Decoupling Training-Free Guided Diffusion by ADMM

Youyuan Zhang[1*] Zehua Liu[2*] Zenan Li[3*] Zhaoyu Li[4*] James J. Clark[1] Xujie Si[4,5]
[1]McGill University [2]The University of Hong Kong [3]Nanjing University
[4]University of Toronto [5]CIFAR AI Chair, MILA
youyuan.zhang@mail.mcgill.ca, liuzehua@connect.hku.hk, lizn@smail.nju.edu.cn
{zhaoyu,six}@cs.toronto.edu, james.clark1@mcgill.ca

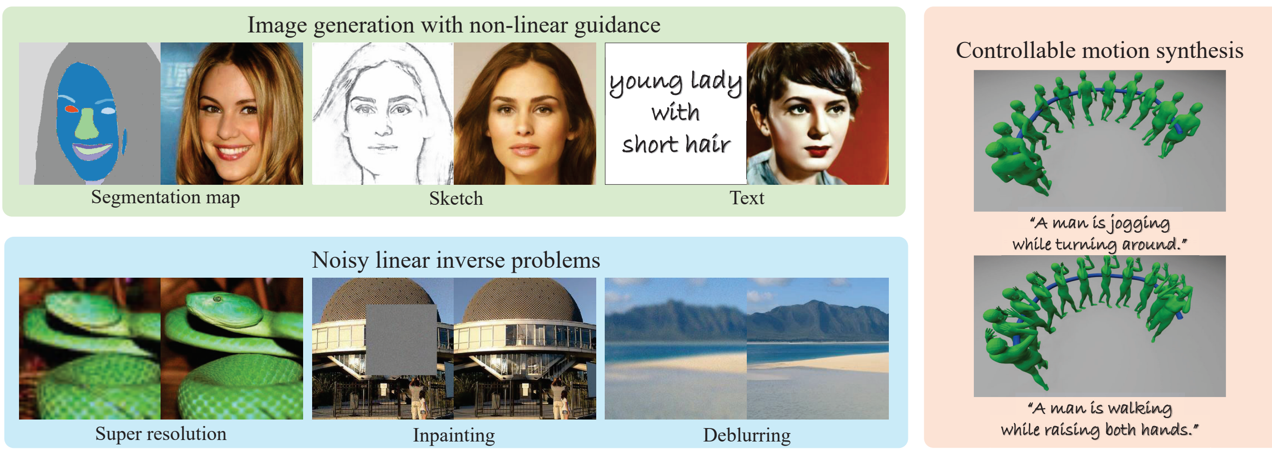


Figure 1. **Illustrated results of *ADMMDiff* on diverse conditional generation tasks.** *ADMMDiff* effectively guides the generation process of diffusion models using training-free guidance functions, producing high-quality samples that adhere closely to the specified conditions.

Abstract

In this paper, we consider the conditional generation problem by guiding off-the-shelf unconditional diffusion models with differentiable loss functions in a plug-and-play fashion. While previous research has primarily focused on balancing the unconditional diffusion model and the guided loss through a tuned weight hyperparameter, we propose a novel framework that distinctly decouples these two components. Specifically, we introduce two variables $\mathbf{x}$ *and* $\mathbf{z}$*, to represent the generated samples governed by the unconditional generation model and the guidance function, respectively. This decoupling reformulates conditional generation into two manageable subproblems, unified by the constraint* $\mathbf{x} = \mathbf{z}$*. Leveraging this setup, we develop a new algorithm based on the Alternating Direction Method of Multipliers (ADMM) to adaptively balance these components. Additionally, we establish the equivalence between the diffusion reverse step and the proximal operator of ADMM and provide a detailed convergence analysis of our algorithm under certain mild assumptions. Our experiments demonstrate that our proposed method* ADMMDiff *consistently generates high-quality samples while ensuring strong adherence to the conditioning criteria. It outperforms existing methods across a range of conditional generation tasks, including image generation with various guidance and controllable motion synthesis.*

1. Introduction

Diffusion models [22, 47, 50] have emerged as a potent paradigm in generative modeling, demonstrating versatility across various domains, including image synthesis [45, 46], 3D object generation [30, 40], natural language processing [29, 33], and time-series analysis [3, 42]. In many of these areas, there is an escalating interest in conditional generation, where the generative process is guided not only by the diffusion model but also by external information such as text prompts [1, 21, 27, 28, 41], segmentation maps [37, 59], sketches [37, 52, 59], etc. In this paper, we focus on solving this task by combining an unconditional diffusion model with a differentiable guidance function that evaluates condi-

*Authors contributed equally to this work.

tion satisfaction in a *plug-and-play* fashion [8, 18, 48], which generates the desired samples without additional training.

Specifically, our objective is to generate samples from the prior distribution $p(\mathbf{x})$ that satisfy conditions $\mathbf{y}$, approximating the posterior $p(\mathbf{x}|\mathbf{y})$. Using Bayes' rule, the posterior can be expressed as $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$, where $p(\mathbf{x})$ is modeled by a pretrained diffusion model and $p(\mathbf{y}|\mathbf{x})$ is approximated by a differentiable loss function. This formulation naturally divides the conditional generation task into two distinct components. However, effectively integrating these components poses significant challenges due to their different objectives. For example, the model may generate diverse but condition-violating samples or produce samples that satisfy the conditions but lack diversity or quality. Existing approaches [8, 18, 20, 48, 56, 58] mainly introduce a weight hyperparameter to strike a balance between the unconditional diffusion model and the guidance function. Nonetheless, determining an optimal parameter is non-trivial and highly dependent on the specific task, making it difficult to generalize across different problem settings.

