LABridge: Text–Image Latent Alignment Framework via Mean-Conditioned OU Process

Huiyang Shao[1,2] **Xin Xia**[2,*] **Yuxi Ren**[2] **Xing Wang**[2] **Xuefeng Xiao**[2,†]

[1]Tsinghua University [2]ByteDance Seed

Abstract

Diffusion models have emerged as state-of-the-art in image synthesis.However, it often suffer from semantic instability and slow iterative denoising. We introduce Latent Alignment Framework (LABridge), a novel Text–Image Latent Alignment Framework via an Ornstein–Uhlenbeck (OU) Process, which explicitly preserves and aligns textual and visual semantics in an aligned latent space. LABridge employs a Text-Image Alignment Encoder (TIAE) to encode text prompts into structured priors that are directly aligned with image latents. Instead of a homogeneous Gaussian, Mean-Conditioned OU process smoothly interpolates between these text-conditioned priors and image latents, improving stability and reducing the number of denoising steps. Extensive experiments on standard text-to-image benchmarks show that LABridge achieves better text–image alignment metric and competitive FID scores compared to leading diffusion baselines. By unifying text and image representations through principled latent alignment, LABridge paves the way for more efficient, semantically consistent, and high-fidelity text to image generation.

1 Introduction

Diffusion models [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020, Song et al., 2021] represent a significant advancement in generative modeling, demonstrating state-of-the-art performance in diverse tasks such as text generation [Li et al., 2022, Wu et al., 2023], high-fidelity image synthesis [Rombach et al., 2022, Ramesh et al., 2022, Ho et al., 2022, Shao et al., 2023, Lin et al., 2025], image restoration [Blattmann et al., 2023, Brooks et al., 2024], and 3D content creation [Liu et al., 2023, Evans et al., 2024]. These models typically operate by defining a forward diffusion process that gradually adds noise to data, transforming it into a simple prior distribution (often Gaussian), and then learning a reverse process to generate data by iteratively denoising samples drawn from this prior. The mathematical foundation often relies on stochastic differential equations (SDEs) [Song et al., 2021] or discrete-time Markov chains, enabling powerful sampling and manipulation strategies like accelerated generation [Mei et al., 2024], timestep analysis, and model distillation [Shao et al., 2025, Xie et al., 2024, Nguyen et al., 2024, Kang et al., 2024].

Existing Challenges

Despite their success, standard diffusion models face limitations, particularly in text-to-image generation, stemming largely from their reliance on a fixed, often unstructured prior:

1. **Instability and Ambiguity from Homogeneous Priors:** Conventional methods map the entire diverse data distribution $q_{\text{data}}(\boldsymbol{x})$ to a single, fixed prior, typically $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. This forces distinct semantic concepts (e.g., "cat," "dog," "landscape") onto the same simple latent structure. This collapsing of priors can lead to instability, as the reverse process must disentangle these varied

*Corresponding Author † Project Leader

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-1719

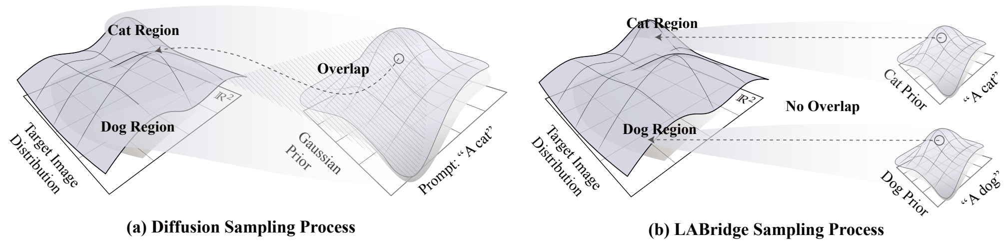

Figure 1: Comparision of sampling process between traditional diffusion and LABridge

semantics from a homogeneous starting point. Furthermore, it can create ambiguity in score estimation ($\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$), as paths originating from different initial data points might overlap significantly in the latent space near $t = T$, making the learned score an average that lacks precision for any specific semantic direction. We provide analysis in Appendix E.3

