EMControl: Adding Conditional Control to Text-to-Image Diffusion Models via Expectation-Maximization

He Wang, Longquan Dai*, Jinhui Tang

Nanjing University of Science and Technology, China
{wanghe, dailongquan, jinhuitang}@njust.edu.cn

Abstract

Recent advances in diffusion models focus on efficiently handling conditional generative tasks without extra training. The process involves decomposing the result into two components: 1. unconditional sample, generated in the absence of conditions; 2. condition correction, adjusting unconditional sample to include the guidance image. This adjustment is quantified by the pixel-level measure, where the latent is decoded back into a pixel image, and the forward operator translates the noisy image into the guidance domain for comparison with the guidance image. To enhance the fidelity of condition correction, we propose a learnable latent forward operator, focusing on latent-space consistency with the expectation that this latent-space consistency approximates the pixel-level fidelity measure. The encoder translates the guidance image into the latent space, and a correctional operator is proposed to rectify model mismatching in the latent guidance model. The determination of the condition term and the correction estimation is akin to solving a blind inverse problem. Our EMControl employs the Expectation-Maximization (EM) algorithm to solve the blind inverse problem during the reverse sampling process. This technique ensures that samples, once consistent with the guidance, are accurately mapped back onto the noisy data manifold, adhering to the data's inherent distribution. The EMControl has proven its effectiveness by delivering superior performance in conditional diffusion generation tasks compared to previous approaches. Moreover, its application to multiple-condition scenarios underscores its versatility and robustness across a range of generative tasks.

Introduction

Over the past few years, diffusion models (Sohl-Dickstein et al. 2015; Song and Ermon 2019; Ho, Jain, and Abbeel 2020; Song et al. 2021b) have achieved remarkable success, particularly in the domain of conditional diffusion models (Dhariwal and Nichol 2021; Rombach et al. 2022; Zhang, Rao, and Agrawala 2023), due to their powerful expressive and re-editing capabilities. These models have demonstrated exceptional performance across a range of generative tasks, such as: image generation (Nichol and Dhariwal 2021; Song and Ermon 2020; Song et al. 2021a), image inpainting (Chung et al. 2023b), person synthesis (Shen et al. 2023; Shen and Tang 2024), image editing (Choi et al. 2021).

Conditional diffusion models generally employ two techniques: classifier-guided (Dhariwal and Nichol 2021) and classifier-free (Ho and Salimans 2021) diffusion models. Despite their effectiveness, these methods encounter challenges related to learning cost and model generality, as they require additional training and data for conditional generation. Recent advances (Chung et al. 2022; Zhu et al. 2023; Yu et al. 2023; Bansal et al. 2024; Yang et al. 2024a) have addressed these issues by developing training-free methods that leverage off-the-shelf loss guidance.

Yu et al. (2023) decomposed the latent code $\boldsymbol{z}_t$ into two components: unconditional sample $\boldsymbol{u}_t$ and conditional correction $\boldsymbol{\phi}$. Given the guidance image $\boldsymbol{I}_c$, their objective is to minimize $\|\mathcal{A}(\mathcal{D}(\boldsymbol{z}_t))-\boldsymbol{I}_c\|_2$, where $\boldsymbol{z}_t = \mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t) = \boldsymbol{u}_t+\boldsymbol{\phi}$, $\mathcal{D}(\boldsymbol{z}_t)$ is the decoder, and $\mathcal{A}(\boldsymbol{x}_t)$ denotes the forward operator that maps the noisy image $\boldsymbol{x}_t$ to the guidance domain. However, there is the divergence between the clean data manifold and the noisy latent $\boldsymbol{z}_t$. Existing approaches typically utilize an off-the-shelf forward operator $\mathcal{A}(\boldsymbol{x}_t)$, which is trained on clean images like $\mathcal{D}(\boldsymbol{z}_0)$. Applying it to noisy images $\boldsymbol{x}_t = \mathcal{D}(\boldsymbol{z}_t)$ often leads to suboptimal performance. To mitigate this, researchers (Chung et al. 2023b; Yu et al. 2023; Bansal et al. 2024) have developed methods to refine $\boldsymbol{\phi}$, minimizing the gap between $\mathcal{A}(\boldsymbol{x}_t)$ and $\boldsymbol{I}_c$.

Our approach advocates for a paradigm shift in guidance, moving from the pixel domain to the latent domain to refine $\boldsymbol{\phi}$. Central to this method is the introduction of a learnable latent forward operator $\mathfrak{A}_\theta(\boldsymbol{z}_t, t)$, designed to minimize expression $\min_{\boldsymbol{\theta}}\|\mathfrak{A}_\theta(\boldsymbol{z}_t, t) - \mathcal{E}(\boldsymbol{I}_c)\|_2$. Here, $\mathcal{E}(\boldsymbol{I}_c)$ represents the latent encoding of the guidance image $\boldsymbol{I}_c$. Note that this approach operates within the latent domain, contrasting with the pixel-domain comparison $\|\mathcal{A}(\mathcal{D}(\boldsymbol{z}_t)) - \boldsymbol{I}_c\|_2$.

Both $\mathcal{A}(\boldsymbol{x}_t)$ and $\mathfrak{A}(\boldsymbol{z}_t, t)$ are often simplified, omitting complexities of systems that are difficult to capture. Inversion algorithms (Bertero and Boccacci 2020), suffer from forward model mismatches, leading to biased reconstructions. For example, in non-blind deconvolution results (Cho, Wang, and Lee 2011), these biases manifest themselves as reconstruction artifacts due to the use of an inaccurate blur kernel. Similarly, the latent forward operator $\mathfrak{A}(\boldsymbol{z}_t, t)$ also exhibits mismatches in conditional diffusion models. To mitigate the impact of model mismatch, we propose a correctional opera-

*Corresponding author.

tor $\mathfrak{C}''_{\boldsymbol{\psi}}(\boldsymbol{z}_t) = \boldsymbol{z}_t + \boldsymbol{\psi}$, which takes the correction term $\boldsymbol{\psi}$ to refine the data consistency $\|\mathfrak{A}_{\boldsymbol{\theta}}(\mathfrak{C}''_{\boldsymbol{\psi}}(\boldsymbol{z}_t), t) - \mathcal{E}(\boldsymbol{I}_c)\|_2$.