To address these limitations, we propose a novel framework that fundamentally rethinks the interaction between the unconditional diffusion model and the guidance function in conditional generation. Specifically, we decouple these two components by representing the samples from the diffusion model as $\mathbf{x}$ and introducing an auxiliary variable $\mathbf{z}$ that represents samples refined by the guidance function. This formulation allows us to transform the conditional generation problem into two more tractable subproblems, connected by the constraint $\mathbf{x} = \mathbf{z}$. Within this setup, we introduce a dual variable and develop a new algorithm based on the Alternating Direction Method of Multipliers (ADMM) [16, 17] for conditional generation. Our algorithm sidesteps the use of the weight hyperparameter and allows for a more natural and adaptive balancing between the diffusion model and the guidance function. Additionally, we theoretically demonstrate that the proximal operator of the diffusion term $-\log p(\mathbf{x})$ in the ADMM can be effectively approximated by the diffusion reverse process and further provide a rigorous proof of the convergence of our algorithm under generic assumptions regarding the distributions $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{z})$.

We evaluate our proposed framework on various conditional generation tasks in different domains (see Figure 1 for illustration). In image synthesis tasks, our method can accept both non-linear semantic parsing models and linear measurement functions as guidance conditions, and achieves superior performance in terms of both image quality and condition satisfaction. In addition, our method is capable of generalizing in motion domain and guide text-condition motion diffusion models to follow specific trajectories.

2. Preliminary

2.1. Diffusion Model

Given any datapoint $\mathbf{x}_0 \sim p(\mathbf{x})$, the *forward diffusion process* adds small amount of Gaussian noise to the sample by T steps [22], producing a sequence of noisy samples $\mathbf{x}_t, t = 1, \dots, T$:

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\{\beta_t \in (0,1)\}_{t=1}^T$ is a pre-defined variance schedule to control the scales of Gaussian noise. By introducing auxiliary variables $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$ for $t = 1, 2, \dots, T$, we have

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

For the distribution $p_t(\mathbf{x}_t)$ induced by adding noise to $p(\mathbf{x})$, the score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ can be approximated by training a time-conditioned neural network $s_{\boldsymbol{\theta}}(\mathbf{x}_t, t) \colon \mathcal{X} \times [0, T] \to \mathcal{X}$ that tries to denoise the noisy sample $\mathbf{x}_t$ using the denoising score matching objective [50]:

$$L_t(\theta) = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_t}\left[\|s_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \mathbf{x}_0\|_2^2\right],$$

With sufficient capacity of a denoising model s_θ, the *reverse diffusion process* is obtained using Tweedie's formula [14]:

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) := \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t + \beta_t s_\theta(\mathbf{x}_t, t))$$

To approximately obtain samples, the Langevin dynamic is often introduced [5, 49], forming the following update rule:

$$\tilde{\mathbf{x}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t + \beta_t s_\theta(\mathbf{x}_t, t)) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}\boldsymbol{\epsilon}.$$

2.2. Proximal Operator and ADMM

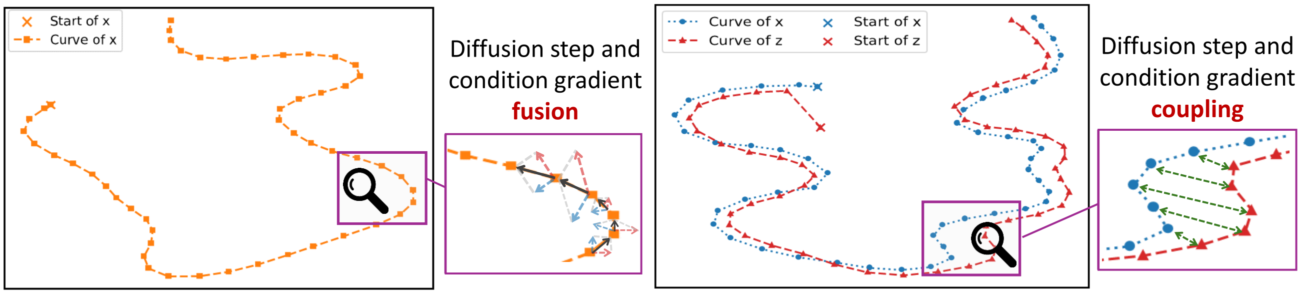
For any l-weakly convex function g (i.e., the function satisfying that $g(\mathbf{x}) + (l/2)\|\mathbf{x}\|^2$ is convex), its proximal operator with $\lambda \in (0, 1/l)$ is defined by [36, 39]:

$$\operatorname{prox}_{\lambda g}(\mathbf{x}) := \arg\min_{\mathbf{x}'} \left\{g(\mathbf{x}') + \frac{1}{2\lambda}\|\mathbf{x}' - \mathbf{x}\|^2\right\}.$$

Note that the mapping $\mathbf{x}' \to g(\mathbf{x}') + (1/2\lambda)\|\mathbf{x}' - \mathbf{x}\|^2$ is strongly convex for any $\mathbf{x}'$, and thus the proximal operator is $\frac{1}{1-\lambda\ell}$-Lipschitz [44, Proposition 12.19].

Alternating Direction Method of Multipliers (ADMM) is a classic method [16, 17], which efficiently addresses the following constraint minimization problem.

$$\begin{aligned} \min_{\mathbf{x}, \mathbf{z}} \quad & f(\mathbf{x}) + g(\mathbf{z}), \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{z}. \end{aligned}$$



(a) Geometry of single trajectory denoising methods.
(b) Geometry of ADMM-based method for decoupling guidance gradient.

Figure 2. **Geometrical illustration of ADMM-based method.** Compared with classic guided diffusion frameworks which directly perturbs reverse diffusion steps with guidance gradients, ADMM-based method decouples the guidance gradient from the reverse diffusion trajectory and allows more flexibility to explore guidance conditions.

Furthermore, the corresponding augmented Lagrangian function is defined as

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}; \boldsymbol{\mu}) := f(\mathbf{x}) + g(\mathbf{z}) + \langle \boldsymbol{\nu}, \mathbf{x} - \mathbf{z} \rangle + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|^2.$$

Based on the augmented Lagrangian function and plugging in the proximal operator, the Gauss-Seidel update of ADMM at the t-th iteration can be formulated by

$$\begin{cases} \mathbf{x}_{t+1} = \mathrm{prox}_{\rho f}(\mathbf{z}_t - \frac{1}{\rho}\boldsymbol{\nu}_t); \\ \mathbf{z}_{t+1} = \mathrm{prox}_{\rho g}(\mathbf{x}_{t+1} + \frac{1}{\rho}\boldsymbol{\nu}_t); \\ \boldsymbol{\nu}_{t+1} = \boldsymbol{\nu}_k + \rho(\mathbf{x}_{t+1} - \mathbf{z}_{t+1}). \end{cases}$$

ADMM not only guarantees the convergence, but also achieves a satisfactory convergent rate when subproblems involving the proximal operator can be solved efficiently [23, 32, 34]. Moreover, as a primal-dual method, ADMM possesses a nice property that can dynamically control the equilibrium between the objective function and the constraint satisfaction in the optimization process, by balancing the primal and dual residuals [4, 15].