2. **Inefficient Sampling:** As shown in Fig. 1 (a), the diffusion sampling process follows a curved path. Without strong guidance, especially in early steps, the process can be slow, requiring many iterations (NFE - Number of Function Evaluations) to converge to a high-fidelity image that accurately reflects the conditioning. We provide analysis in Appendix D.3
3. **Weak Text-Vision Alignment:** While conditioning mechanisms inject textual information, the fundamental diffusion process still operates between the image manifold and Gaussian prior. This indirect connection can limit the precise alignment between the generated image and complex or nuanced text prompts, especially for out-of-distribution concepts (analysis in Appendix D.2).

Our Innovations: LABridge

To overcome these challenges, we introduce LABridge, a framework designed to enhance text-vision alignment and accelerate sampling in diffusion models. LABridge leverages two core ideas: a dedicated encoder TIAE to create structured, text-conditioned priors, and an OU diffusion process to connect image latents directly to these priors.

1. **Text Encoder for Structured Priors:** We employ a encoder to process text prompts (y) and generate corresponding latent representations $\boldsymbol{\mu}_T(y)$. The latent $\boldsymbol{\mu}_T(y)$ acts as a structured, text-specific prior mean in the aligned latent space of an image autoencoder. The TIAE ensures that semantically similar texts map to nearby priors, preserving semantic structure.
2. **OU Diffusion Process for Alignment and Stability:** We adopt OU process explicitly models the stochastic path between the image latent $\boldsymbol{x}_0$ (obtained from a VAE encoder) and a distribution centered around the text-specific prior $\boldsymbol{\mu}_T(y)$, i.e., $\mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2\boldsymbol{I})$. The OU process, known for its mean-reverting property, naturally pulls the state towards the target mean $\boldsymbol{\mu}_T(y)$, inherently promoting stability and alignment.

LABridge offers a new perspective: it frames text-to-image generation as learning a stochastic process between the image latent manifold and a manifold of text-conditioned priors. This explicit alignment via the process mechanism enhances semantic consistency and, by providing a better starting point and direction, significantly speeds up the sampling process.

In summary, our contributions are threefold:

- We propose LABridge, a novel framework utilizing an encoder TIAE to generate structured text-conditioned priors $\boldsymbol{\mu}_T(y)$ and an OU diffusion process to align them with image latents $\boldsymbol{x}_0$.
- We demonstrate that the OU process mechanism, combined with structured priors, improves text-vision alignment and inherently accelerates sampling by providing a more certain trajectory, supported by theoretical analysis.
- We validate LABridge experimentally, showing significant improvements in sampling and text-image consistency metrics while maintaining competitive image fidelity compared to baselines.

2 Preliminaries

In this section, we introduce the essential background concepts. Additional technical details, and extended definitions are deferred to Appendix A.

2.1 Diffusion Processes and Score Matching

Diffusion Process. Let $\boldsymbol{x} \in \mathbb{R}^d$ follow the data distribution $q_{\text{data}}(\boldsymbol{x})$. A forward diffusion process defines a sequence of latent variables $\{\boldsymbol{x}_t\}_{t\in[0,T]}$ starting from $\boldsymbol{x}_0 \sim q_{\text{data}}(\boldsymbol{x})$ and evolving towards a simple prior distribution $p_T(\boldsymbol{x}_T) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ as t goes from 0 to T. This evolution is often described by a stochastic differential equation (SDE) [Song et al., 2021]:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)\, dt + g(t)\, d\boldsymbol{w}_t, \tag{1}$$

where $\boldsymbol{f}(\cdot, t)$ is the drift function, $g(t)$ is the diffusion coefficient, and $\boldsymbol{w}_t$ is a standard Wiener process. Generating new data involves reversing this process. The corresponding reverse-time SDE is given by [Anderson, 1982]:

$$d\boldsymbol{x}_t = \left[\boldsymbol{f}(\boldsymbol{x}_t, t) - g(t)^2\, \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] dt + g(t)\, d\bar{\boldsymbol{w}}_t, \tag{2}$$

where $\bar{\boldsymbol{w}}_t$ is a Wiener process running backward in time, and $p_t(\boldsymbol{x}_t)$ is the marginal probability density of $\boldsymbol{x}_t$. The crucial term is the score function $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$.

