Blind Inverse Problem Solving Made Easy by Text-to-Image Latent Diffusion

Michail Dontas[1*], Yutong He[1*], Naoki Murata[2], Yuki Mitsufuji[2,3], J. Zico Kolter[1,4], Ruslan Salakhutdinov[1]
Carnegie Mellon University[1] Sony AI[2] Sony Group Corporation[3] Bosch Center for AI[4]
{mdontas, yutonghe, zkolter, rsalakhu}@cs.cmu.edu, {naoki.murata, yuhki.mitsufuji}@sony.com

Abstract

Blind inverse problems, where both the target data and forward operator are unknown, are crucial to many computer vision applications. Existing methods often depend on restrictive assumptions such as additional training, operator linearity, or narrow image distributions, thus limiting their generalizability. In this work, we present LADiBI, a training-free framework that uses large-scale text-to-image diffusion models to solve blind inverse problems with minimal assumptions. By leveraging natural language prompts, LADiBI jointly models priors for both the target image and operator, allowing for flexible adaptation across a variety of tasks. Additionally, we propose a novel posterior sampling approach that combines effective operator initialization with iterative refinement, enabling LADiBI to operate without predefined operator forms. Our experiments show that LADiBI is capable of solving a broad range of image restoration tasks, including both linear and nonlinear problems, on diverse target image distributions.

*Equal contribution

1. Introduction

Inverse problems are fundamental to many computer vision tasks, where the goal is to recover unknown data $\boldsymbol{x}$ from observed measurements $\boldsymbol{y}$. Formally, these problems can be expressed as $\boldsymbol{y} = \mathcal{A}_\phi(\boldsymbol{x}) + \boldsymbol{n}$, where $\mathcal{A}$ is a operator representing the forward process parametrized by ϕ, and $\boldsymbol{n}$ is the measurement noise. These problems are critical in fields such as medical imaging, computational photography and signal processing, as they are inherently tied to real-world scenarios like image decompression, deblurring, and super-resolution [13, 17, 41, 43]. Recently, deep learning-based methods [7, 8, 10, 22, 30, 44] have emerged as powerful tools for solving inverse problems. In particular, diffusion models offer an elegant solution that incorporates pre-trained models in diverse conditional settings, providing flexible and scalable frameworks for reconstructing high-quality data by guiding the generation process through the measurement constraints [3, 14, 19, 36, 37, 42].

However, most of these efforts focus on inverse problems where the operator $\mathcal{A}_\phi$ is assumed to be *known*. In practice, the forward operator is often *unknown*, leading to what are termed *blind inverse problems*. Solving these blind inverse

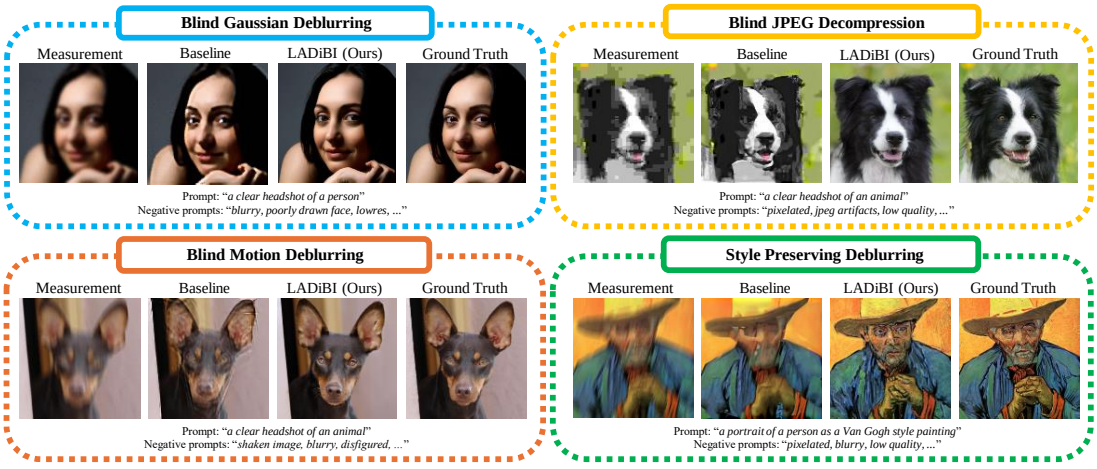

Figure 1. Our proposed **LADiBI** is a training-free blind inverse problem solving algorithm using large pre-trained text-to-image diffusion models. LADiBI is applicable to a wide variety of image distribution as well as operators with minimal modeling assumptions imposed.

problems remains a significant challenge due to their ill-posed nature. Consequently, current methods address this problem by introducing various assumptions about the prior distributions of the target data and unknown operator. These assumptions are typically incorporated through a combination of the following strategies: (1) using explicit formulas or specific neural network architectures to hand-craft inductive biases, (2) simplifying the operator with linear assumptions, (3) restricting target distributions to narrow image classes, (4) collecting measurement, operator or image datasets to train models tailored to specific tasks. While effective in certain cases, these solutions often entail significant computational and manual overhead, limited flexibility, and substantial barriers to practical deployment due to complex model training, costly data curation, unrealistic assumptions, and laborious hyperparameter tuning. These limitations raise a critical question: *can we design an algorithm that applies to diverse operators and target image distributions without additional training or data collection?*

To tackle this challenge, we formulate this problem as a Bayesian inference problem, where the optimal solution is to sample from the posterior $p(\boldsymbol{x}, \mathcal{A}_\phi|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{A}_\phi)p(\boldsymbol{x}, \mathcal{A}_\phi)$. State-of-the-art diffusion-based methods often simplify this problem by assuming independence between the target image and operator distributions [8, 23, 27, 34]. Although this independence assumption reduces the modeling difficulty, we challenge its validity by observing the correlation between the target image distribution and the operator distribution, and argue that this assumption heavily restricted the flexibility of the current diffusion-based approaches. To address this problem effectively, we first propose to estimate the joint distribution $p(\boldsymbol{x}, \mathcal{A}_\phi)$ instead. While this estimation seems intractable on the first glance, there actually exists a surprisingly simple solution – large-scale pre-trained text-to-image diffusion models already capture diverse target data and measurement distributions. Moreover, many inverse problems can be described using natural language, allowing tasks like deblurring or super-resolution to be expressed through prompts such as "high-definition, clear image" for the target and "blurry, low-quality" for the measurement. By simply prompting with classifier-free guidance [15], we can approximate the score of this challenging joint prior distributions across a wide range of image and operator types using the same base model, significantly extending the flexibility of the current diffusion-based framework.

In addition to a strong prior, an effective posterior sampling technique is also crucial for meeting measurement constraints. Existing methods generally fall into two categories: Gibbs or EM-style alternating optimization [23, 27], which suits diffusion-based setups without additional training but is sensitive to operator initialization and tuning, and a two-step approach that first estimates the operator followed by solving the non-blind problem [34], which often requires training and strong assumptions. To obtain the best of both worlds, we propose a hybrid approach: first, we obtain a reasonable initial estimate of the operator parameters, and then apply the alternating optimization to iteratively refine both operator parameters and data estimates throughout the diffusion process. Obtaining a reliable initial operator estimate can be difficult for broad functional classes like neural networks. To address this, we introduce a novel initialization scheme that leverages the pseudo-supervision signals from multiple lower-quality data estimation generated by fast posterior diffusion sampling. This approach eliminates the assumption of a specific form of operators, allowing for nonlinear blind inverse problem solving and flexible operator parametrization.

Combining the strong prior with the effective posterior sampling technique, we introduce **L**anguage-**A**ssisted **Di**ffusion for **B**lind **I**nverse problems (**LADiBI**), a training-free method that leverages large-scale text-to-image diffusion models to solve a broad range of blind inverse problems with minimal assumptions. LADiBI can be directly applied across diverse data distributions and allows for easy specification of task-specific modeling assumptions through simple prompt adjustments. Algorithm 1 provides an overview of LADiBI, which can be easily adapted from the standard inference algorithm used in popular text-to-image diffusion models. Unlike existing methods, LADiBI requires no model retraining or reselection for different target data distributions or operator functions; instead, all prior parameterization is encoded directly in the prompt, which users can adjust as needed. Notably, we do not assume linearity of the operator, making LADiBI, to the best of our knowledge, the most generalizable approach to blind inverse problem solving in image restoration.