3. Conditional Generation by ADMM

We frame the conditional generation problem as the maximum likelihood problem:

$$\max_{\mathbf{x}} \log p(\mathbf{x}|\mathbf{y}) = \log p(\mathbf{x}) + \log p(\mathbf{y}|\mathbf{x}) - \log p(\mathbf{y}),$$

where $\log p(\mathbf{y})$ is constant for a specific condition $\mathbf{y}$, $p(\mathbf{x})$ and $p(\mathbf{y}|\mathbf{x})$ are approximated by an off-the-shelf diffusion model (denoted by $q_{\boldsymbol{\phi}}(\mathbf{x})$) and a pre-defined condition guidance function (denoted by $c_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})$), respectively. Putting these together, the conditional generation problem can be rewritten as the following variational version:

$$\max_{\mathbf{x}} \log q_{\boldsymbol{\phi}}(\mathbf{x}) + \log c_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y}).$$

Furthermore, by *decoupling* the two objectives with an auxiliary variable $\mathbf{z}$, we can reformulate the problem as

$$\max_{\mathbf{x},\mathbf{z}} \log q_{\boldsymbol{\phi}}(\mathbf{x}) + \log c_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{y}), \quad \text{s.t.} \quad \mathbf{x} = \mathbf{z}. \tag{P}$$

From this formulation, the sample generating and condition satisfying are assigned to two different variables $\mathbf{x}$ and $\mathbf{z}$, and they are connected by the constraint $\mathbf{x} = \mathbf{z}$. In this sense, the conditional generation is converted into a *coupling* problem, i.e., fusing $\mathbf{x}$ and $\mathbf{z}$ in the generative process [10, 31], such that $\mathbf{x}$ and $\mathbf{z}$ maximizes $\log q_{\boldsymbol{\phi}}(\mathbf{x})$ and $\log c_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{y})$ under the equality constraint $\mathbf{x} = \mathbf{z}$.

Rewriting the maximization problem to the minimization, the corresponding augmented Lagrangian function is

$$\mathcal{L}_\rho(\mathbf{x}, \mathbf{z}; \boldsymbol{\nu}) := -\log q_{\boldsymbol{\phi}}(\mathbf{x}) - \log c_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{y}) + \langle \boldsymbol{\nu}, \mathbf{x} - \mathbf{z} \rangle + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|^2.$$

Based on the augmented Lagrangian form, we encode the constraint as a penalty term, paired by the Lagrange multiplier (i.e., the dual variable $\boldsymbol{\nu}$), as well as a coefficient $\rho > 0$, to control the intensity of the constraint enforcement. Next, by applying the ADMM update, we have[1]

$$\begin{cases} \mathbf{x}_t = \mathrm{prox}_{(-\rho \log q_{\boldsymbol{\phi}})}(\mathbf{z}_{t+1} - \frac{1}{\rho}\boldsymbol{\nu}_t); \\ \mathbf{z}_t = \mathrm{prox}_{(-\rho \log c_{\boldsymbol{\theta}})}(\mathbf{x}_t + \frac{1}{\rho}\boldsymbol{\nu}_t); \\ \boldsymbol{\nu}_t = \boldsymbol{\nu}_{t+1} + \rho(\mathbf{x}_t - \mathbf{z}_t); \end{cases}$$

In particular, the dual variable $\boldsymbol{\nu}$ and the coefficient ρ are *dynamically* adapted (by conducting gradient descent in ADMM) based on the satisfaction degree of constraint $\mathbf{x} = \mathbf{z}$, which essentially controls the progress of coupling.

Compared to existing training-free guided diffusion approaches, ADMM provides a more flexible paradigm by

[1]To keep the consistency with the index order of reverse diffusion process, the iteration index t is also defined to range from T to 0.

decoupling the guidance gradient from the reverse diffusion trajectory. Figure 2 demonstrates an intuitive illustration of existing methods and our approach. Existing approaches directly modify reverse diffusion steps with guidance gradients, while the ADMM-based method gradually couples two trajectories of reverse diffusion and condition satisfaction.

However, the computations of the proximal operators in ADMM are still intractable due to the non-convex and non-analytic functions $q_{\boldsymbol{\phi}}(\mathbf{x})$ and $c_{\boldsymbol{\theta}}(\mathbf{z}, \mathbf{y})$. To overcome this issue, we propose using the inexact version instead, in which the proximal operator of $\mathbf{x}$ can be approximated by a single reverse step of the diffusion model, and the proximal operator of $\mathbf{z}$ can be approximated by a few steps of gradient descent with respect to the condition guidance function. To achieve an approximation of the proximal operator of $-\log q_{\boldsymbol{\phi}}$, we have the following proposition.

Proposition 1. *Let* $\mathbf{x}_t = \sqrt{\bar{\alpha}}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, *be a noisy point generated from the datapoint* $\mathbf{x}_0 \sim p(\mathbf{x})$ *by the diffusion forward process. Then, the point* $\tilde{\mathbf{x}}_{t-1}$ *derived from the diffusion reverse step at* $\mathbf{x}_t$*, i.e.,*

$$\tilde{\mathbf{x}}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t + \beta_t s_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\right) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\beta_t \boldsymbol{\epsilon},$$

is a first-order approximation to the proximal operator of negative log-likelihood $-\frac{1}{\rho}\log q_{\boldsymbol{\phi}}(\mathbf{x})$ *at* $\mathbf{x}_t$ *with* $\rho = \frac{\beta}{1-\beta}$.