Score Matching. In practice, the true score $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ is unknown and is approximated by a time-dependent neural network $\boldsymbol{s}_\theta(\boldsymbol{x}_t, t)$, often conditioned on additional information y (like text embeddings), denoted $\boldsymbol{s}_\theta(\boldsymbol{x}_t, t, y)$. This network is trained by minimizing a score matching objective [Hyvärinen and Dayan, 2005, Vincent, 2011]. For many diffusion processes (like VP and VE), the conditional score $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ is tractable. A common training objective is:

$$\mathcal{L}_{\text{SM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,T), \boldsymbol{x}_0\sim q_{\text{data}}, \boldsymbol{\epsilon}\sim\mathcal{N}(0,I)} \left[\lambda(t) \left\| \boldsymbol{s}_\theta(\alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, t, y) - \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) \right\|^2\right], \tag{3}$$

where $p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I})$ defines the transition kernel, and $\lambda(t)$ is a weighting function. For this Gaussian kernel, $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = -(\boldsymbol{x}_t - \alpha_t \boldsymbol{x}_0)/\sigma_t^2 = -\boldsymbol{\epsilon}/\sigma_t$. This leads to the widely used noise prediction objective:

$$\mathcal{L}_{\text{denoise}}(\theta) = \mathbb{E}_{t, \boldsymbol{x}_0, \boldsymbol{\epsilon}} \left[\lambda'(t) \left\| \boldsymbol{\epsilon}_\theta(\alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}, t, y) - \boldsymbol{\epsilon} \right\|^2\right], \tag{4}$$

where $\boldsymbol{\epsilon}_\theta$ is the network predicting the noise $\boldsymbol{\epsilon}$, related to the score network by $\boldsymbol{s}_\theta(\boldsymbol{x}_t, t, y) = -\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, y)/\sigma_t$.

2.2 Diffusion Bridge Models

While standard diffusion maps data to a fixed prior, a diffusion bridge connects two specified endpoint distributions, $p_0(\boldsymbol{x}_0)$ and $p_T(\boldsymbol{x}_T)$, which can both be complex. This is particularly relevant for tasks involving paired data $(\boldsymbol{x}_0, \boldsymbol{x}_T) \sim q_{\text{data}}(\boldsymbol{x}_0, \boldsymbol{x}_T)$, such as image translation or, in our case, aligning image latents $\boldsymbol{x}_0$ with text-derived priors $\boldsymbol{\mu}_T$.

Stochastic Bridges via h-Transform. Given a forward SDE Eq. (1), Doob's h-transform provides a way to condition the process to start at $\boldsymbol{x}_0$ at $t = 0$ and end exactly at $\boldsymbol{x}_T = \boldsymbol{y}$ at $t = T$. The resulting bridge SDE is:

$$d\boldsymbol{x}_t = \left[\boldsymbol{f}(\boldsymbol{x}_t, t) + g(t)^2\, \nabla_{\boldsymbol{x}_t} \log p_{T|t}(\boldsymbol{y} \mid \boldsymbol{x}_t)\right] dt + g(t)\, d\boldsymbol{w}_t, \tag{5}$$

where $p_{T|t}(\boldsymbol{x}_T \mid \boldsymbol{x}_t)$ is the transition probability density of the original SDE from time t to T. The term $\boldsymbol{h}(\boldsymbol{x}_t, t, \boldsymbol{y}, T) = \nabla_{\boldsymbol{x}_t} \log p_{T|t}(\boldsymbol{y} \mid \boldsymbol{x}_t)$ is the "guidance" term ensuring the endpoint constraint. For linear SDEs with Gaussian transitions (like VP, VE, OU), this term is often tractable.