We evaluate LADiBI against state-of-the-art baselines across several blind image restoration tasks, including two linear inverse problems (motion deblurring and Gaussian deblurring) and a nonlinear inverse problem (JPEG decompression), on various image distributions, as illustrated in Figure 1. In the linear cases, our method matches the performance of the state-of-the-art approaches, which require extensive assumptions and model specifications. Additionally, LADiBI is the only method tested that successfully performs JPEG decompression without any information about the task, such as the compression algorithm, the quantization table or factors, relying solely on observations of the compressed images.

2. Background & Related Works

Diffusion for Inverse Problem Solving Diffusion models, also known as score-based generative models [16, 38], generate clean data samples $\boldsymbol{x}_0$ by iteratively refining noisy samples $\boldsymbol{x}_t$ using a time-dependent score function

Table 1. A conceptual comparison between our proposed LADiBI and the existing literature.

Method Family	Method	Prior Type	Diverse Image Prior	No Additional Training	Flexible Operator
Optimization-based	Pan-ℓ_0 [28]	Inductive bias	×	✓	×
	Pan-DCP [29]	Inductive bias	×	✓	×
Unsupervised	SelfDeblur [30]	Inductive bias	×	×	✓
Supervised	MPRNet [44]	Discriminative	×	×	✓
	DeblurGANv2 [22]	GAN	×	×	✓
Diffusion-based	BlindDPS [8]	Pixel diffusion	×	×	×
	GibbsDDRM [27]	Pixel diffusion	×	✓	×
	LADiBI (Ours)	Text-to-image latent diffusion	✓	✓	✓

$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$. This score function is usually parametrized as a noise predictor $\epsilon_\theta(\boldsymbol{x}_t, t)$ and can be used to produce the clean data samples by iteratively denoising the noisy samples from the noise prediction. Particularly, a popular sampling algorithm DDIM [35] adopts the update rule

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\boldsymbol{x}_t, t)}{\sqrt{\bar{\alpha}_t}} \right)}_{\text{intermediate estimation of } \boldsymbol{x}_0, \text{ denoted as } \boldsymbol{x}_{0|t}} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(\boldsymbol{x}_t, t) + \sigma_t \epsilon \tag{1}$$

that consists of an intermediate estimation of the clean data in order to perform fast sampling.

Because of this iterative sampling process and its close connection to the score function, many efforts attempt to use unconditionally pretrained diffusion models for conditional generation [11, 26, 37, 38], especially inverse problem solving [7, 9, 19, 36], without additional training. Generally, when attempting to sample from the conditional distribution $p(\boldsymbol{x}|\boldsymbol{y})$, they decompose its score as

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{y}) = \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) + \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y}|\boldsymbol{x}_t) \tag{2}$$

Since $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ can be obtained from an unconditionally pre-trained diffusio model, these methods usually aim at deriving an accurate approximation for $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{y}|\boldsymbol{x}_t)$.

As text-to-image latent diffusion models become increasingly popular and powerful [2, 33, 45], recent work has explored using these large language-conditioned models for inverse problem-solving. In particular, MPGD [14] addresses noisy inverse problems by leveraging the natural manifold preserving property of the latent diffusion models. Specifically, it modifies the intermediate clean latent estimate $\boldsymbol{z}_{0|t}$ using the following posterior update rule

$$\boldsymbol{z}_{0|t} = \boldsymbol{z}_{0|t} - c_t \nabla_{\boldsymbol{z}_{0|t}} \|\boldsymbol{y} - \mathcal{A}_\phi(\mathbf{D}(\boldsymbol{z}_{0|t}))\|_2^2 \tag{3}$$

where $\mathbf{D}(\boldsymbol{z}_{0|t})$ represents the decoded intermediate clean image estimation and c_t is the step size hyperparameter to account for the unknown noise level σ. The L_2 loss $\|\boldsymbol{y} - \mathcal{A}_{\hat{\phi}}(\mathbf{D}(\boldsymbol{z}_{0|t}))\|_2^2$ is induced by the additive Gaussian noise assumption from conventional inverse problem setting. Although these diffusion-based methods performs well on diverse data distributions and tasks, they all require the operator $\mathcal{A}_{\hat{\phi}}$ to be known.

Blind Inverse Problem Blind inverse problems aim to recover unknown data $\boldsymbol{x} \in \mathbb{R}^d$ from measurements $\boldsymbol{y} \in \mathbb{R}^m$, typically modeled as:

$$\boldsymbol{y} = \mathcal{A}_\phi(\boldsymbol{x}) + \boldsymbol{n} \tag{4}$$

where $\mathcal{A}_\phi : \mathbb{R}^d \to \mathbb{R}^m$ is the forward operator parameterized by unknown function ϕ, and $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2 I) \in \mathbb{R}^m$ represents measurement noise with variance $\sigma^2 I$. Blind inverse problems are more difficult due to the joint estimation of $\boldsymbol{x}$ and $\mathcal{A}_\phi$, and are inherently ill-posed without further assumptions. As a result, existing methods typically incorporate different assumptions about the priors of the target image data as well as the unknown operator.

Conventional methods use hand-crafted functions as image and operator prior constraints [5, 20, 24, 28, 29]. They often obtain these functional constraints by observing certain characteristics (e.g. clear edges and sparsity) unique to certain distributions that are usually considered as the target data distribution (e.g. clear and well-calibrated photos of the real world) and the operator distribution (e.g. some blurring kernels). However, not only are these hand-made functions not generalizable, they also often require significant manual tuning efforts for each individual image.

With the rise of deep learning, neural network has become a popular choice for parameterizing priors [12, 21, 22, 30, 40, 44]. In particular, Kupyn et al. [21, 22], Zamir et al. [44] use supervised learning on data-measurement pairs to jointly learn the image-operator distributions, and Ren et al. [30], Ulyanov et al. [40] apply unsupervised learning solely on measurements by leveraging deep convolutional networks' inherent image priors. These methods offer significant improvement over traditional approaches. However, their learning procedures require separate data collection and model training for each image distribution

and task, making them resource-intensive. Although some attempt to generalize by training on a wide range of data and measurements, our experiments show that they still struggle to adapt to unseen image or operator distributions.

Recently, inspired by advances in diffusion models for inverse problem solving, numerous efforts have incorporated the generative priors from pre-trained diffusion models. However, most of these methods still lack generalizability [8, 23, 27, 34]. For instance, Chung et al. [8], Sanghvi et al. [34] require training separate operator priors, while Murata et al. [27] remains training-free but rely on the operator kernel's SVD for feasible optimization. These approaches generally depend on linear assumptions about the operator and well-trained diffusion models tailored to specific image distributions, limiting their ability to generalize across diverse image and operator types.

Concurrent to our work, two methods [1, 39] attempt to address these limitations. However, Tu et al. [39] only slightly relaxes the operator constraints to linear function with additive masking. It also only uses pixel diffusion rather than more powerful latent diffusion models. Anonymous [1] relies on hand-crafted update rules for each task, lacking a truly generalizable framework. More importantly, neither method leverages the flexibility provided by text prompts in large text-to-image diffusion models.

Table 1 summarizes the conceptual difference between out method and popular approaches in current literature. By leveraging large-scale pre-trained text-to-image latent diffusion models and our new posterior sampling algorithm, our method offers the most generalizability across diverse image and operator distributions with no additional training.

3. Method

In this paper, we aim to tackle the problem of blind inverse problem solving defined in Equation 4. Our solution (Algorithm 1) has the following desiderata: (1) **No additional training:** our method should not require data collection or model training, (2) **Adaptability to diverse image priors:** the same model should apply to various image distributions, (3) **Flexible operators:** no assumptions about the operator's functional form, such as linearity or task-specific update rules, should be necessary. To make this problem feasible, we assume access to open-sourced pre-trained models.

To tackle this problem, we first formulate it as a Bayesian inference problem where the optimal solution is to sample from the posterior

$$p(\boldsymbol{x}, \mathcal{A}_\phi | y) \propto p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{A}_\phi) p(\boldsymbol{x}, \mathcal{A}_\phi). \tag{5}$$

This formulation allows us to decompose this problem into two parts: first obtaining a reliable estimation of the prior $p(\mathcal{A}_\phi)$ and then maximizing the measurement likelihood $p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{A}_\phi)$. This makes diffusion-based framework the ideal approach as its posterior sampling process naturally separates these two stages (by Equation 2).