The goal of conditional diffusion models is to generate samples that are not only realistic but also adhere to specific control signal encoded from certain conditions. To formalize this objective, we define two key elements: 1, $q_{\boldsymbol{\phi}}(\boldsymbol{z}_t)$ is the implicit distribution, resulting from the transformation $\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t)$ on the unconditional sample $\boldsymbol{u}_t$; 2, $p(\boldsymbol{z}_t|\boldsymbol{\psi}, \mathcal{E}(\boldsymbol{I}_c))$ is the posterior distribution of the latent variables $\boldsymbol{z}_t$, given a correction term $\boldsymbol{\psi}$ and an encoded guidance $\mathcal{E}(\boldsymbol{I}_c)$. Our approach involves minimizing the Kullback-Leibler (KL) divergence between these two distributions. The target is to identify the condition $\boldsymbol{\phi}$ and the correction $\boldsymbol{\psi}$, thereby aligning the implicit distribution with the posterior distribution that reflects the desired conditions. We attribute this minimization as the blind inverse problem (Gao et al. 2021; Gan et al. 2023; Chung et al. 2023a).

EMControl addresses the blind inverse problem involving two parameters, $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, by employing the Expectation-Maximization (EM) algorithm. Utilizing the guidance $\mathcal{E}(\boldsymbol{I}_c)$, the EM algorithm simultaneously estimates the condition $\boldsymbol{\phi}$ and correction $\boldsymbol{\psi}$. This tandem estimation aims to reconstruct the observed measurements. The adoption of a variational inference-based framework enables the efficient handling and optimization of complex probabilistic distributions.

EMControl is a versatile framework applicable to various conditional diffusion models described by the latent forward operators $\mathfrak{A}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$. In the experimental section, we showcase the effectiveness of our EMControl sampling in adding conditional control to text-to-image diffusion models, where we significantly outperform other methods. Additionally, we also demonstrate the generality of EMControl by applying it to multiple conditional generation tasks using text-to-image diffusion models.

Related Work

In this section, we offer a review of the existing literature on conditional diffusion models (CDMs) and inverse algorithm.

Training-required CDMs are primarily split into two categories. The first category includes classifier-guided diffusion models (Dhariwal and Nichol 2021), which leverage a pre-existing diffusion model to train a time-dependent classifier aimed at approximating the posterior probability $p(y|x_t)$. The second category comprises classifier-free diffusion models, exemplified by ControlNet (Zhang, Rao, and Agrawala 2023) and its derivatives (Qin et al. 2023; Yang et al. 2024a). These models skip the classifier training step, opting instead to train a conditional denoiser directly using paired data. Although they guarantee high-fidelity and realistic sample generation, they incur the unavoidable costs of training time.

Training-free CDMs have emerged to address these challenges, enabling conditional generation without training. These models use existing classifiers to approximate the gradient $\nabla_{\boldsymbol{z}_t} \log(y|\boldsymbol{z}_t)$. Chung et al. (2022) employ Tweedie's formula to tackle linear inverse problems, a concept later generalized for broader conditional generation by Chung et al. (2023b), Yu et al. (2023), and Bansal et al. (2024). Song et al. (2023) aim to mitigate bias through multiple samples from an imprecise Gaussian distribution but faces the computational demands of Monte Carlo simulations. Zhu et al. (2023) apply guidance to a pristine data sample $\boldsymbol{z}_0$ and project it onto the intermediate data manifold $\boldsymbol{z}_t$. He et al. (2024) refine this approach by incorporating an auto-encoder to enforce guidance within the tangent space of the clean data manifold. Yang et al. (2024a) adeptly manage manifold deviations without imposing stringent assumptions, thereby enhancing both quality of the samples and efficiency of the process.

Inverse Problems reconstruct an original image from corrupted measurements influenced by a forward operator. They are divided into two categories: 1. Non-blind inverse problems (Bertero and Boccacci 2020): The forward operator is known, facilitating image reconstruction. 2. Blind inverse problems (Gao et al. 2021; Gan et al. 2023; Chung et al. 2023a): These are more challenging due to the unknown forward operator, requiring its estimation alongside the image. Blind inverse problems present a significant challenge as they are ill-posed, potentially leading to multiple solutions. To identify the most likely solution, additional information such as image priors is crucial. However, current research in this area has mainly targeted specific scenarios, such as spatially-invariant blind deconvolution (Levin et al. 2011) and CT (Xie et al. 2021) with simple rotational errors, limiting the generalizability of the results to more intricate situations.

Method

In this section, we describe our EMControl sampling to add conditional control to text-to-image diffusion models.

Latent Guidance

We introduce three operators to facilitate the transition from pixel guidance to latent guidance. These include the latent forward operator $\mathfrak{A}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$, the conditional operator $\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t) = \boldsymbol{u}_t + \boldsymbol{\phi}$, and the correctional operator $\mathfrak{C}''_{\boldsymbol{\psi}}(\boldsymbol{z}_t) = \boldsymbol{z}_t + \boldsymbol{\psi}$. This transformation enables the conversion of the pixel-based guidance measure $\|\mathcal{A}(\mathcal{D}(\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t))) - \boldsymbol{I}_c\|_2$ to its latent counterpart $\|\mathfrak{A}_{\boldsymbol{\theta}}(\mathfrak{C}''_{\boldsymbol{\psi}}(\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t)), t) - \mathcal{E}(\boldsymbol{I}_c)\|_2$. With the objective function given by Eq. (1), we train our latent forward model $\mathfrak{A}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$ using paired data $(\boldsymbol{z}_t, \mathcal{E}(\boldsymbol{I}_c))$, where $\boldsymbol{z}_t$ is the noisy latent code of the natural image, the guidance image $\boldsymbol{I}_c$ is extract from the natural image using detection algorithms as implemented by ControlNet (Zhang, Rao, and Agrawala 2023).

$$\min_{\boldsymbol{\theta}} \|\mathfrak{A}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t) - \mathcal{E}(\boldsymbol{I}_c)\|^2 \tag{1}$$

Given this setup, it is reasonable to expect that the latent guidance $\mathcal{E}(\boldsymbol{I}_c)$ will closely mirror the original pixel guidance.

The transition to this new latent guidance offers several distinct advantages: 1. **Computational Convenience**: Edge detection in generated images is straightforward, yet generating segmentation or depth maps without additional networks remains a challenge. The latent forward model $\mathfrak{A}_{\theta}(\boldsymbol{z}_t, t)$ is not inherently equipped to distinguish among different control modalities such as edges or depth. 2. **Computational Efficiency**: The latent representation of control guidance is compact, enabling the final output to be influenced with a

minimal number of control signals. This approach is more efficient than direct pixel-level manipulation due to the reduced complexity of the control inputs. 3. **Computational Robustness**: The diffusion network creates intermediate images $\boldsymbol{z}_t$ that are inherently noisy. Existing networks are difficult to produce accurate segmentation or depth maps. In contrast, our network is trained across various noise levels, ensuring the generation of clear latent control representations, and thereby enhancing the robustness of the results.

