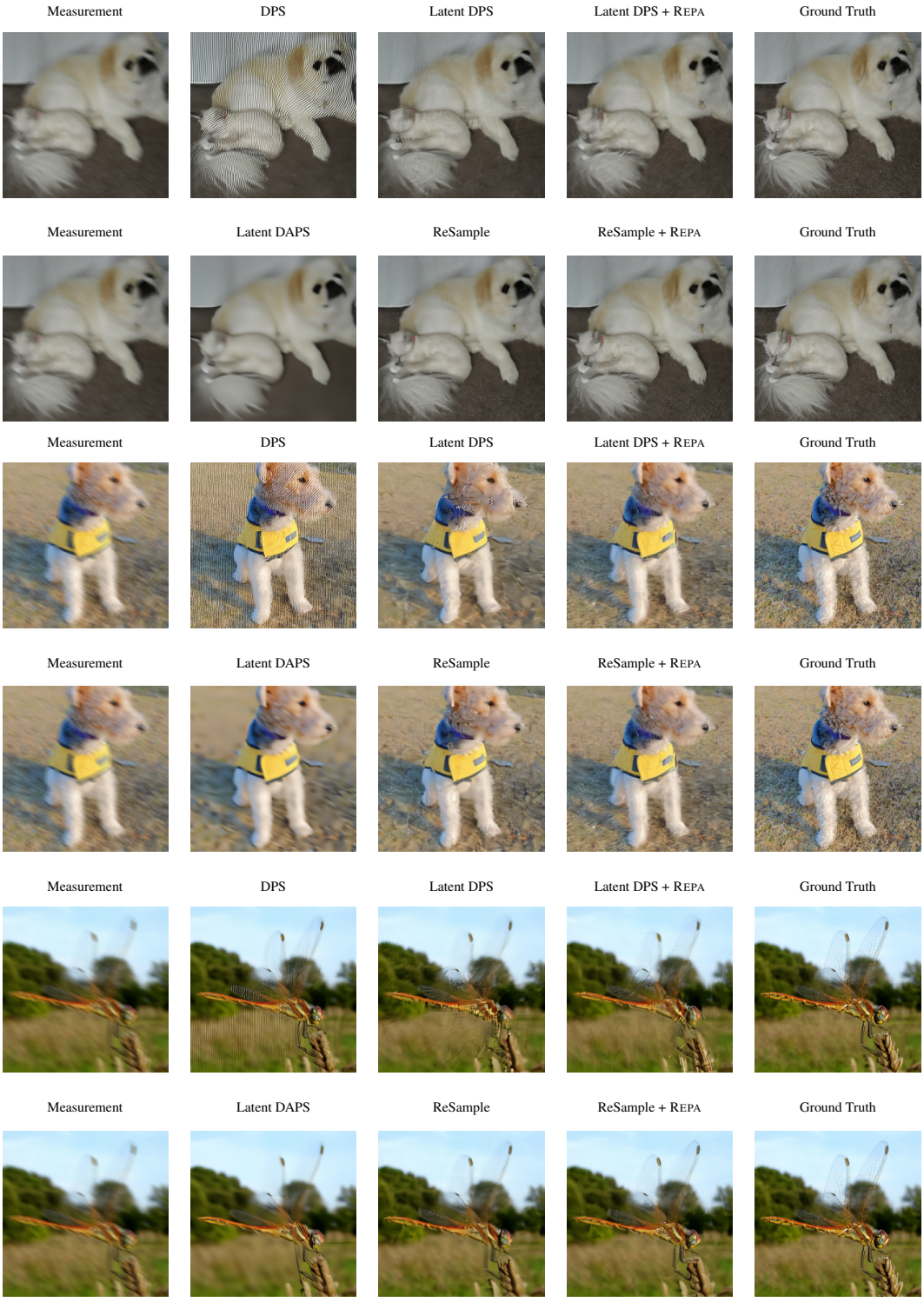


Figure 10: Qualitative comparison for Motion Deblurring. Each row shows: Measurement, a baseline method (DPS or DAPS), a latent solver (Latent DPS or ReSample), its REPA-enhanced version, and the Ground Truth.

Figure 11: Qualitative comparison for box inpainting on FFHQ dataset.