Remarks. The proof of this proposition aligns with the theoretical analysis in [60], which provides a strong foundation for interpreting reverse diffusion as a proximal operator.

According to Proposition 1 illustrating the connection between the proximal operator and diffusion reverse process, we can directly approximate the update of $\mathbf{x}$ using the off-the-shelf diffusion model. However, the update of $\mathbf{z}$ is still challenging, because the condition guidance function is generally implemented by an off-the-shelf classifier, which may not be explicitly trained on samples along the forward diffusion trajectory but rather on the original datapoint $\mathbf{x}_0$. Therefore, we first use Tweedie's formula as an estimation of $\mathbf{z}_0$:

$$\tilde{\mathbf{z}}_0(\mathbf{z}_t) = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{z}_t + (1-\bar{\alpha}_t)\boldsymbol{s}_t\right),$$

Then, we compute the loss function $c_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}_0, \mathbf{y})$ and apply the gradient descent with respect to $\mathbf{z}_t$ as the first-order approximation of the proximal operator:

$$\mathbf{z}_{t-1} = \mathbf{z}_t - \eta\rho(\mathbf{z}_t - \mathbf{x}_t - \frac{1}{\rho}\boldsymbol{\nu}_t) + \eta\nabla_{\mathbf{z}_t}\log c_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}_0(\mathbf{z}_t), \mathbf{y}).$$

4. Algorithm and Theoretical analysis

By integrating the update steps for both $\mathbf{x}$ and $\mathbf{z}$, we derive the complete ADMM-based conditional generation algorithm, as outlined in Algorithm 1. To further minimize the approximation error of the proximal operator, we extend the gradient descent process to K_t iterations at each step, ensuring more accurate optimization and improved convergence.

Algorithm 1 Training-Free Guided Diffusion by ADMM

Input: pretrained diffusion model $s_\theta(\mathbf{x}_t, t)$ and guidance function $c_\phi(\mathbf{x}, \mathbf{y})$; initial points $\mathbf{x}_T, \mathbf{z}_T$ sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\nu}_T = \mathbf{0}$, step size $\eta > 0$.

for $t = T, T-1, \ldots, 0,$ **do**

Inexact update of $\mathbf{x}$ *by reverse diffusion*

$\hat{\mathbf{x}}_{t-1} = \mathbf{z}_t - \frac{1}{\rho}\boldsymbol{\nu}_t$;

$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_t + \beta_t s_\theta(\hat{\mathbf{x}}_{t-1}, t)\right) + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}}\beta_t\boldsymbol{\epsilon}$.

Inexact update of $\mathbf{z}$ *by gradient descent*

for $k = 0, 1, \ldots, K_t$ **do**

$\mathbf{z}_t^{(k+1)} = \mathbf{z}_t^{(k)} - \eta\rho(\mathbf{z}_t^{(k)} - \mathbf{x}_{t-1} - \boldsymbol{\nu}_t)$
$\quad + \eta\nabla_{\mathbf{z}}\log c_\theta(\tilde{\mathbf{z}}_0(\mathbf{z}^{(k)}), \mathbf{y})$;

end for

$\mathbf{z}_{t-1} = \mathbf{z}_t^{(K_t+1)}$;

Update of the dual variable $\boldsymbol{\nu}$

$\boldsymbol{\nu}_{t-1} = \boldsymbol{\nu}_t + \rho(\mathbf{x}_{t-1} - \mathbf{z}_{t-1})$;

end for

Compared to the standard ADMM method, Algorithm 1 offers greater flexibility via solving the minimization subproblem inexactly. However, analyzing the convergence of this ADMM-based algorithm encounters a significant challenge due to that the non-convex nature of $\log q_{\boldsymbol{\phi}}(\mathbf{x})$ and $\log c_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{z})$ hinders the usage of the standard variational inequality framework for the ADMM in this context [4]. In this paper, we facilitate that the subproblems become convex programming problems when the hyperparameter ρ is sufficiently small. Based on this observation, we can prove the convergence of our proposed algorithm under some mild assumptions on the objective functions.

Sufficient decreasing property. Proposition 1 states that the proximal operator of negative log-likehood $-\rho\log q_{\boldsymbol{\phi}}(\mathbf{x})$ can be well approximated by using the diffusion reverse process. Then, with a proper assumption, we can derive the sufficient decreasing property in the following.

Theorem 1. *Assume that (1) there exists a constant* $\delta > 0$*, such that* $|\log p(\mathbf{x}) - \log q_{\boldsymbol{\phi}}(\mathbf{x})| \leq \frac{\delta}{2}$ *holds for any* $\mathbf{x}$*; and (2)* $-\log q_{\boldsymbol{\phi}}(\mathbf{x})$ *is L-smooth. Then, there exists* $\delta_t > 0$*, such that*

$$-\log p(\mathbf{x}_{t-1}) + \frac{1}{2\rho}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|^2 \leq \min_{\mathbf{x}} \left\{-\log p(\mathbf{x}) + \frac{1}{2\rho}\|\mathbf{x} - \mathbf{x}_t\|^2\right\} + \delta_t.$$

Remarks. The detailed formation of δ_t is provided in Appendix A. Theorem 1 essentially states that the gap between

Method	Segmentation map Distance ↓	Segmentation map FID ↓	Sketch Distance ↓	Sketch FID ↓	Text Distance ↓	Text FID ↓
DPS [8]	2199.8	57.38	50.74	67.21	10.46	57.13
LGD-MC [48]	2073.1	46.10	34.33	65.99	10.72	44.04
FreeDoM [58]	1696.1	53.08	33.29	70.97	10.83	55.91
MPGD [20]	1922.5	43.97	35.32	60.56	10.70	43.98
ADMMDiff	**1586.2**	**30.18**	**32.28**	**42.43**	**10.08**	**43.84**

Table 1. Comparison of non-linear guided image synthesis on CelebA-HQ dataset with different guidance conditions. **Bold** indicates the best. We use the ℓ_2 distance between parsing models' outputs from generated samples and reference images when testing on segmentation and sketch guidance, and use the ℓ_2 distance between text and image embeddings when testing on text guidance. We compute the FID score from the statistics of the all the generated images and reference image set. Results illustrate that our method outperforms all the baselines in term of both image quality and guidance satisfaction.