To address the potential discrepancy between training and inference datasets that may result in model mismatch, we propose the incorporation of a correctional operator: $\mathfrak{C}''_{\boldsymbol{\psi}}(\boldsymbol{z}_t)$. The correction term $\boldsymbol{\psi}$ is carefully tuned throughout the sampling process to decrease the domain gap, ensuring that the inference data are more closely aligned with the training data.

Cost Function

Song and Ermon (2019) and Song et al. (2021b) are predicated on score matching, focusing on the estimation of the score function $\nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t)$. Within the DDPM sampling process (Ho, Jain, and Abbeel 2020), these models systematically infer the preceding state $\boldsymbol{z}_{t-1}$ from the current state $\boldsymbol{z}_t$ using the following formula (2):

$$\boldsymbol{z}_{t-1} = (1 + \frac{1}{2}\beta_t)\boldsymbol{z}_t + \beta_t \nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t) + \sqrt{\beta_t}\boldsymbol{\epsilon}, \quad (2)$$

Here, $\boldsymbol{\epsilon}$ represents Gaussian noise sampled at random, and β_t is a predefined parameter.

Song et al. (2021b) proposed to control the generated results with a guidance $\boldsymbol{I}_c$ by modifying the score function as $\nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t|\boldsymbol{I}_c)$. Utilizing Bayes' theorem, this conditional score function can be decomposed into two components:

$$\nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t|\boldsymbol{I}_c) = \nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t) + \nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{I}_c|\boldsymbol{z}_t)$$

where the first term $\nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t)$ is estimated using a pre-trained unconditional score estimator $s(\boldsymbol{z}_t, t)$, and the second term represents the contribution for conditional diffusion modeling. The conditional sampling is then defined as:

$$\boldsymbol{z}_{t-1} = \boldsymbol{u}_t + \rho_t \nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{I}_c|\boldsymbol{z}_t),$$

where $\boldsymbol{u}_t = (1 + \frac{1}{2}\beta_t)\boldsymbol{z}_t + \beta_t \nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t) + \sqrt{\beta_t}\boldsymbol{\epsilon}$ which is same to Eq (2) and ρ_t serves as a scaling factor.

Let $q_{\boldsymbol{\phi}}(\boldsymbol{z}_t)$ represent the implicit distribution arising from the transformation $\boldsymbol{z}_t = \mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t)$, where $\boldsymbol{u}_t$ is derived from the unconditional sampling formula (2), and let $p(\boldsymbol{z}_t|\boldsymbol{\psi}, \mathcal{E}(\boldsymbol{I}_c))$ denote the posterior distribution given by:

$$p(\boldsymbol{z}_t|\boldsymbol{\psi}, \mathcal{E}(\boldsymbol{I}_c)) \propto \exp\left(-\frac{1}{2\hat{\rho}^2}\|\mathfrak{A}_{\boldsymbol{\theta}}(\mathfrak{C}''_{\boldsymbol{\psi}}(\boldsymbol{z}_t), t) - \mathcal{E}(\boldsymbol{I}_c)\|_2\right).$$

To approximate the distribution $p(\boldsymbol{z}_t|\boldsymbol{\psi}, \mathcal{E}(\boldsymbol{I}_c))$ with $q_{\boldsymbol{\phi}}(\boldsymbol{z}_t)$, we aim to minimize the Kullback-Leibler (KL) divergence between these two distributions. This is achieved by finding the parameters $\boldsymbol{\phi}, \boldsymbol{\psi}$ that yield the smallest divergence, as expressed by the following equation:

$$\underset{\boldsymbol{\phi},\boldsymbol{\psi}}{\operatorname{argmin}} \operatorname{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_t) \parallel p(\boldsymbol{z}_t|\boldsymbol{\psi}, \mathcal{E}(\boldsymbol{I}_c))) \quad (3)$$

This cost function is similar to the one used in blind inverse problems (Gan et al. 2023; Chung et al. 2023a) that is solved by the EM algorithm.

Algorithm 1: EMControl Sampling

Input: noise $\boldsymbol{z}_T \sim \mathcal{N}(0, I)$, guidance image $\boldsymbol{I}_c$, guidance scale λ, standard deviations ρ, $\hat{\rho}$, $\tilde{\rho}$, $\bar{\rho}$.
for $t = T$ **to** 1 **do**
  $\boldsymbol{\epsilon}^{\{n\}} \sim \mathcal{N}(0, I)$
  $\boldsymbol{u}_{t-1}^{\{n\}} = (1 + \frac{1}{2}\beta_t)\boldsymbol{z}_t + \beta_t \nabla_{\boldsymbol{z}_t} \log p(\boldsymbol{z}_t) + \sqrt{\beta_t}\boldsymbol{\epsilon}^{\{n\}}$
  $\boldsymbol{\psi}^{(0)} = 0, L = 3$
  for $l = 1$ **to** L **do**
    $\boldsymbol{\phi}^{(l)} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} (\frac{1}{2\hat{\rho}^2}\|\mathfrak{A}_{\boldsymbol{\theta}}(\mathfrak{C}''_{\boldsymbol{\psi}^{(l-1)}}(\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_{t-1}^{\{n\}}))) - \mathcal{E}(\boldsymbol{I}_c)\|_2 - \frac{1}{2\rho^2}\|\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_{t-1}^{\{n\}}) - (\mathbb{E}(\boldsymbol{u}_{t-1}) + \boldsymbol{\phi})\|_2 + \frac{1}{2\tilde{\rho}^2}\|\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_{t-1}^{\{n\}}) - \mathbb{E}(\boldsymbol{u}_{t-1})\|_2)$ according to Eq (4).
    $\boldsymbol{\psi}^{(l)} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}} \frac{1}{N} \sum_{n=1}^{N} (\frac{1}{2\hat{\rho}^2}\|\mathfrak{A}_{\boldsymbol{\theta}}(\mathfrak{C}''_{\boldsymbol{\psi}}(\mathfrak{C}'_{\boldsymbol{\phi}^{(l)}}(\boldsymbol{u}_{t-1}^{\{n\}}))) - \mathcal{E}(\boldsymbol{I}_c)\|_2 + \frac{1}{2\tilde{\rho}^2}\|\boldsymbol{\psi}\|_2)$ according to Eq (5).
  end for
  $\boldsymbol{z}_{t-1} = \boldsymbol{u}_{t-1}^{\{0\}} + \lambda \cdot \boldsymbol{\phi}^{(L)}$
end for
Return x_0

Expectation-Maximization

We propose an EM approach that optimizes the objective (3) to recover the conditional parameters $\boldsymbol{\phi}$ of the conditional model $\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t) = \boldsymbol{u}_t + \boldsymbol{\phi}$ using latent guidance $\mathcal{E}(\boldsymbol{I}_c)$. Once learned, the updated conditional model can then be used to estimate the posterior distribution of the correction parameter $\boldsymbol{\psi}$ for the correctional model $\mathfrak{C}''_{\boldsymbol{\psi}}(\boldsymbol{z}_t) = \boldsymbol{z}_t + \boldsymbol{\psi}$. Our iterates between two stages that are inspired by the standard EM algorithm: (1) an E-step that learns a distribution, $q_{\boldsymbol{\phi}^{(l)}}(\boldsymbol{z}_t)$, to approximate the posterior distribution of $\boldsymbol{z}_t$ given the current condition term $\boldsymbol{\psi}^{(l-1)}$, and (2) an M-step that solves for correction estimation $\boldsymbol{\psi}^{(l)}$ that maximizes the expected value of the log likelihood function, with respect to the posterior distribution $q_{\boldsymbol{\phi}^{(l)}}(\boldsymbol{z}_t)$ estimated in the prior E-step.