Algorithm 1 Our algorithm LADiBI

/* Initialize latent with encoded measurement $\boldsymbol{y}$ & SDEdit */
$\boldsymbol{z}_{T_s} = \sqrt{\bar{\alpha}_{T_s}}\mathbf{E}(y) + \sqrt{1-\bar{\alpha}_{T_s}}\epsilon_{T_s}$ for $\epsilon_{T_s} \sim \mathcal{N}(0, I)$
Initialize $\hat{\phi}$ with a fixed operator prior or Algorithm 2
for $t = T_s, \dots, 1$ **do**
  /* Use time-traveling for more stable results */
  for $j = 1, \dots, M$ **do**
    /* Use CFG to obtain accurate prior */
    Calculate $\hat{\epsilon}_\theta(\boldsymbol{z}_t, t)$ with Eq. 7
    $\boldsymbol{z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}_\theta(\boldsymbol{z}_t, t))$
    if $j \equiv 0 \pmod 2$ **then**
      /* Perform MPGD with the estimated $\mathcal{A}_{\hat{\phi}}$ */
      Update $\boldsymbol{z}_{0|t}$ with Eq. 8
    end if
    $\boldsymbol{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{z}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(\boldsymbol{z}_t, t) + \sigma_t\epsilon_t$ for $\epsilon_t \sim \mathcal{N}(0, I)$
    $\boldsymbol{z}_t = \sqrt{\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}\boldsymbol{z}_t + \sqrt{1-\frac{\bar{\alpha}_t}{\bar{\alpha}_{t-1}}}\epsilon'_t$ for $\epsilon'_t \sim \mathcal{N}(0, I)$
  end for
  /* Periodically update $\hat{\phi}$ with $\boldsymbol{z}_{0|t}$ and Eq. 9 */
  if $t \equiv 0 \pmod 5$ **then**
    for $k = 1, \dots, K$ **do**
      $\hat{\phi} = \text{Adam}(\hat{\phi}, L_\phi(\boldsymbol{z}_{0|t}, \hat{\phi}))$
    end for
  end if
end for
return $\boldsymbol{x}_0 = \mathbf{D}(\boldsymbol{z}_0)$

3.1. Obtaining a better Prior

Most existing diffusion-based methods [1, 8, 23, 27, 34, 39] avoid modeling the joint prior and instead further decompose the goal distribution as $p(\boldsymbol{x}, \mathcal{A}_\phi|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{A}_\phi)p(\boldsymbol{x})p(\mathcal{A}_\phi)$, under the assumption that the target data distribution $p(\boldsymbol{x})$ independently from the operator distribution $p(\mathcal{A}_\phi)$. In this way, they can separately model the image and operator distribution by pre-trained diffusion models or fixed distributions and thus bypass the combinatorial complexity. However, is this truly a fair assumption?

Consider the task of deblurring and super-resolution. A low-resolution blurry measurement may correspond to a high-resolution blurry image (e.g. an out of focus DSLR photo) if the operator is a downsampler, or to a low-resolution clear image if the operator is a blur kernel. Similarly, different target image distribution can imply different operators: for example, the estimated blur kernel should differ if the target is a realistic photo versus an Impressionist painting, even when both share the same measurement.

These observations suggest that $p(\boldsymbol{x})$ and $p(\mathcal{A}_\phi)$ should be correlated, and previous methods can make this independence assumption because they implicitly restrict the types of operations and target images they handle. In other

words, while this independence assumption reduces modeling complexity, it also significantly limits the flexibility of the resulting algorithms. Consequently, most existing methods require additional training or parametrization of the operator prior and need to switch the image prior models when the target distribution changes.

Therefore, to build a unified training-free algorithm, we should directly model $p(\boldsymbol{x}, \mathcal{A}_\phi)$, or in the case of diffusion models, $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t, \mathcal{A}_\phi)$. On the first glance, modeling $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t, \mathcal{A}_\phi)$ appears intractable. However, with large-scale pre-trained text-to-image diffusion models now widely available, we have a surprisingly simple solution:

In practice, the target data distribution (e.g. "a high-resolution photo") and the forward operation (e.g. "Gaussian blur") can often be effectively described in natural language. We can specify the target distribution by using its description as a positive prompt. Specifying the operator distribution is less direct, as these pre-trained models are designed to represent image distributions rather than operators. However, by leveraging classifier-free guidance (CFG) [15], we can implicitly define the operator distribution through negative prompts. For example, if the operator is Gaussian blur, an effective negative prompt can be "blurry image, low quality". Formally, denote positive prompt as $\boldsymbol{c}_+$ and negative prompt as $\boldsymbol{c}_-$, we can approximate $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t, \mathcal{A}_\phi)$ as

$$\begin{aligned}\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t, \mathcal{A}_\phi) \approx \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{c}_-) \\ + \gamma(\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{c}_+) - \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{c}_-))\end{aligned} \quad (6)$$

where $\gamma > 1$ is a weighting hyperparameter. This is equivalent to approximating $p(\boldsymbol{x}, \mathcal{A}_\phi)$ as a distribution that is proportional to $p(\boldsymbol{x}|\boldsymbol{c}_-)^{1-\gamma} p(\boldsymbol{x}|\boldsymbol{c}_+)^\gamma$.

When using latent diffusion models parameterized by θ, this corresponds to having the empirical noise prediction

$$\hat{\epsilon}_\theta(\boldsymbol{z}_t, t) = \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_-) + \gamma(\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_+) - \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_-)) \quad (7)$$

This way, our method not only models this seemingly extremely intractable joint distribution in a remarkably simple way, it also enable wide range of applications with diverse target image and operator distributions by leveraging the knowledge in large-scale pre-trained text-to-image diffusion models like Stable Diffusion[31].

3.2. Obtaining a better Posterior

In addition to a strong prior, our final output image should also satisfy the measurement constraint.Given the problem setup in Equation 4, the measurements are subject to additive Gaussian noise $\boldsymbol{n}$, hence $\log p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{A}_\phi) = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathcal{A}_\phi(\boldsymbol{x})\|_2^2$. When $\mathcal{A}_\phi$ is known, we can use the MPGD update rule in Equation 3 to perform posterior sampling.

However, since $\mathcal{A}_\phi$ is not known, the true parameters ϕ is often approximated by another set of parameters $\hat{\phi}$. This approximation is usually addressed by one of the two strategies: an alternating optimization scheme that jointly approximates $\boldsymbol{x}$ and ϕ, or obtaining a reliable $\hat{\phi}$ first then solving a non-blind inverse problem. The first approach is well-suited for training-free settings as it can be well integrated with diffusion-based methods, but it is often highly sensitive to $\hat{\phi}$ initialization and tuning. The second approach can perform well if a strong $\hat{\phi}$ is obtained, though it usually requires training and additional assumptions or restrictions.

We propose to combine the two strategies above: we first obtain a reliable initial $\hat{\phi}$, and then perform an alternating optimization to iteratively refine both the operator parameter and data estimations throughout the diffusion process.