Methods	DPS [8]	LGD-MC [48]	FreeDoM [58]	MPGD [20]	**ADMMDiff**
CLIP Score↑	24.5	24.3	25.9	25.1	**26.85**

Table 2. Comparison of textual alignment on unconditional CelebA-HQ using text guidance. **Bold** indicates the best. The CLIP score is computed using CLIP-L/14. Results illustrate that our method achieves the best performance.

the two proximal operators of $-\log q_\phi(\mathbf{x})$ and $-\log p(\mathbf{x})$ can be properly bounded if the ground-truth distribution $p(\mathbf{x})$ can be well estimated by the variational distribution $q_\phi(\mathbf{x})$ based on the diffusion model. This result paves the way for the sequel convergence analysis of Algorithm 1.

Convergence analysis. Now we are ready to establish the convergence analysis of Algorithm 1. By combining the result of Theorem 1 and some mild assumptions of smoothness, we can derive the theoretical results as follows.

Theorem 2. *Assume that $\log q_\phi(\mathbf{x})$ and $\log c_{\boldsymbol{\theta}}(\mathbf{z};\mathbf{y})$ are both L-smooth, $\sum_t \delta_t < +\infty$, and $\rho \leq \frac{1}{6L}$. Let the sequence generated by Algorithm 1 be $\{(\mathbf{x}_t, \mathbf{z}_t, \boldsymbol{\nu}_t)\}_{t=T}^{0}$. If $\eta \leq \frac{1}{\rho+L}$ and $\sum_t 2^{-K_t} < +\infty$, then $(\mathbf{x}_0, \mathbf{z}_0)$ converges to the stationary point of problem (P)*

$$\lim_{T\to\infty} \|\mathbf{z}_1 - \mathbf{z}_0\| = 0, \quad \lim_{T\to\infty} \|\mathbf{x}_1 - \mathbf{x}_0\| = 0,$$

with a convergent rate

$$\min_{j\in\{T,\dots,0\}} \left\{\|\mathbf{z}_{j+1} - \mathbf{z}_j\|^2 + \|\mathbf{x}_{j+1} - \mathbf{x}_j\|^2\right\} = o\left(\frac{1}{T}\right).$$

Remarks. A more detailed version and the corresponding proof are provided in Appendix A. The theorem shows that Algorithm 1 can successfully converges to the stationary point with a sublinear rate. The basic idea to prove it is to construct a suitable Lyapunov function which metricizes the primal residual and the dual residual.

5. Experiments

We evaluate the efficacy of our method on three tasks using different diffusion models and guidance conditions. Namely, we evaluate on image synthesis tasks using both non-linear and linear guidance functions and controllable motion generation tasks using trajectory guidance.

5.1. Image Generation with Non-linear Guidance

We first apply our method to guide unconditional image diffusion models pretrained on CelebA-HQ human face dataset [35]. We employ three different guidance conditions: segmentation map, sketch, and text. The experimental settings are consistent with [58]. For the segmentation map guidance, we use BiSeNet [57] to compute the face parsing maps. For the sketch guidance, we use pretrained sketch generator from AODA [54]. For textual guidance, we leverage CLIP-B/16 text and image encoders to get the embeddings.

We compare the proposed method with four baseline methods, i.e., Diffusion Posterior Sampling (DPS) [8], Loss-Guided Diffusion with Monte Carlo (LGD-MC) [48], Training-Free Energy-Guided Diffusion Models (FreeDoM) [58], and Manifold Preserving Guided Diffusion (MPGD) [20]. For the LGD-MC, we set the size of Monte-Carlo sampling by $n = 10$. We adopt Algorithm 1 and 3 of [20] and apply pixel diffusion models with VQGAN's autoencoder when evaluating MPGD.

For each condition type from segmentation map, sketch, and text prompt, we select 1000 conditions for evaluation. Table 1 compares ℓ_2 distances and FID scores of our method with the baselines. When evaluating the guidance of text prompts, we additionally compute the CLIP scores in Table 2 as a more generic metric to evaluate text-image alignment. The results demonstrate that *ADMMDiff* achieves state-of-the-art performance on non-linear guided image generation tasks in terms of both image quality and guidance following.

5.2. Linear Inverse Problems

We next evaluate our method on linear inverse problems, instantiated by super-resolution, inpainting and deblurring tasks. The experimental settings follow [8]. We select