E-step solves for a distribution $q_{\boldsymbol{\phi}}(\boldsymbol{z}_t)$ that approximates well the posterior distribution $p(\boldsymbol{z}_t|\boldsymbol{\psi}^{(l-1)}, \mathcal{E}(\boldsymbol{I}_c))$. For a batch size N, $\boldsymbol{u}_t^{\{n\}}$ is sampled from Eq (2) with different noise $\boldsymbol{\epsilon}^{\{n\}}$, $\boldsymbol{z}_t^{\{n\}} = \mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t^{\{n\}})$, $\mathbb{E}(\boldsymbol{u}_t) \approx \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{u}_t^{\{n\}}$ is the expectation of $\boldsymbol{u}_t$, we have

$$\begin{aligned}\boldsymbol{\phi}^{(l)} &= \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \operatorname{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}_t) \parallel p(\boldsymbol{z}_t|\mathcal{E}(\boldsymbol{I}_c), \boldsymbol{\psi}^{(l-1)})) \\ &\approx \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \frac{1}{N}\sum_{n=1}^{N} \log q_{\boldsymbol{\phi}}(\boldsymbol{z}_t^{\{n\}}) - \log p(\boldsymbol{z}_t^{\{n\}}) \\ &\qquad - \log p(\mathcal{E}(\boldsymbol{I}_c)|\boldsymbol{z}_t^{\{n\}}, \boldsymbol{\psi}^{(l-1)})\end{aligned} \quad (4)$$

Here $\log q_{\boldsymbol{\phi}}(\boldsymbol{z}_t) \propto -\frac{1}{2\rho^2}\|\boldsymbol{z}_t - (\mathbb{E}(\boldsymbol{u}_t) + \boldsymbol{\phi})\|_2$ is the estimated distribution used to approximate $p(\boldsymbol{z}_t|\mathcal{E}(\boldsymbol{I}_c), \boldsymbol{\psi}^{(l-1)})$, $\log p(\boldsymbol{z}_t) \propto -\frac{1}{2\tilde{\rho}^2}\|\boldsymbol{z}_t - \mathbb{E}(\boldsymbol{u}_t)\|_2$ is a prior that indicates $\boldsymbol{z}_t$ should not deviate the expectation $\mathbb{E}(\boldsymbol{u}_t)$ too much, $\log p(\mathcal{E}(\boldsymbol{I}_c)|\boldsymbol{z}_t, \boldsymbol{\psi}^{(l-1)})$ is the data likelihood. When assuming the latent guidance $\mathcal{E}(\boldsymbol{I}_c)$ experience $i.i.d$ additive Gaussian noise, we have $\log p(\mathcal{E}(\boldsymbol{I}_c)|\boldsymbol{z}_t, \boldsymbol{\psi}^{(l-1)}) \propto \frac{1}{2\bar{\rho}^2}\|\mathcal{E}(\boldsymbol{I}_c) - \mathfrak{A}_{\boldsymbol{\theta}}(\mathfrak{C}''_{\boldsymbol{\psi}^{(l-1)}}(\boldsymbol{z}_t))\|_2$.

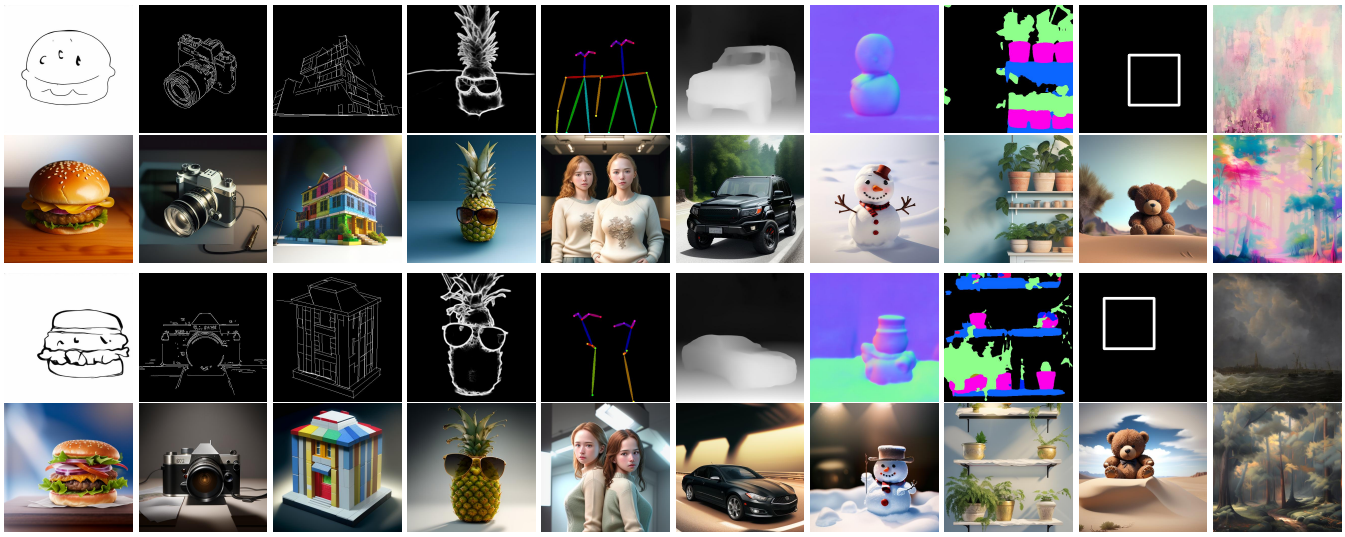

Figure 1: Visual Demonstration for Single Condition. Rows 1 & 3 depict the guidance that serves as the reference for the generation. Rows 2 & 4 exhibit the output, providing a side-by-side comparison that illustrates the alignment with the condition.