Algorithm 2 General Operator Initialization

for $j = 1, \ldots, M$ **do**
  /* Initialize latent batch with encoded $\boldsymbol{y}$ & SDEdit */
  $\epsilon_{T_s}^{(i)} \sim \mathcal{N}(0, I)$ for $i = 1, 2, \ldots, N$
  $\boldsymbol{z}_{T_s}^{(i)} = \sqrt{\bar{\alpha}_{T_s}}\mathbf{E}(y) + \sqrt{1 - \bar{\alpha}_{T_s}}\epsilon_{T_s}^{(i)}$
  /* Use MPGD to obtain latent estimations $\{\boldsymbol{z}_{0|t}^{(i)}\}_{i=1}^N$ */
  for $t = T_j, \ldots, 1$ and all i in parallel **do**
    Calculate $\hat{\epsilon}_\theta(\boldsymbol{z}_t^{(i)}, t)$ with Eq. 7
    $\boldsymbol{z}_{0|t}^{(i)} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{z}_t^{(i)} - \sqrt{1 - \bar{\alpha}_t}\hat{\epsilon}_\theta(\boldsymbol{z}_t^{(i)}, t))$
    if $j \neq 1$ **then**
      Update $\boldsymbol{z}_{0|t}^{(i)}$ with Eq. 8
    end if
    $\boldsymbol{z}_{t-1}^{(i)} = \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{z}_{0|t}^{(i)} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(\boldsymbol{z}_t^{(i)}, t) + \sigma_t\epsilon_t^{(i)}$ for $\epsilon_t^{(i)} \sim \mathcal{N}(0, I)$
  end for
  /* Update $\hat{\phi}$ with $\{\boldsymbol{z}_{0|t}^{(i)}\}_{i=1}^N$ and Eq. 9 */
  for $k = 1, \ldots, K$ **do**
    $\hat{\phi} = \text{Adam}(\hat{\phi}, \frac{1}{N}\sum_{i=1}^N L_\phi(\boldsymbol{z}_0^{(i)}, \hat{\phi}))$
  end for
end for
return $\hat{\phi}$

$\hat{\phi}$ Initialization In some cases, initializing the operator parameters is straightforward – for example, for a slight blur, an identity function can be a reasonable stating point. Alternatively, reliable operator priors, such as the pre-trained prior in BlindDPS [8] or the Gaussian prior in GibbsDDRM [27] for linear operators, can also provide effective initializations for simple kernels. However, when the operator requires complex functions like neural networks, obtaining good initializations becomes more challenging.

To address this, we propose a new algorithm for general operator initialization by drawing further inspiration from alternating optimization algorithms. Note that since the goal here is only to initialize $\hat{\phi}$, the quality of intermediate $\boldsymbol{x}$ estimates is less critical as long as they provide a useful signal. Unlike other diffusion-based methods that alternate optimization targets at each diffusion step, we use SDEdit [26] and MPGD [14] with very few diffusion steps

to quickly obtain a batch of $\boldsymbol{x}$ estimates. We then perform maximum likelihood estimation (MLE) on these estimates to update $\hat{\phi}$ using Adam optimizer. As detailed in Algorithm 2, repeating this process can leverage the diffusion prior to quickly converge to a reliable starting point for $\hat{\phi}$.

Notice that we only assume that ϕ can be approximated by differentiable functions, allowing $\mathcal{A}_{\hat{\phi}}$ to be parameterized by general model families such as neural networks. Moreover, as mentioned earlier, any well-performing operator priors can be seamlessly integrated into our framework.

Iterative Refinement After initializing $\hat{\phi}$, we then perform another alternating optimization scheme to refine both the operator and the data. In particular, we alternate between updating $\boldsymbol{z}_{0|t}$ using MPGD guidance with $\mathcal{A}_{\hat{\phi}}$ fixed and updating the MLE estimation of $\hat{\phi}$ using Adam with $\boldsymbol{z}_{0|t}$ fixed throughout the diffusion process.

Since unlike GibbsDDRM, which uses Langevin dynamics to update the operator, we solve for a local optimum. Therefore, it's reasonable not to update the operator too frequently. Empirically, we find that periodic updates to the operator in combination of the time-traveling technique [25, 42] yield the best results

In addition, since MPGD supports any differentiable loss function, we can incorporate techniques like regularization to further improve the visual quality. In practice, we use

$$\boldsymbol{z}_{0|t} = \boldsymbol{z}_{0|t} - c_t \nabla_{\boldsymbol{z}_{0|t}} (\|\boldsymbol{y} - \mathcal{A}_{\hat{\phi}}(\mathbf{D}(\boldsymbol{z}_{0|t}))\|_2^2 + \lambda_{\boldsymbol{z}} \text{LPIPS}(\mathbf{D}(\boldsymbol{z}_{0|t}), \boldsymbol{y})) \tag{8}$$

as the MPGD update rule and

$$L_\phi(\boldsymbol{z}_{0|t}, \hat{\phi}) = \|\boldsymbol{y} - \mathcal{A}_{\hat{\phi}}(\mathbf{D}(\boldsymbol{z}_{0|t}))\|_2^2 + \lambda_\phi \|\hat{\phi}\|_1 \tag{9}$$

as the Adam objective. Here $\text{LPIPS}(\cdot)$ denotes the LPIPS distance between two images and $\|\cdot\|_1$ denotes the L_1 regularization term. $\lambda_{\boldsymbol{z}}$ and λ_ϕ are adjustable hyperparameters.

By combining large-scale language-conditioned diffusion priors, effective operator initializations, and the iterative refinement process described above, we introduce LADiBI (Algorithm 1), a new training-free algorithm for blind inverse problem solving that supports diverse target image distributions and flexible operators.

4. Experiments

4.1. Experimental Setup

We empirically verify the performance of our method with two linear deblurring tasks, Gaussian deblurring and motion deblurring, and a non-linear restoration task, JPEG decompression. Here we provide the details of our experiments.

Datasets We conduct quantitative comparisons on the datasets FFHQ 256 × 256 [18] and AFHQ-dog 256 × 256 [6]. We use a size of 1000 images for FFHQ and 500 for AFHQ, following the setting in Murata et al. [27].

Baselines We compare our method against other state-of-the-art approaches as baselines. Specifically, we choose Pan-l0 [28] and Pan-DCP [29] as the optimization-based method, SelfDeblur [30] as the self-supervised approach, PRNet [44] and DeblurGANv2 [22] as the supervised baselines, and BlindDPS [8] and GibbsDDRM [27] as the diffusion-based methods. All baselines are experiments using their open-sourced code and pre-trained models, whose details are provided in the appendix.

Implementation Details We conduct all experiments an NVIDIA A6000 GPU. We use Stable Diffusion 1.4 [31], DDIM 200-step noise scheduling, and $T_s = 150, M = 4$ for all tasks. We set positive prompts as *"a clear headshot of a person"* and *"a clear headshot of an animal"* for FFHQ and AFHQ experiments respectively. We adjust negative prompts according to each individual task, and the details will be provided in the following sections.

Evaluation Metrics We evaluate our method and the baselines with three quantitative metrics: LPIPS [46], PSNR and KID [4]. Each metric has its own significance and reveals a different set of characteristics about performance. In particular, PSNR directly measures the pixel-by-pixel accuracy of the reconstructed image compared to the original, while LPIPS measures how similar two images are in a way that reflects human perception. Finally, KID evaluates the image fidelity on a distribution level.

4.2. Blind Linear Deblurring

We first test our method against the baselines on linear deblurring, which is the design space of most baselines. We evaluate all methods on two types of blur kernels: Gaussian blur and motion blur.

Setup To generate measurements, we apply randomly generated motion blur kernels with intensity 0.5 and Gaussian blur kernels with standard deviation 3.0 respectively. The final measurements are derived by applying a pixel-wise Gaussian noise with $\sigma = 0.02$. We use a 61×61 convolutional matrix as $\hat{\phi}$ and we initialize it as a Gaussian kernel of intensity 6.0 following Murata et al. [27]. We adjust the negative prompts to the potential negative image artifacts induced by the degradation operation, such as *"lowres, blurry, DSLR effect, poorly drawn face, deformed, bad proportions, ..."* for Gaussian blur, *"shaken image, morbid, mutilated, text in image, disfigured, ..."* for motion blur. Full prompts are included in the appendix.

Results Table 2 and Figure 2 show results on blind linear deblurring. Our method matches the performance the state-of-the-art method GibbsDDRM, which is explicitly designed to solving linear problems using SVDs. In fact, GibbsDDRM's highly specialized design makes it so sensitive that even small discrepancies (e.g. kernel padding) between their modeling assumption and the ground truth

Table 2. Quantitative results on blind linear deblurring tasks. "GibbsDDRM" denotes the results from directly running their open-sourced code and "GibbsDDRM*" denotes the results after adjusting their code to match with the ground truth kernel padding.