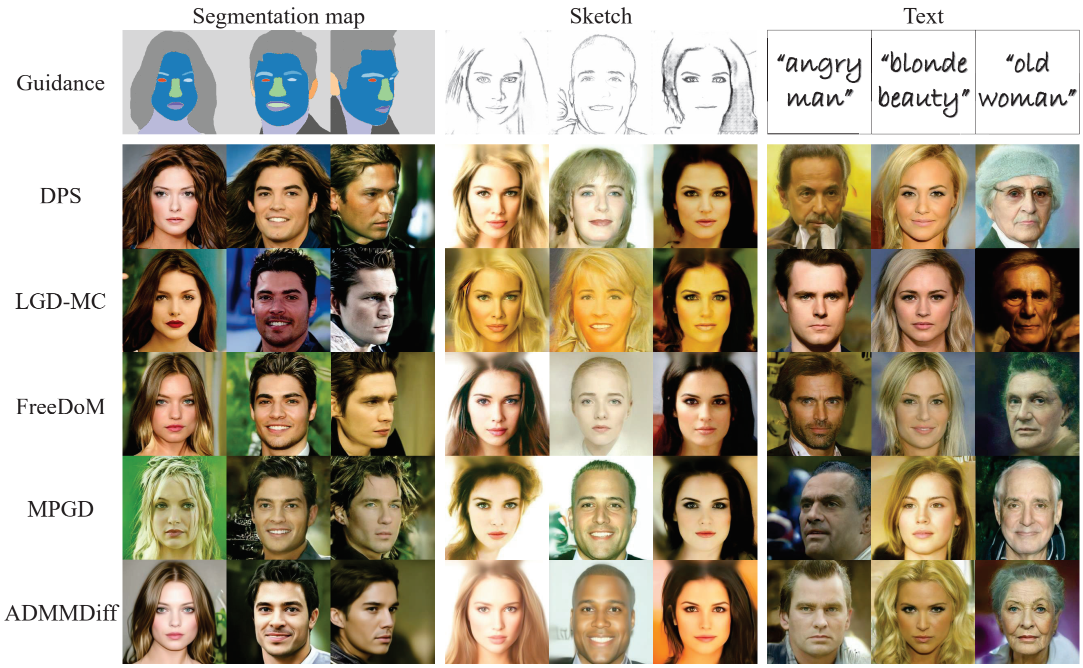


Figure 3. Qualitative comparison on CelebA-HQ in three conditional image synthesis tasks: (a) segmentation maps to human faces; (b) sketches to human faces; (c) text prompts to human faces. Our method offers comparable image quality and advantage in the degree of satisfaction of the conditions.

	SR ($\times 4$)		**Inpaint (box)**		**Inpaint (random)**		**Deblur (gauss)**		**Deblur (motion)**	
Method	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓	FID ↓	LPIPS ↓
					FFHQ DATASET					
ADMM-TV	110.6	0.428	68.94	0.322	181.5	0.463	186.7	0.507	152.3	0.508
Score-SDE [50]	96.72	0.563	60.06	0.331	76.54	0.612	109.0	0.403	292.2	0.657
PnP-ADMM [7]	66.52	0.353	151.9	0.406	123.6	0.692	90.42	0.441	89.08	0.405
MCG [9]	87.64	0.520	40.11	0.309	29.26	0.286	101.2	0.340	310.5	0.702
DDRM [26]	62.15	0.294	42.93	0.204	69.71	0.587	74.92	0.332	-	-
DPS [8]	39.35	0.214	33.12	0.168	**21.19**	0.212	44.05	0.257	39.92	0.242
ADMMDiff	**26.8**	**0.203**	**24.97**	**0.160**	22.04	**0.111**	**27.98**	**0.251**	**25.11**	**0.239**
					IMAGENET DATASET					
ADMM-TV	130.9	0.523	87.69	0.319	189.3	0.510	155.7	0.588	138.8	0.525
Score-SDE [50]	170.7	0.701	54.07	0.354	127.1	0.659	120.3	0.667	98.25	0.591
PnP-ADMM [7]	97.27	0.433	78.24	0.367	114.7	0.677	100.6	0.519	89.76	0.483
MCG [9]	144.5	0.637	39.74	0.330	39.19	0.414	95.04	0.550	186.9	0.758
DDRM [26]	59.57	0.339	45.95	**0.245**	114.9	0.665	63.02	0.427	-	-
DPS [8]	50.66	0.337	38.82	0.262	35.87	0.303	62.72	0.444	56.08	0.389
ADMMDiff	**49.97**	**0.331**	**37.47**	0.262	**34.80**	**0.207**	**51.3**	**0.414**	**55.99**	**0.364**

Table 3. Quantitative evaluation (FID, LPIPS) of solving linear inverse problems on FFHQ 256×256-1k validation dataset and ImageNet 256×256-1k validation dataset. **Bold** indicates the best. Underline indicates the second best.

the FFHQ validation set [24] and the ImageNet validation set [11] as two benchmark datasets. We adopt two unconditional diffusion models pre-trained from [8, 12] for these two datasets, respectively. For linear inverse problems which have a measurement function of the form $\mathbf{y} = \boldsymbol{A}\mathbf{x} + \epsilon$, we compute the loss as $\ell_2 = \|\boldsymbol{A}\hat{\mathbf{x}}_0 - \mathbf{y}\|_2^2$, where $\hat{\mathbf{x}}_0$ represents the estimated clean image at each reverse step and $\mathbf{y}$ is the provided noisy measurement depending on the task. We compare our method with various baselines including diffusion posterior sampling (DPS) [8], denoising diffusion restoration models (DDRM) [26], manifold constrained gradients (MCG) [9], plug-and-play alternating

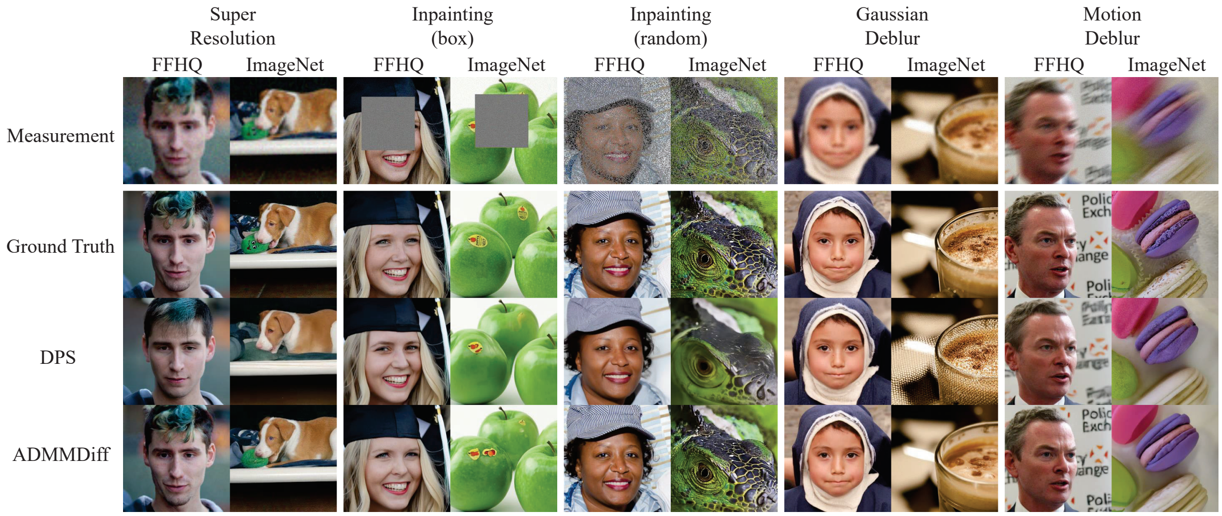


Figure 4. Qualitative ablation of protected concept distillation, targeting "ship" as the concept to be forgotten. The results showcase the model's outputs on the target concept ("ship") and unrelated concepts, highlighting the effectiveness of the forgetting mechanism