M-step uses $\phi^{(l)}$, from the prior E-step, to update ψ, the parameters of the correctional model $\mathfrak{C}''_{\psi}(\boldsymbol{z}_t^{\{n\}}) = \boldsymbol{z}_t^{\{n\}} + \psi$. This is achieved by stochastically solve

$$\psi^{(l)} \approx \underset{\psi}{\text{argmax}} \frac{1}{N} \sum_{n=1}^{N} \log p(\mathcal{E}(\boldsymbol{I}_c) | \boldsymbol{z}_t^{\{n\}}, \psi) + \log p(\psi) \tag{5}$$

where the prior $p(\psi)$ is be used to encourage the parameter of the correctional model to remain close to an initial model $\tilde{\psi}$ by defining $\log p(\psi) \propto \frac{1}{2\tilde{\rho}^2} \|\psi\|_2 + c$ with $\tilde{\psi} = 0$.

EMControl Sampling

Upon successfully training $\mathfrak{A}_{\theta}(\boldsymbol{z}_t, t)$, we proceed to apply the EM algorithm as detailed in Alg 1. This algorithm alternates between estimating ϕ and ψ. Specifically, steps 3 and 4 generate N estimates for $\boldsymbol{u}_{t-1}^{\{n\}}$, while steps 6 and 7 are dedicated to performing E-step and M-step of the EM process. Our experiments have demonstrated that a mere three iterations are adequate to secure commendable results. The final conditional outcome $\boldsymbol{z}_t$ is derived by adding the product of conditional correction ψ and guidance scale λ to the unconditional sample $\boldsymbol{u}_{t-1}^{\{0\}}$.

Experiments

We implement EMControl across various conditions, including canny edge (Canny 1986), depth map (Yang et al. 2024b), normal map (Vasiljevic et al. 2019), M-LSD lines (Gu et al. 2022), HED edge (Xie and Tu 2015), semantic segmentation (Cheng et al. 2022), skeleton (Cao et al. 2017), sketch, object location (Redmon et al. 2016) and style guidance (Radford et al. 2021). In this section, we present the generated results and provide a comparison with existing methods to demonstrate the effectiveness of our approach.

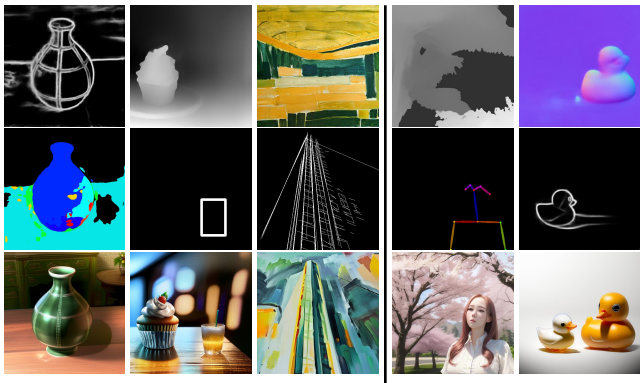

Figure 2: Visual Demonstration for Multiple Conditions. There are two approaches for multiple condition generation: 1, dual latent guidance networks; 2, single latent guidance with control adaptation. Columns 1-3 display the results generated by the first method. Columns 4-5 present the results from the second method, featuring the use of ControlNet and T2I-Adapter for controlled image generation.

Experimental Setup

Our model for the latent forward network $\mathfrak{A}_{\theta}(\boldsymbol{z}_t, t)$ is based on U-Net (Ronneberger, Fischer, and Brox 2015). During training, the model commenced with the SDv1.5 checkpoint and was trained for 20 hours on a single NVIDIA RTX3090 GPU. A batch size of 1 was utilized alongside the AdamW optimizer at a learning rate of 1e-5, where the inputs, including images and condition, were scaled down to 512×512 pixels. For EMControl sampling, we employed the DDPM (Ho, Jain, and Abbeel 2020) scheduler across 20 time steps.

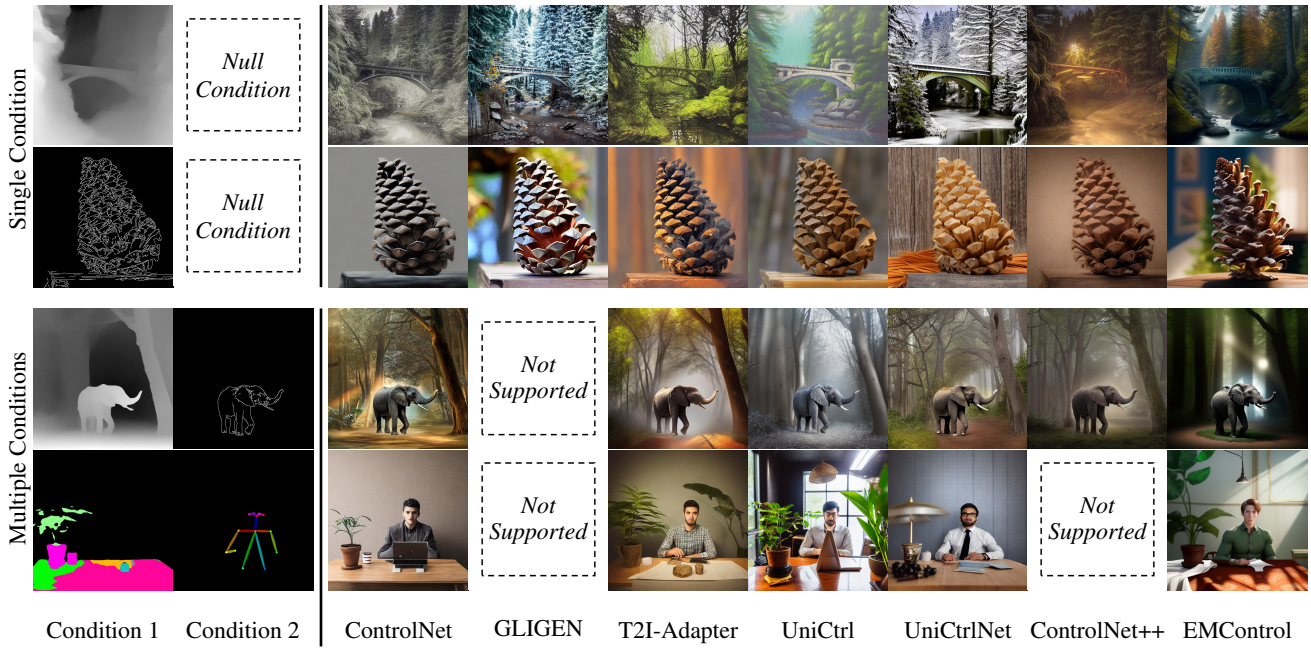

Figure 3: Visual Comparison for Training-Required Methods. The top two rows showcase the generation results under single conditions. The bottom two rows present the generation results under multiple conditions.