Method	FFHQ						AFHQ					
	Motion			Gaussian			Motion			Gaussian		
	LPIPS ↓	PSNR ↑	KID ↓	LPIPS ↓	PSNR ↑	KID ↓	LPIPS ↓	PSNR ↑	KID ↓	LPIPS ↓	PSNR↑	KID ↓
SelfDeblur	0.732	9.05	0.1088	0.733	8.87	0.0890	0.742	9.04	0.0650	0.736	8.84	0.0352
MPRNet	0.292	22.42	0.0467	0.334	23.23	0.0511	0.324	22.09	0.0382	0.379	21.97	0.0461
DeblurGANv2	0.309	**22.55**	0.0411	0.325	26.61	0.0227	0.323	**22.74**	0.0350	0.340	**27.12**	0.0073
Pan-l0	0.389	14.10	0.1961	0.265	20.68	0.1012	0.414	14.16	0.1590	0.276	21.04	0.0320
Pan-DCP	0.325	17.64	0.1323	0.235	24.93	0.0490	0.371	17.63	0.1377	0.297	25.11	0.0263
BlindDPS	0.246	20.93	0.0153	0.216	25.96	0.0205	0.393	20.14	0.0913	0.330	24.79	0.0268
GibbsDDRM	0.293	20.52	0.0746	0.216	27.03	0.0430	0.303	19.44	0.0265	0.257	24.01	**0.0040**
LADiBI (Ours)	0.230	20.96	**0.0084**	0.197	21.08	**0.0068**	**0.262**	21.20	**0.0132**	**0.204**	24.33	0.0065
GibbsDDRM*	**0.199**	22.36	0.0309	**0.155**	**27.65**	0.0252	0.278	19.00	0.0180	0.224	21.62	**0.0040**

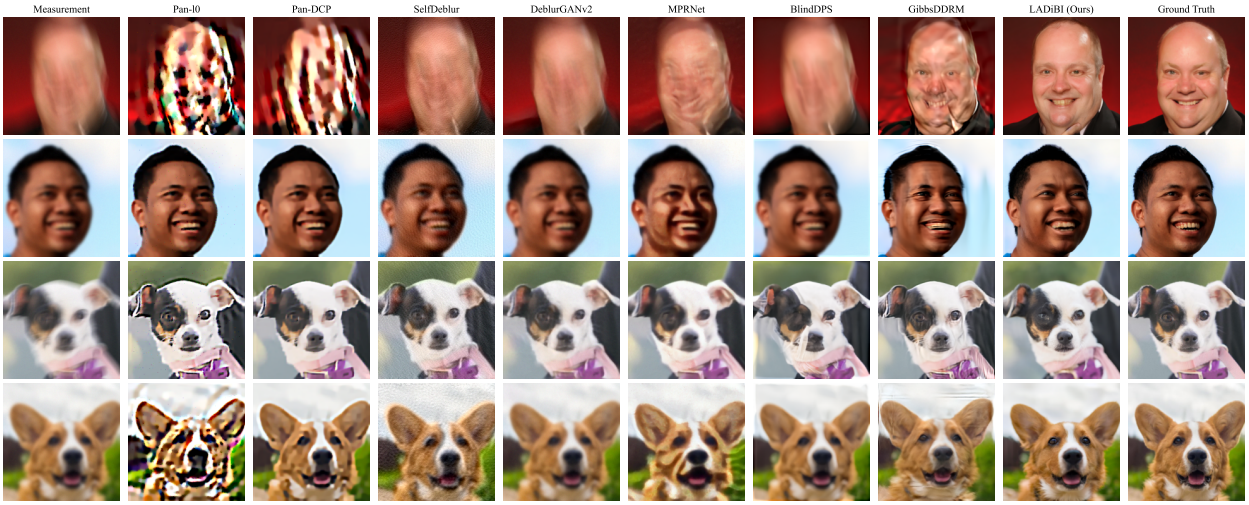

Figure 2. Qualitative results on blind linear deblurring tasks. From top to bottom we showcase examples from motion deblur on FFHQ, Gaussian deblur on FFHQ, motion deblur on AFHQ, and Gaussian deblur on AFHQ respectively.

operator can result in significant performance degradation. Moreover, although BlindDPS and GibbsDDRM use diffusion models that are trained on the exact data distributions tested, our method can still match or outperform them with the large-scale pre-trained model. In addition, even though supervised methods can obtain higher PSNR, our method produces the fewest artifacts, which is also validated by the LPIPS and KID scores. Overall, our method offers a consistently competitive performance on linear tasks, even though our method is designed for more general applications.

4.3. Blind JPEG Decompression

We further compare our method with baselines on JPEG decompression, a particularly challenging task due to the non-linear, non-differentiable nature of JPEG compression. Unlike traditional settings, our experiments provide no task-specific knowledge, such as the compression algorithm, quantization table, or factors – algorithms rely solely on the measurement images.

Table 3. Quantitative results on blind JPEG decompression task.

Method	FFHQ			AFHQ		
	LPIPS ↓	PSNR ↑	KID ↓	LPIPS ↓	PSNR ↑	KID ↓
SelfDeblur	0.676	8.84	0.1959	0.703	8.89	0.1173
MPRNet	0.785	5.99	0.8008	0.769	5.93	0.6809
DeblurGANv2	0.473	21.48	0.2207	0.502	**21.76**	0.1527
BlindDPS	0.431	**21.55**	0.1791	0.397	20.87	0.2108
GibbsDDRM	0.841	13.91	0.2915	0.775	14.59	0.2915
LADiBI (Ours)	**0.268**	21.40	**0.0172**	**0.315**	21.12	**0.0216**

Setup We generate measurements using JPEG compression with a quantization factor of 2. We use a neural network with a 3-layer U-net [32] to parametrize the operator. The operator is initialized using Algorithm 2 with $M = 8$ and $N = 4$. We use phrases including *"pixelated, low quality, jpeg artifacts, deformed, bad proportions, ..."* as negative prompt, whose full version is provided in the appendix.

Results Table 3 and Figure 3 present the results on the JPEG decompression task. It is evident through both quantitative and qualitative results that our method is the only one

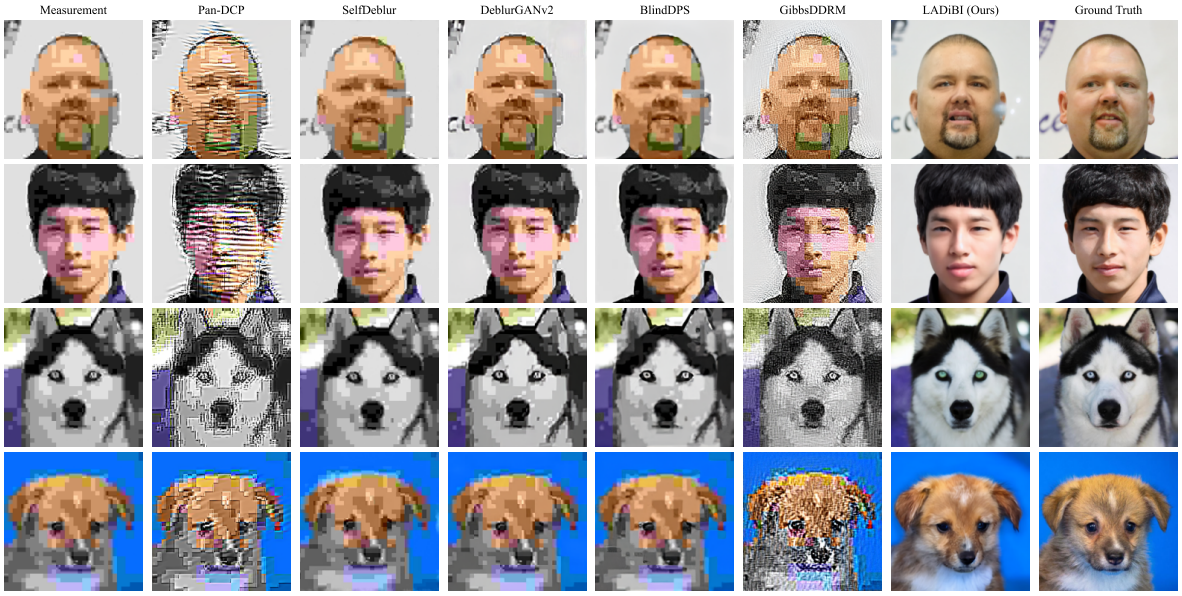

Figure 3. Qualitative results on the blind JPEG decompression task.

Table 4. Ablation study on our proposed algorithm.