Method	“walking” Obj. ↓	Emb. ↓	“raising hands” Obj. ↓	Emb. ↓	“jogging around” Obj. ↓	Emb. ↓	“backwards” Obj. ↓	Emb. ↓
Unconditional [51]	7.139	1.48	3.367	7.80	8.357	5.17	3.344	4.96
GMD w/o inpainting [25]	0.095	4.62	0.297	7.93	0.104	5.54	0.223	8.94
LGD-MC [48]	0.107	4.67	0.272	8.13	0.109	5.45	0.237	8.83
ADMMDiff	**0.042**	**3.38**	**0.064**	**7.28**	**0.057**	**5.23**	**0.050**	**8.30**

Table 4. Comparison on controllable motion generation task using trajectory guidance. Our method achieves the best guidance performance with lowest objective value. In terms of textual alignment, the embedding distance of our method is the closest to the unconditional model and outperforms the two guided motion diffusion baselines. **Bold** indicates the best.

direction method of multipliers (PnP-ADMM) [7], total-variation sparsity regularized optimization method (ADMM-TV) and Score-SDE [50].

On both FFHQ and ImageNet validation set, we test on 1000 images using perceptual metrics including Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) distance, and standard distortion metrics including peak-signal-to-noise-ratio (PSNR) and structural similarity index (SSIM). In Table 3, we compare our method with baselines on FID and LPIPS. Among the 10 tasks across the two dataset. The results show that our methods achieve state-of-the-art performances on FID score for 9 out of 10 tasks, LPIPS for 9 out of 10 tasks. We also report PSNR and SSIM of our method and baselines in Appendix C. Though our method is not specially designed to solve linear inverse problems, it still outperforms most of the baselines on different tasks. In Figure 4, we further provide visual comparison of our method against single trajectory method DPS. It can be observed that, compared with DPS which may output blurry images, our proposed method stably gives results with superior image quality and fidelity to original images.

5.3. Guided Motion Generation

In this section, we extend our algorithm to controllable human motion generation task. We use Guided Motion Diffusion (GMD) [51] without guidance as the base motion diffusion model. To control the general behavior of human motion, we employ the text prompt input as prior and test on fine-grained trajectory guidance. In specific, we manually configure the temporal motion trajectory and guide the motion generation to follow the trajectory. The loss of the trajectory guidance is defined by the sum of the ℓ_2 distances between $\text{root}(\hat{\mathbf{x}}_0^i)$ and y^i. For a given time step t, the loss is defined as $\ell_2 = \sum_{i<N} \|\mathbf{y}^i - \text{root}(\hat{\mathbf{x}}_0^i)\|_2^2$ where $\mathbf{y}^i$ is the i-th position in the trajectory projected on the ground plane and $\text{root}(\hat{\mathbf{x}}_0^i)$ is the root position of the i-th motion. The overall objective is to align the semantic of the text prompt while minimizing the loss function to follow the configured trajectory. Following [48], we evaluate the performance using the *Objective* and *Embedding* metrics. The *Objective* metric measures the loss function ℓ_2 defined above. The *Embedding* metric measures the ℓ_2 distance between the motion embedding and the text embedding from the pretrained text-motion

encoders of T2M [19]. The validity of using the embedding distance to reflect text-motion alignment is demonstrated in Appendix B.3.1 in [48].

To evaluate the metrics, we choose four prompts: (1) "a person is walking"; (2) "a person is walking while raising both hands"; (3) "a person is jogging while turning around"; (4) "a person is walking backwards". For each text prompt, we sample 10 trajectories as path conditions using the trajectory diffusion model in GMD [25] and calculate the average value of each metric. We compare our method with GMD [25] and Loss-Guided Diffusion with Monte Carlo (LGD-MC) [48]. For LGD-MC, we set the size of Monte-Carlo sampling by $n = 10$. For GMD with guidance, we remove the trajectory inpainting post-processing for fair comparison. The results in table 4 show that our method achieves both the lowest embedding and objective distance compared with two guided motion generation baselines. Figure 5 compares the visual results of our method with GMD, where our method is shown to better follow the designed motion trajectory. While previous work [48] has already shown that baseline methods can do simple path following like straight lines, our visualization proves the superiority of our method in following more complicated trajectories. The comparison demonstrates that our method can satisfy the guidance when it is significantly different from the motion diffusion prior and meanwhile preserve the high-level semantics entailed by the motion diffusion prior.

6. Related Works

6.1. Training-Free Guided Diffusion

Diffusion-based training-free guidance methods [2, 8, 9, 12, 18, 25, 26, 48, 50, 58] aims to leverage unconditional diffusion models for conditional generation. A special case of the guidance tasks is linear inverse problems, where the guidance loss is derived from a linear forward model. A serious of methods [7–9, 13, 50], are specially designed to solve such linear guidance. Other guidance tasks use general differential loss functions as guidance. Methods target at general loss guidance [2, 8, 12, 20, 48, 50, 58], aims to generate samples that satisfy both the prior diffusion of unconditional diffusion models and the guidance alignment measured by the loss of guidance functions.

Guiding unconditional diffusion models typically requires the estimation of $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$ at each denoising step. Early works directly measure the gradient by training time-dependent classifiers or loss functions. Such classifiers or loss functions are task-dependent and hard to acquire. To improve the flexibility and leverage off-the-shelf guidance functions pretrained on clean samples, DPnP [18] proposes to leverage the score-matching as a plug-and-play objective to achieve the alignment with diffusion prior while minimizing the loss function for guidance satisfication. DPS [8] measures the intractable term $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y}|\mathbf{x}_t)$ through an intermediate estimation of $\hat{x}_0$ at each denoising step. A few works follow the sampling-time guidance methods introduced by DPS. Among them, LGD-MC [50] improves the point estimation of $\hat{x}_0$ with Monte Carlo. UGD [2] and FreeDoM [58] use time-travel strategy to solve the poor guidance problem in large domain. MPGD [20] focuses on the issue of manifold preservation and leverages pretrained autoencoders or latent space to make generated samples lie on the clean data manifold. Recent research [56] also shows that most existing approaches aim to guide a single reverse diffusion trajectory can be unified by a framework with different choices of hyperparameters. Compared with them, our method evolves two trajectories with constraints to balance the objectives of prior and guidance satisfication.