We trained our model on approximately 156,000 images from COCO2017 (Lin et al. 2014), covering a range of tasks. The generation of conditional images was facilitated by the same detection algorithms employed by ControlNet (Zhang, Rao, and Agrawala 2023). For the aspect of style guidance, we further integrated approximately 81,000 images from Wiki-Art (Tan et al. 2019). The reference style was extracted from the style images using CLIP's (Radford et al. 2021) image feature extractor.

Visual Demonstration

Here, we present visual results for both single and multiple conditions to illustrate the capabilities of EMControl.

Single Condition Generation: Fig 1 presents the results of EMControl, guided by a single latent network, across diverse image generation prompts. The prompts, from left to right, include: "a hamburger on cutting board", "a quaint camera on table", "a colorful house built of toy blocks", "a pineapple wearing sunglasses", "two students wearing sweater", "a cool black car running on road", "a cute snowman in vivid details", "the green plants on white metal shelf", "a teddy bear in desert", "an artist painting of forest", respectively. Our model demonstrates compatibility with ControlNet's range of conditional types and extends its functionality with enhanced support for object placement and stylistic guidance. The outputs are in close alignment with the text prompts and guidance images, confirming the model's efficacy and versatility in conditional image synthesis.

Multiple Condition Generation: We developed two multiple condition generation approaches: 1, dual latent guidance networks: This method employs separate latent guidance networks to independently address multiple conditions. 2, single latent guidance with control adaptation: This method leverages a singular latent network that is complemented by a controlnet-like network, each dedicated to handling separate conditions.

Figure 4: Visual Comparison for Training-Free Methods. Our outputs demonstrate improved alignment with the conditions, reflecting enhanced quality in the generated results.

In Fig. 2, the initial three columns present the results of the first approach, highlighting multiple latent guidance models that can synergize effectively. In the first column, the model uses a segmentation map to provide a basic structure, such as a bottle, and refines it with a HED map for detailed enhancement. In the second column, a depth map shapes a cake, while an object map introduces juice, merging to form a cohesive image. In the third column, we apply the aesthetics of a style image to an M-LSD map, imbuing a building with an artistic touch. All these demonstrations confirm the model's versatility in handling a variety of conditional inputs, reflecting its robust and efficient performance.

	Depth	Canny	Segmentation	HED	Skeleton	Normal	M-LSD	Location	Sketch
FID↓									
ControlNet	19.3821	**16.5397**	21.9932	20.1351	56.7648	27.9317	**20.5671**	-	-
GlIGEN	23.2931	24.8427	27.4012	26.3622	52.5637	27.7452	-	**24.2495**	-
T2i-Adapter	23.9011	17.1326	**21.7609**	-	**35.6791**	-	-	-	28.7961
UniCtrl	24.2794	19.1032	29.6965	20.0217	40.3571	29.6139	-	29.7812	-
UniCtrlNet	24.9583	18.0131	22.7149	18.5472	65.3125	-	27.8137	-	**24.0166**
ControlNet++	**17.4087**	20.0955	24.5436	**16.2718**	-	-	-	-	-
EMControl	25.1024	25.2631	25.1132	24.6031	43.0931	**27.3594**	26.8112	33.6821	26.4697
CLIP text-image score↑									
ControlNet	0.2813	0.2862	0.2829	0.2844	0.2597	0.2734	0.2876	-	-
GlIGEN	0.2983	0.2933	0.2846	0.2792	0.2601	0.2694	-	0.2778	-
T2i-Adapter	0.2995	0.3026	0.3001	-	**0.3094**	-	-	-	0.2711
UniCtrl	**0.3054**	0.3011	**0.3053**	0.2995	0.3073	0.2959	-	**0.2973**	-
UniCtrlNet	0.3001	0.3037	0.3048	0.3023	0.2814	-	0.2867	-	0.2968
ControlNet++	0.3023	**0.3074**	0.2986	0.3011	-	-	-	-	-
EMControl	0.2919	0.3035	0.3012	**0.3045**	0.2931	**0.2961**	**0.2951**	0.2861	**0.3011**
CLIP aesthetic score↑									
ControlNet	5.1792	5.2112	5.3003	5.2543	5.2403	5.0786	5.3069	-	-
GlIGEN	5.1389	5.0507	4.9296	4.9535	4.8968	4.8253	-	5.3711	-
T2i-Adapter	5.1032	5.1264	4.9684	-	5.2947	-	-	-	4.8532
UniCtrl	5.3511	5.1646	**5.3918**	5.1691	**5.4810**	5.1049	-	5.2628	-
UniCtrlNet	5.0147	5.0032	5.0565	5.0039	4.9613	-	4.9741	-	5.0108
ControlNet++	5.3010	5.1209	4.9357	5.1203	-	-	-	-	-
EMControl	**5.4137**	**5.4137**	5.3896	**5.4392**	5.4769	**5.4139**	**5.4293**	**5.3965**	**5.4767**
Condition Fidelity↑	MSE↓	SSIM↑	mIoU↑	SSIM↑	mAP↑	MSE↓	SSIM↑	mAP↑	SSIM↑
ControlNet	88.8735	0.4363	0.4234	0.5837	0.4394	86.8683	0.7546	-	-
GlIGEN	81.2194	0.4011	0.2476	0.4096	0.2013	90.2429	-	0.2431	-
T2i-Adapter	94.5428	0.3996	0.2342	-	**0.4986**	-	-	-	0.3872
UniCtrl	88.8363	**0.5341**	0.3285	0.3628	0.2487	104.0974	-	**0.2728**	-
UniCtrlNet	99.3754	0.4682	0.3048	0.6034	0.2175	-	0.7328	-	**0.6537**
ControlNet++	86.7210	0.5395	**0.5456**	**0.6813**	-	-	-	-	-
EMControl	**86.6693**	0.4476	0.4034	0.4896	0.4173	**70.5733**	**0.7692**	0.2631	0.6273

Table 1: Quantitative Comparison. The superior outcomes are highlighted in bold. The symbol "-" signifies that the approach lacks a publicly accessible model for evaluation. EMControl has attained the leading rank across several criteria.

In the latter two columns of Fig 2, we present the outcomes from our second approach. Here, the initial condition is managed by a latent guidance network, while the subsequent condition is addressed by a controlnet-like network, exemplified by ControlNet (Zhang, Rao, and Agrawala 2023) and T2i-Adapter (Mou et al. 2024). This integration highlights the seamless compatibility of our method with a variety of conditional models, facilitating sophisticated multi-conditional image generation. Moreover, EMControl's compatibility with any model sharing the same encoder expands its utility across numerous community-tailored versions. This universal adaptability underscores the approach's versatility and its promising prospects for broader applications.

Qualitative Comparison

In this section, we conduct a qualitative comparison between our method and other approaches to reveal the distinct advantages of our approach.