Removed Component	LPIPS ↓	PSNR ↑	KID ↓
Text prompts	0.500	18.88	0.0483
MPGD guidance	0.470	17.80	0.0434
$\hat{\phi}$ Initialization	0.355	19.74	0.0250
LADiBI	**0.315**	**21.12**	**0.0216**

capable of enhancing the fidelity of the images and maintaining consistency to measurements. Unlike the baselines, which struggle with this task due to their limited posterior formulations, our flexible framework adapts to approximate this challenging operator and produces high quality images.

4.4. Ablation Studies

To showcase the effectiveness of each part of our algorithm, we conduct ablation study on AFHQ dataset with JPEG decompression task. In particular, we test the importance of suitable ad-hoc prompts, the use of MPGD guidance, and the operator initialization for neural network $\mathcal{A}_{\hat{\phi}}$. in-between diffusion steps. We keep SDEdit in all settings to encode the measurement information for fair comparisons. As shown in Table 4, each of the aforementioned components plays a significant role in our scheme and is indispensable for producing high quality results.

4.5. Style Preserving Deblurring

In order to demonstrate our method's ability to solve blind inverse problems from a wider range of image distributions, we present indicative examples of Gaussian and motion deblurring on Monet paintings in Figure 4. We use the same negative prompts as in the previous experiments and *"a portrait of a person as a Monet style painting"* as positive prompts. We notice that, although the images portray human faces, the baseline method using models trained on realistic human face images cannot accurately solve the problem, while our method effectively generates images consistent with both the measurement image and the painting style present in the ground truth image.

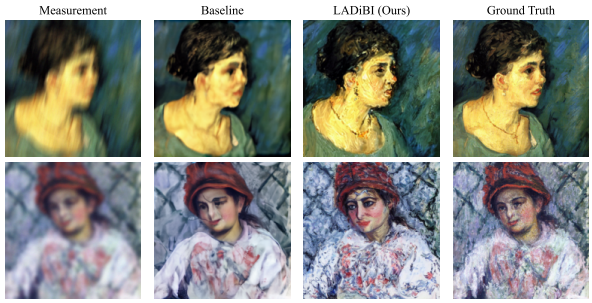

Figure 4. Qualitative results on blind deblurring Monet paintings.

5. Conclusion

In this work, we propose LADiBI, a new training-free algorithm to solving blind inverse problems in image restoration using a large-scale pre-trained text-to-image diffusion model. Our method leverages text prompts as well as posterior guidance on intermediate diffusion steps to guide the generation process towards clean reconstructions based on degraded measurements, in cases where the degradation operation is unknown. Experimental results demonstrate that LADiBI achieves competitive performance on tasks with diverse operator and is also capable of covering a wider range of image distributions compared to baselines.

Acknowledgment

This work is sponsored by Sony AI and is also supported in part by the ONR grant N000142312368.

References

[1] Anonymous. Blind inversion using latent diffusion priors. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. under review. 4

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2023. 3

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 843–852, 2023. 1

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *Proc. ICLR*, 2018. 6

[5] T.F. Chan and Chiu-Kwong Wong. Total variation blind deconvolution. *IEEE Transactions on Image Processing*, 7(3): 370–375, 1998. 3

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[7] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Proc. Adv. Neural Inform. Process. Syst.*, pages 25683–25696, 2022. 1, 3

[8] Hyungjin Chung, Jeongsol Kim, Sehui Kim, and Jong Chul Ye. Parallel diffusion models of operator and image for blind inverse problems. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 4, 5, 6, 15

[9] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proc. ICLR*, 2023. 3, 15

[10] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Fast diffusion sampler for inverse problems by geometric decomposition. *arXiv preprint arXiv:2303.05754*, 2023. 1

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Proc. Adv. Neural Inform. Process. Syst.*, pages 8780–8794, 2021. 3

[12] Yosef Gandelsman, Assaf Shocher, and Michal Irani. " double-dip": unsupervised image decomposition via coupled deep-image-priors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11026–11035, 2019. 3

[13] Hayit Greenspan. Super-resolution in medical imaging. *The Computer Journal*, 52(1):43–63, 2009. 1

[14] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 3, 5, 15

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 5

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Proc. Adv. Neural Inform. Process. Syst.*, 33:6840–6851, 2020. 2

[17] Jithin Saji Isaac and Ramesh Kulkarni. Super resolution techniques for medical image processing. In *2015 International Conference on Technologies for Sustainable Development (ICTSD)*, pages 1–6, 2015. 1

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, pages 4401–4410, 2019. 6

[19] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Proc. Adv. Neural Inform. Process. Syst.*, 2022. 1, 3

[20] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240, 2011. 3

[21] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 3

[22] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 1, 3, 6

[23] Charles Laroche, Andrés Almansa, and Eva Coupeté. Fast diffusion em: a diffusion model for blind inverse problems with application to deconvolution. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2, 4

[24] Anat Levin, Yair Weiss, Fredo Durand, and William T. Freeman. Understanding and evaluating blind deconvolution algorithms. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1964–1971, 2009. 3

[25] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proc. CVPR*, pages 11461–11471, 2022. 6

[26] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3, 5

[27] Naoki Murata, Koichi Saito, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. GibbsDDRM: A partially collapsed gibbs sampler for solving blind inverse problems with denoising diffusion restoration. In *International Conference on Machine Learning*, 2023. 2, 3, 4, 5, 6, 15

[28] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. l_0 -regularized intensity and gradient prior for deblurring text images and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(2):342–355, 2017. 3, 6

[29] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Deblurring images via dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(10):2315–2328, 2018. 3, 6

[30] Dongwei Ren, Kai Zhang, Qilong Wang, Qinghua Hu, and Wangmeng Zuo. Neural blind deconvolution using deep priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3, 6

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 5, 6

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 7, 11

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494. Curran Associates, Inc., 2022. 3

[34] Yash Sanghvi, Yiheng Chi, and Stanley H Chan. Kernel diffusion: An alternate approach to blind deconvolution. *arXiv preprint arXiv:2312.02319*, 2023. 2, 4

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. ICLR*, 2021. 3

[36] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *Proc. ICLR*, 2022. 1, 3

[37] Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *Proc. Int. Conf. Mach. Learn.*, pages 32483–32498. PMLR, 2023. 1, 3

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. ICLR*, 2021. 2, 3

[39] Siwei Tu, Weidong Yang, and Ben Fei. Taming generative diffusion for universal blind image restoration. *arXiv preprint arXiv:2408.11287*, 2024. 4

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 3

[41] J. Wang and K. Huang. Medical image compression by using three-dimensional wavelet transformation. *IEEE Transactions on Medical Imaging*, 15(4):547–554, 1996. 1

[42] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. *arXiv:2303.09833*, 2023. 1, 6

[43] Lu Yuan, Jian Sun, Long Quan, and Heung-Yeung Shum. Image deblurring with blurred/noisy image pairs. In *ACM SIGGRAPH 2007 Papers*, page 1–es, New York, NY, USA, 2007. Association for Computing Machinery. 1

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 1, 3, 6

[45] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, and Junmo Kim. Text-to-image diffusion models in generative ai: A survey, 2024. 3

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, pages 586–595, 2018. 6

A. Additional Implementation Details

We offer more detailed information regarding the implementation setup for Algorithms 1 and 2. All experiments are conducted using an NVIDIA A6000 GPU. In an effort to make our methodology generalizable and extensible to other inverse problems, we attempt to maintain the same hyper-parameter values across tasks wherever possible. Each hyper-parameter value has been chosen after conducting preliminary experiments for a specific range and opting for the value that offers the best performance. Indicative examples of such experiments are shown in C. In particular, we run Algorithm 1 for $T_s = 150$ timesteps while performing 4 repetitions as part of the time-traveling strategy. We encode the measurement as $\boldsymbol{x}_{T_s}$ by applying the forward diffusion process up to timestep 800. In parallel with the reverse diffusion process, we update $\hat{\phi}$ every 5 timesteps, each time conducting $K = 150$ gradient steps. We adust the c_t and the λ_ϕ parameters of Equations 8 and 9 to 30 and 2 respectively. In terms of the operator initialization process described by Algorithm 2, we make use of a batch of $N = 4$ samples and run $M = 8$ iterations, each comprising $T_j = 60$ timesteps.

Additionally, schematic overviews of Algorithms 1 and 2 are presented in Figures 5 and 6.