6.2. ADMM for Image-Based Tasks

Several works [6, 7, 53, 55] have explored the application of ADMM in image-based tasks, leveraging its ability to handle complex optimization problems efficiently. These methods primarily build upon the PnP-ADMM [7], which introduces a flexible framework where off-the-shelf image denoising algorithms are used as proximal operators within ADMM for image restoration tasks. In contrast, our approach is the first to incorporate diffusion models as the proximal operator for $\mathbf{x}_t$, taking advantage of their generative capabilities. Furthermore, we utilize Tweedie's formula to estimate $\mathbf{z}_0$, enabling our method to handle various conditional generation tasks beyond traditional restoration problems. This combination not only extends the applicability of ADMM to generative modeling but also ensures high-quality outputs by effectively balancing data fidelity and conditional alignment. Additionally, we provide a rigorous convergence analysis for the proposed method, offering new theoretical insights into its performance. This analysis highlights the robustness and efficiency of our approach, establishing a solid foundation for its application in diverse image-based generation tasks.

7. Conclusion and Future Work

In this paper, we propose a novel framework for training-free guided diffusion that decouples the unconditional diffusion model from the guidance function, introducing a new ADMM-based algorithm to effectively solve and couple the resulting subproblems. Our method, *ADMMDiff*, demonstrates a strong theoretical foundation with convergence guarantees and consistently outperforms existing approaches across various tasks, including guided image generation and motion synthesis. In future work, we plan to explore techniques to further accelerate the sampling process, enabling our method more effective and efficient for real-time applications. Additionally, we aim to incorporate more guidance functions to handle more complex and diverse conditional generation tasks.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 1

[2] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023. 8

[3] Marin Biloš, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, and Stephan Günnemann. Modeling temporal data as continuous functions with process diffusion. *arXiv preprint arXiv:2211.02590*, 2022. 1

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 3, 4

[5] Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 75 (5):056707, 2007. 2

[6] Stanley H Chan. Performance analysis of plug-and-play admm: A graph signal processing perspective. *IEEE Transactions on Computational Imaging*, 5(2): 274–286, 2019. 8

[7] Stanley H Chan, Xiran Wang, and Omar A Elgendy. Plug-and-play admm for image restoration: Fixed-point convergence and applications. *IEEE Transactions on Computational Imaging*, 3(1):84–98, 2016. 6, 7, 8, 16

[8] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 2, 5, 6, 8, 16

[9] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022. 6, 8, 16

[10] Frank Den Hollander. Probability theory: The coupling method. *Lecture notes available online (http://websites. math. leidenuniv. nl/probability/lecturenotes/CouplingLectures. pdf)*, 2012. 3

[11] 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*, pages 248–255. Ieee, 2009. 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 6, 8

[13] Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024. 8

[14] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106 (496):1602–1614, 2011. 2

[15] Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010. 3

[16] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & mathematics with applications*, 2(1):17–40, 1976. 2

[17] Roland Glowinski and Americo Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975. 2

[18] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022. 2, 8

[19] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022. 8

[20] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. *arXiv preprint arXiv:2311.16424*, 2023. 2, 5, 8

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[23] Bo Jiang, Shiqian Ma, and Shuzhong Zhang. Alternating direction method of multipliers for real and complex polynomial optimization models. *Optimization*, 63(6):883–898, 2014. 3

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[25] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2151–2162, 2023. 7, 8

[26] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022. 6, 8, 16

[27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1

[28] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 1

[29] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343, 2022. 1

[30] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023. 1

[31] Torgny Lindvall. *Lectures on the coupling method*. Courier Corporation, 2002. 3

[32] WB Liu and JW Barrett. Numerical methods for nonlinear variational problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 28(725-744), 1994. 3

[33] Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[34] Sindri Magnússon, Pradeep Chathuranga Weeraddana, Michael G Rabbat, and Carlo Fischione. On the convergence of alternating direction lagrangian methods for nonconvex structured optimization problems. *IEEE Transactions on Control of Network Systems*, 3(3):296–309, 2015. 3

[35] 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. *arXiv preprint arXiv:2108.01073*, 2021. 5

[36] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965. 2

[37] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 1

[38] Yurii Nesterov et al. *Lectures on convex optimization*. Springer, 2018. 12

[39] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and trends® in Optimization*, 1(3):127–239, 2014. 2

[40] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 1

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[42] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *International Conference on Machine Learning*, pages 8857–8868. PMLR, 2021. 1

[43] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971. 12

[44] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*. Springer Science & Business Media, 2009. 2

[45] 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*, pages 10684–10695, 2022. 1

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[47] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised

learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 1

[48] 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 *International Conference on Machine Learning*, pages 32483–32498. PMLR, 2023. 2, 5, 7, 8

[49] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2, 6, 7, 8, 16

[51] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 7

[52] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1

[53] Xiran Wang and Stanley H Chan. Parameter-free plug-and-play admm for image restoration. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1323–1327. IEEE, 2017. 8

[54] Xiaoyu Xiang, Ding Liu, Xiao Yang, Yiheng Zhu, Xiaohui Shen, and Jan P Allebach. Adversarial open domain adaptation for sketch-to-photo synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1434–1444, 2022. 5

[55] Yoshinao Yazaki, Yuichi Tanaka, and Stanley H Chan. Interpolation and denoising of graph signals using plug-and-play admm. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5431–5435. IEEE, 2019. 8

[56] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *arXiv preprint arXiv:2409.15761*, 2024. 2, 8

[57] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 5

[58] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23174–23184, 2023. 2, 5, 8

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[60] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023. 4