Training-required Methods: We perform qualitative assessments comparing single-condition and multi-condition control methods to provide a comprehensive analysis that underscores the strengths of our approach. In the upper rows of Fig. 3, a single condition control comparison is displayed with models including ControlNet (Zhang, Rao, and Agrawala 2023), GLIGEN (Li et al. 2023), T2I-Adapter (Mou et al. 2024), UniCtrl (Qin et al. 2023), UniCtrlNet (Zhao et al. 2024), and ControlNet++ (Li et al. 2024). We focus on depth and canny images to evaluate against prompts such as "a bridge among the forests" and "a pine cone on the table". Our method matches these in generation quality while optimizing training resource usage. In lower rows of Fig. 3, we assess models that handle multi-condition inputs: ControlNet, T2I-Adapter, UniCtrl, UniCtrlNet, and ControlNet++. The prompts are "a bridge among the forests" and "a pine cone on the table". The label "Not Supported" indicates that published weights for certain methods are unavailable. Our method continues to exhibit strong performance in terms of generation quality and alignment with the given conditions.

Training-free Methods: We zero in on semantic segmentation and style guidance for comparison, highlighting the constraints of current training-free methods. Our approach is compared to Freedom (Yu et al. 2023), UniGuid (Bansal et al. 2024), and DSG (Yang et al. 2024a), which operate without additional training but may not match our versatility.

	Training-required Methods						Training-free Methods			
	ControlNet	GLIGEN	T2I-Adapter	UniCtrl	UniCtrlNet	ControlNet++	FreeDom	UniGuid	DSG	EMControl
Training(GPU Hours)	500	1000	192	5000	6900	60	-	-	-	20
Sampling(Seconds)	3	7	3	5	4	3	115	4524	74	17

Table 2: Efficiency Comparison. "-" indicates that authors of these method does not provide this value. EMControl strikes a balance between training efficiency and sampling speed. Compared to training-required methods, it demands a modest investment in training. Conversely, when juxtaposed with methods devoid of training, EMControl delivers accelerated sampling capabilities.

Fig. 4 juxtaposes our method using the prompts "park" and "chairs". It plots our superior generation quality over other training-free techniques and hints at our swifter inference times, explored further in the Efficiency Comparison section. These enhancements -which cover an expanded range of conditions, improved result quality, and faster inference - underscore the substantial advantages of our methodology, indicating its vast application potential.

Quantitative Comparison

To evaluate the performance of different methods, we used the COCO2017 validation set comprising 5,000 image-text pairs. The uniformity in the evaluation was ensured by conducting all experiments at a resolution of 512×512, using the DDPM scheduler over the 20 time steps. Our quantitative comparison encompassed six training-required methods: ControlNet, GLIGEN, T2I Adapter, UniCtrl, UniCtrlNet, and ControlNet++. Training-free methods were not included in this quantitative assessment due to their extended inference times and narrower versatility in managing a wide array of conditions.

We employed a suite of metrics—FID (Heusel et al. 2017), CLIP text-image score (Radford et al. 2021), and CLIP aesthetic score (Schuhmann et al. 2022)—to evaluate performance, as outlined in Table 1. The FID score can be ascribed to the assessment being based on a limited sample of 2.7k images from the COCO2017 test set. Our method holds its own against existing methods in terms of FID and CLIP text-image score, often outperforming them in CLIP aesthetic scoring across various conditions. To further ensure a thorough evaluation, we gauged Condition Fidelity by analyzing the correlation between the input conditions and those identified in the images rendered by our diffusion models. For this purpose, we deployed SSIM (Structural Similarity) (Wang et al. 2004), mAP (mean Average Precision) (Everingham et al. 2010), MSE (Mean Squared Error) (Sara, Akter, and Uddin 2019) and mIoU (Mean Intersection over Union) (Rezatofighi et al. 2019) as our evaluative tools. Our method proved adept at preserving the fidelity of conditions, as indicated by robust performance across all metrics utilized.

Efficiency Comparison

Here, we focus on an efficiency analysis of our method, highlighting the training and inference times. We compare it with both training-dependent and training-independent methods to highlight the advantages of our approach. Efficiency metrics were calculated under a segmentation condition scenario, with 20 timesteps allocated for training-dependent methods and 500 for those independent of training.

As illustrated in Table 2, our method substantially cuts down on training time compared to training-required methods. For instance, ControlNet++ needs 60 hours of training, thanks to its fine-tuning from a pre-trained ControlNet model. This superior efficiency is due to the training of a measurement model, which is less complex than a generative network. Moreover, our method also excels in inference speed, surpassing traditional training-free methods, mainly because it controls at the latent level instead of the pixel level.

Influence of Guidance Scale

EMControl achieves conditional sampling by adjusting the unconditional sample $\boldsymbol{u}_{t-1}$ through a conditional adjustment $\boldsymbol{\phi}$ scaled by the guidance scale λ, as detailed in Alg 1 line 10. Fig 5 (a) provides a condition image, prompted with "a wooden house in the forest". As shown in Fig 5(b), when λ is low, the results generated are equivalent to unconditional sampling. Conversely, as illustrated in Fig 5(d), when λ is high, the generated results closely match the conditional image, but may suffer from color distortions. Fig 5(c) demonstrates that with optimal λ, the generated images are both aesthetically pleasing and well-aligned with the conditions.

(a) Condition (b) $\lambda = 0$ (c) $\lambda = 1$ (d) $\lambda = 10$

Figure 5: Visualization of Guidance Scale λ. A diminished λ leads to an unconditional model, while an elevated λ introduces variations in color representation.

Conclusion

We present EMControl, a novel conditional sampling method adept at handling a variety of conditional generation tasks. Our approach features two key innovations: 1. We introduce the latent guidance network $\mathfrak{A}(\boldsymbol{z}_t, t)$ to translate visual guidance into the latent space, refining the conditional generation process. 2. We leverage an EM algorithm to compute the parameters $\boldsymbol{\phi}$ for the conditional operator $\mathfrak{C}'_{\boldsymbol{\phi}}(\boldsymbol{u}_t)$, facilitating the generation of $\boldsymbol{z}_t$ from the unconditional sample $\boldsymbol{\mu}_t$. EMControl's adaptability makes it an ideal plug-and-play addition to conditional diffusion models, requiring only minor code adjustments and minimal computational expense, while delivering substantial performance enhancements.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62372237, 62072238) and the Major Science and Technology Projects in Jiangsu Province under Grant (BG2024042).

References

Bansal, A.; Chu, H.-M.; Schwarzschild, A.; Sengupta, S.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2024. Universal Guidance for Diffusion Models. In *International Conference on Learning Representations*.