Furthermore, there are some parameters in our approach for which employing a task-aware setup strategy is essential for state-of-the-art performance. These are the prompt sentences as well as the architecture of $\hat{\phi}$. Here we provide details with respect to these parameters according to each restoration task:

Motion Deblurring
- Positive prompts: *"a clear headshot of a person/animal"*
- Negative prompts: *"shaken image, motion blur, pixelated, lowres, text, error, cropped, worst quality, blurry ears, low quality, ugly, duplicate, morbid, mutilated, poorly drawn face, mutation, deformed, blurry, dehydrated, blurry hair, bad anatomy, bad proportions, disfigured, gross proportions"*
- $\hat{\phi}$ architecture: A single 61×61 convolutional block with 3 input and 3 output channels.

Gaussian Deblurring
- Positive prompts: *"a clear headshot of a person/animal"*
- Negative prompts: *"blurry, gaussian blur, lowres, text, error, cropped, worst quality, blurry ears, low quality, ugly, duplicate, morbid, mutilated, text in image, DSLR effect, poorly drawn face, mutation, deformed, dehydrated, blurry hair, bad anatomy, bad proportions, disfigured, gross proportions"*
- $\hat{\phi}$ architecture: A single 61×61 convolutional block with 3 input and 3 output channels.

JPEG Decompression
- Positive prompts: *"a clear headshot of a person/animal"*
- Negative prompts: *"pixelated, lowres, text, error, cropped, worst quality, blurry ears, low quality, jpeg artifacts, ugly, duplicate, morbid, mutilated, text in image, DSLR effect, poorly drawn face, mutation, deformed, blurry, dehydrated, blurry hair, bad anatomy, bad proportions, disfigured, gross proportions"*
- $\hat{\phi}$ architecture: A neural network with a typical 3-layer U-net [32] architecture. Each layer consists of 2 convo-

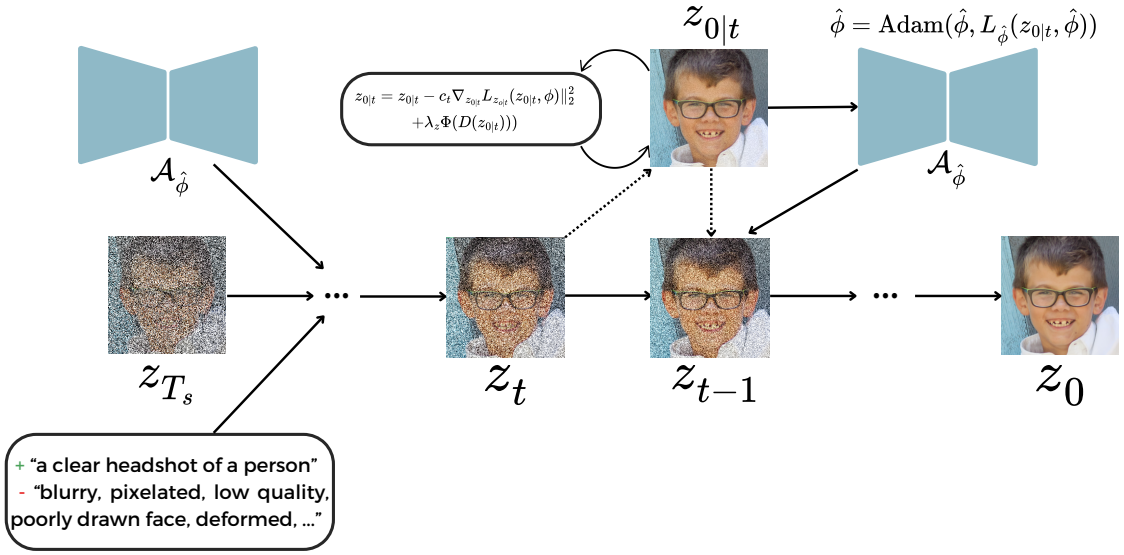

Figure 5. A schematic overview of Algorithm 1.

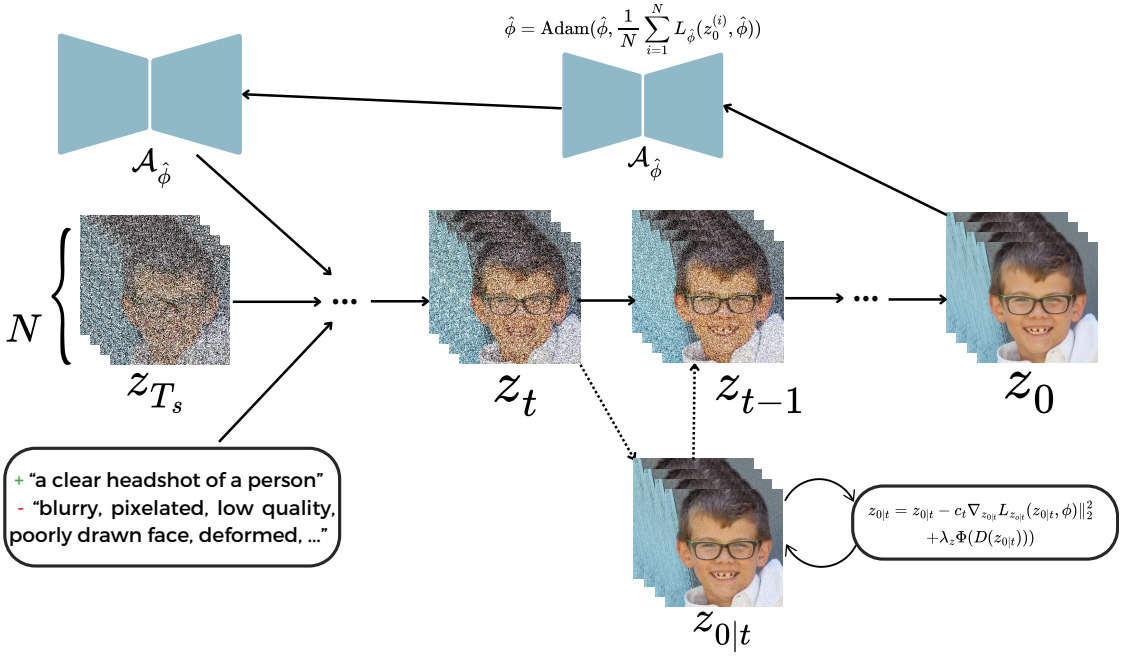

Figure 6. A schematic overview of Algorithm 2.

Table 5. Additional quantitative results on blind JPEG decompression task.

Method	FFHQ			AFHQ		
	LPIPS ↓	PSNR ↑	KID ↓	LPIPS ↓	PSNR ↑	KID ↓
Pan-l0	0.787	12.72	0.2190	0.825	13.15	0.2168
Pan-DCP	0.710	14.91	0.2318	0.673	18.27	0.2472
SelfDeblur	0.676	8.84	0.1959	0.703	8.89	<u>0.1173</u>
MPRNet	0.785	5.99	0.8008	0.769	5.93	0.6809
DeblurGANv2	0.473	<u>21.48</u>	0.2207	0.502	**21.76**	0.1527
BlindDPS	<u>0.431</u>	**21.55**	<u>0.1791</u>	<u>0.397</u>	20.87	0.2108
GibbsDDRM	0.841	13.91	0.2915	0.775	14.59	0.2915
LADiBI (Ours)	**0.268**	21.40	**0.0172**	**0.315**	<u>21.12</u>	**0.0216**

lutional blocks of size 3×3 with ReLU activations and number of input and output channels ranging from 32 to 128.

B. Additional Results

We present more detailed experimental results on all benchmark tasks. In specific, Table 5 shows quantitative results on the JPEG decompression task, including the optimization-based Pan-l0 and Pan-DCP methods. Moreover, Figures 7, 8 and 9 offer more qualitative comparisons against baseline methods on motion deblurring, gaussian deblurring and JPEG decompression respectively. Finally, in Fig. 10 we include comparisons against selected baseline methods on the 3 benchmark tasks using a set of Monet and Van Gogh portrait paintings as ground truth images.

C. Additional Ablation Study

We carefully adjust each parameter of our methodology to the value that offers the best overall results by conducting comparative experiments and in this section we offer some indicative examples of such tests. All hyper-parameter tuning experiments have been performed on the motion deblurring task using the AFHQ dataset.