Bertero, M.; and Boccacci, P. 2020. *Introduction to Inverse Problems in Imaging*. CRC Press.

Canny, J. 1986. A Computational Approach to Edge Detection. *Transactions on Pattern Analysis and Machine Intelligence*.

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *Transactions on Pattern Analysis and Machine Intelligence*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *Conference on Computer Vision and Pattern Recognition*.

Cho, S.; Wang, J.; and Lee, S. 2011. Handling Outliers in Non-Blind Image Deconvolution. In *International Conference on Computer Vision*.

Choi, J.; Kim, S.; Jeong, Y.; Gwon, Y.; and Yoon, S. 2021. ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. In *International Conference on Computer Vision*.

Chung, H.; Kim, J.; Kim, S.; and Ye, J. C. 2023a. Parallel Diffusion Models of Operator and Image for Blind Inverse Problems. In *Conference on Computer Vision and Pattern Recognition*.

Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; and Ye, J. C. 2023b. Diffusion Posterior Sampling for General Noisy Inverse Problems. In *International Conference on Learning Representations*.

Chung, H.; Sim, B.; Ryu, D.; and Ye, J. C. 2022. Improving Diffusion Models for Inverse Problems using Manifold Constraints. In *Advances in Neural Information Processing Systems*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*.

Gan, W.; Shoushtari, S.; Hu, Y.; Liu, J.; An, H.; and Kamilov, U. 2023. Block Coordinate Plug-and-Play Methods for Blind Inverse Problems. In *Advances in Neural Information Processing Systems*.

Gao, A.; Castellanos, J.; Yue, Y.; Ross, Z.; and Bouman, K. 2021. DeepGEM: Generalized Expectation-Maximization for Blind Inversion. In *Advances in Neural Information Processing Systems*.

Gu, G.; Ko, B.; Go, S.; Lee, S.-H.; Lee, J.; and Shin, M. 2022. Towards Light-Weight and Real-Time Line Segment Detection. In *AAAI Conference on Artificial Intelligence*.

He, Y.; Murata, N.; Lai, C.-H.; Takida, Y.; Uesaka, T.; Kim, D.; Liao, W.-H.; Mitsufuji, Y.; Kolter, J. Z.; Salakhutdinov, R.; and Ermon, S. 2024. Manifold Preserving Guided Diffusion. In *International Conference on Learning Representations*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*.

Levin, A.; Weiss, Y.; Durand, F.; and Freeman, W. T. 2011. Efficient Marginal Likelihood Optimization in Blind Deconvolution. In *Conference on Computer Vision and Pattern Recognition*.

Li, M.; Yang, T.; Kuang, H.; Wu, J.; Wang, Z.; Xiao, X.; and Chen, C. 2024. ControlNet++: Improving Conditional Controls with Efficient Consistency Feedback. In *European Conference on Computer Vision*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Conference on Computer Vision and Pattern Recognition*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2I-Adapter: Learning Adapters to Dig Out More Controllable Ability for Text-to-Image Diffusion Models. In *AAAI Conference on Artificial Intelligence*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In *International Conference on Machine Learning*.

Qin, C.; Zhang, S.; Yu, N.; Feng, Y.; Yang, X.; Zhou, Y.; Wang, H.; Niebles, J. C.; Xiong, C.; Savarese, S.; Ermon, S.; Fu, Y.; and Xu, R. 2023. UniControl: A Unified Diffusion Model for Controllable Visual Generation In the Wild. In *Advances in Neural Information Processing Systems*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Conference on Computer Vision and Pattern Recognition*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Conference on Computer Vision and Pattern Recognition*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Conference on Computer Vision and Pattern Recognition*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*.

Sara, U.; Akter, M.; and Uddin, M. S. 2019. Image Quality Assessment through FSIM, SSIM, MSE and PSNR—a Comparative Study. *Journal of Computer and Communications*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models. In *Advances in Neural Information Processing Systems*.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *Conference on Neural Information Processing Systems*.

Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2023. Advancing pose-guided image synthesis with progressive conditional diffusion models. In *International Conference on Learning Representations*.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning*.

Song, J.; Zhang, Q.; Yin, H.; Mardani, M.; Liu, M.-Y.; Kautz, J.; Chen, Y.; and Vahdat, A. 2023. Loss-Guided Diffusion Models for Plug-and-Play Controllable Generation. In *International Conference on Machine Learning*.

Song, Y.; Durkan, C.; Murray, I.; and Ermon, S. 2021a. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems*.

Song, Y.; and Ermon, S. 2020. Improved Techniques for Training Score-Based Generative Models. In *Advances in Neural Information Processing Systems*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021b. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Tan, W. R.; Chan, C. S.; Aguirre, H. E.; and Tanaka, K. 2019. Improved ArtGAN for Conditional Synthesis of Natural Image and Artwork. *Transactions on Image Processing*.

Vasiljevic, I.; Kolkin, N.; Zhang, S.; Luo, R.; Wang, H.; Dai, F. Z.; Daniele, A. F.; Mostajabi, M.; Basart, S.; Walter, M. R.; et al. 2019. DIODE: A Dense Indoor and Outdoor DEpth Dataset. arXiv:1908.00463.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image Quality Assessment: from Error Visibility to Structural Similarity. *Transactions on Image Processing*.

Xie, M.; Liu, J.; Sun, Y.; Gan, W.; Wohlberg, B.; and Kamilov, U. S. 2021. Joint Reconstruction and Calibration Using Regularization by Denoising with Application to Computed Tomography. In *International Conference on Computer Vision Workshops*.

Xie, S.; and Tu, Z. 2015. Holistically-Nested Edge Detection. In *International Conference on Computer Vision*.

Yang, L.; Ding, S.; Cai, Y.; Yu, J.; Wang, J.; and Shi, Y. 2024a. Guidance with Spherical Gaussian Constraint for Conditional Diffusion. In *International Conference on Machine Learning*.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024b. Depth anything: Unleashing the power of large-scale unlabeled data. In *Conference on Computer Vision and Pattern Recognition*.

Yu, J.; Wang, Y.; Zhao, C.; Ghanem, B.; and Zhang, J. 2023. FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model. In *International Conference on Computer Vision*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *International Conference on Computer Vision*.

Zhao, S.; Chen, D.; Chen, Y.-C.; Bao, J.; Hao, S.; Yuan, L.; and Wong, K.-Y. K. 2024. Uni-ControlNet: All-in-One Control to Text-to-Image Diffusion Models. In *Advances in Neural Information Processing Systems*.

Zhu, Y.; Zhang, K.; Liang, J.; Cao, J.; Wen, B.; Timofte, R.; and Gool, L. V. 2023. Denoising Diffusion Models for Plug-and-Play Image Restoration. In *Conference on Computer Vision and Pattern Recognition Workshops*.