Figure 11 presents both the LPIPS and KID score for the value of the timestep at which we encode the measurement using the forward diffusion process. Both metrics form a

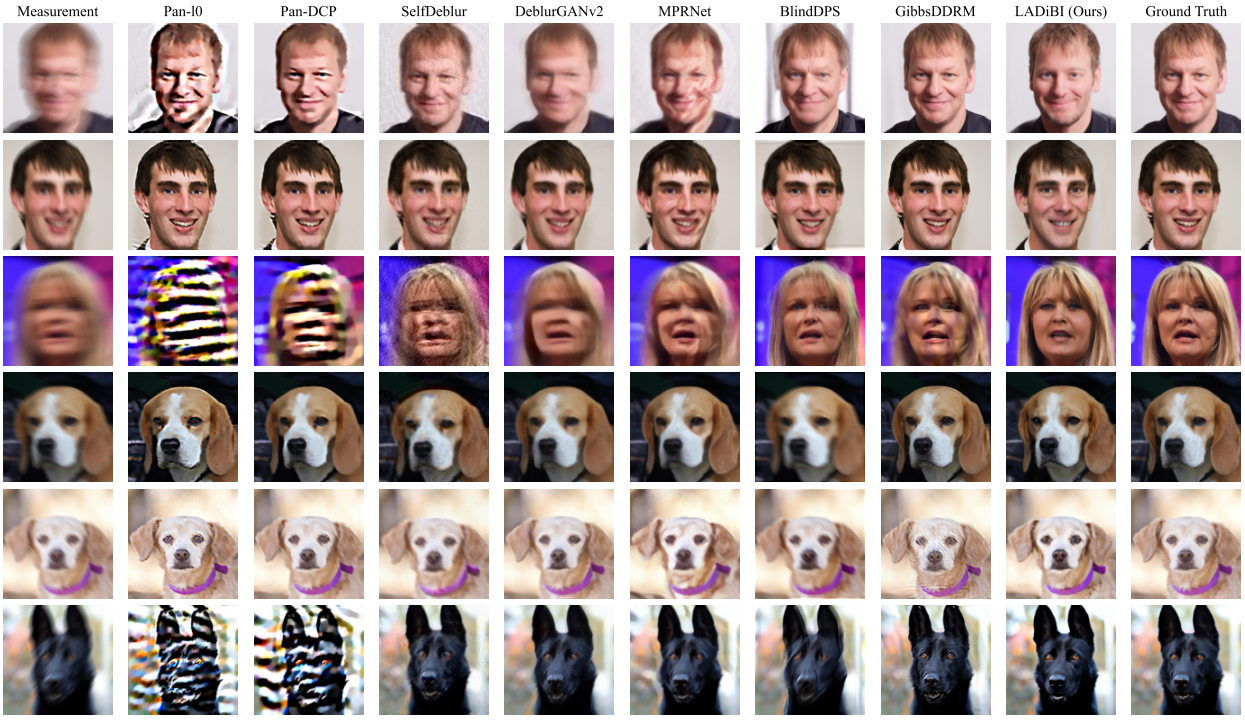

Figure 7. Additional qualitative results on motion deblurring task.

Figure 8. Additional qualitative results on Gaussian deblurring task.

U-shape curve, which results from the following fact about encoding the measurement image; if we make the encoding timestep too large, most of the information about the measurement has been replaced by noise which does not al-

Figure 9. Additional qualitative results on JPEG decompression task.

Figure 10. Additional qualitative results on painting restoration.

low the sampling process to leverage useful features of the degraded image. On the other hand, if the timestep is too small, the sampling process is not equipped with enough scheduled variance to reach the target prior distribution.

In addition, Figure 12 presents the effectiveness of taking advantage of the time-traveling strategy. More time-traveling boosts the overall performance up to a specific value, after which we begin to notice a trade-off between perceptual clarity of the image and fidelity to the target distribution.

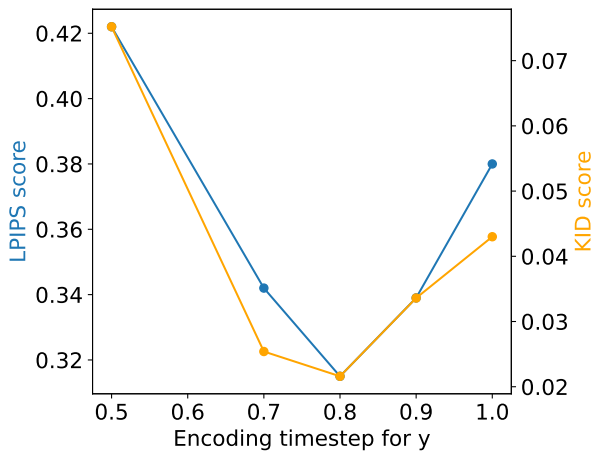

Figure 11. Ablation study on diffusion timestep for encoding the measurement.

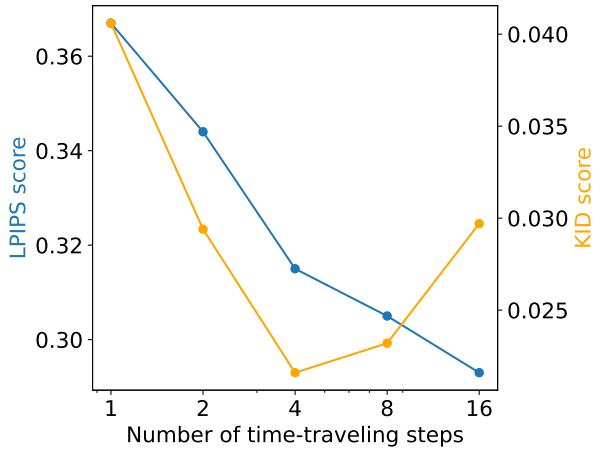

Figure 12. Ablation study on number of time-traveling steps.

With respect to ablation studies, we are interested in evaluating the performance of our algorithm when using a task-agnostic implementation for $\mathcal{A}_{\hat{\phi}}$. Hence, we tested the neural network architecture employed for non-linear tasks at linear inverse problems and we showcase qualitative results in Figure 13. We observe that, although performing worse than LADiBI employed with a linear operator architecture, the task-agnostic method is capable of producing estimates of decent quality while also preserving a design that allows applicability to any image restoration task.

D. Limitations and Future Works

In this section, we discuss the limitation of LADiBI and potential future works to address these problems.

Similar to many training-free diffusion posterior sampling algorithms [8, 9, 14, 27], our method is also sensitive to hyper-parameter tuning. We have provided details about our hyperparameter choices in the previous sections, and we will release our code in a public repository upon publication of this paper.

Figure 13. Comparison between linear and deep operator architecture on linear tasks.

Another notable drawback of our method is that LADiBI requires significantly longer inference time in comparison to the best performing baseline (i.e. GibbsDDRM [27]) in order to obtain high quality restorations. With the general operator initialization, our algorithm can take around 5 minutes to complete on a single NVIDIA A6000 GPU while GibbsDDRM only takes around 30 seconds. Although this is justifiable by the larger optimization space that we operate on, investigating on how to reduce the inference time requirement is an interesting and critical next step for our work.

Lastly we would like to note that, while neural networks, as demonstrated in the previous sections, can serve as a general model family for various operator functional classes and achieve satisfactory results, obtaining state-of-the-art performance for linear tasks within a reasonable inference time still requires resorting to a linear kernel as the estimated operator. Exploring neural network architectures that can easily generalize across different operator functional classes while achieving state-of-the-art results efficiently remains an exciting direction for future work.

E. Broader Impact Statement

Lastly, since our algorithm leverages large pre-trained image generative models, we would like to address the ethical concerns, societal impact as well as the potential harm that can be caused by our method when being used inappropriately.

As a consequence of using large pre-trained text-to-image generative models, our method also inherits potential risks associated with these pre-trained models, including the propagation of biases, copyright infringement and the possibility of generating harmful content. We recognize the significance of these ethical challenges and are dedicated to responsible AI research practices that prevent reinforcing these ethical considerations. Upon the release our code, we are committed to implement and actively update safeguards in our public repository to ensure safer and more ethical content generation practices.