Denoising Diffusion Bridge Models [Zhou et al., 2023]. Instead of exact endpoints, we often want to sample from a conditional distribution $q(\boldsymbol{x}_0 \mid \boldsymbol{x}_T)$ given paired data $(\boldsymbol{x}_0, \boldsymbol{x}_T) \sim q_{\text{data}}$. This can be achieved by reversing a bridge process designed such that its marginals $q(\boldsymbol{x}_0, \boldsymbol{x}_T)$ approximate $q_{\text{data}}(\boldsymbol{x}_0, \boldsymbol{x}_T)$. Zhou et al. [2023] show that the reverse SDE for sampling $\boldsymbol{x}_t$ given $\boldsymbol{x}_T = \boldsymbol{y}$ is:

$$d\boldsymbol{x}_t = \left[\boldsymbol{f}(\boldsymbol{x}_t, t) - g^2(t) \left(\nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_T = \boldsymbol{y}) - \nabla_{\boldsymbol{x}_t} \log p_{T|t}(\boldsymbol{y} \mid \boldsymbol{x}_t)\right)\right] dt + g(t)\, d\bar{\boldsymbol{w}}_t, \tag{6}$$

where $\nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t \mid \boldsymbol{x}_T = \boldsymbol{y})$ is the score of the conditional bridge distribution.

https://doi.org/10.52202/085713-1719

2.3 Mean-Conditioned Ornstein-Uhlenbeck Process

Ornstein-Uhlenbeck (OU) Process. The OU process is a mean-reverting stochastic process often used to model systems returning to equilibrium. Its SDE is:

$$d\boldsymbol{x}_t = \theta\,(\boldsymbol{\mu} - \boldsymbol{x}_t)\,dt + \sigma\,d\boldsymbol{w}_t, \tag{7}$$

where $\theta > 0$ is the rate of mean reversion, $\boldsymbol{\mu}$ is the equilibrium mean, and σ is the volatility. The drift term $\theta(\boldsymbol{\mu} - \boldsymbol{x}_t)$ pulls the state $\boldsymbol{x}_t$ towards $\boldsymbol{\mu}$.

Ornstein-Uhlenbeck Bridge (OUB). An OU process is a conditioned OU process that starts at $\boldsymbol{x}_0$ at $t = 0$ and ends at $\boldsymbol{x}_T = \boldsymbol{y}$ at $t = T$. Its SDE can be derived using Doob's h-transform. For the standard OU process with $\sigma = 1$ starting from $\boldsymbol{x}_0$ at $t = 0$, the transition density to time T is Gaussian: $p_{T|0}(\boldsymbol{x}_T \mid \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{\mu} + (\boldsymbol{x}_0 - \boldsymbol{\mu})e^{-\theta T}, \frac{\sigma^2}{2\theta}(1 - e^{-2\theta T})\boldsymbol{I})$. The OUB derived from this inherits the mean-reverting property but ensures the endpoint constraint. In our work, we use a specific form of OU process connecting $\boldsymbol{x}_0$ to a distribution around $\boldsymbol{\mu}_T(y)$, defined by the transition:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0, \boldsymbol{\mu}_T) = \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{x}_0 e^{-\theta t} + \boldsymbol{\mu}_T(1 - e^{-\theta t}), \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})\boldsymbol{I}\right). \tag{8}$$

This corresponds to an OU process SDE $d\boldsymbol{x}_t = \theta(\boldsymbol{\mu}_T - \boldsymbol{x}_t)dt + \sigma d\boldsymbol{w}_t$. This structure is key to LABridge. The parameter θ controls the strength of reversion towards the text-conditioned prior mean $\boldsymbol{\mu}_T$. OUB processes can encompass VP and VE under specific parameter choices.

3 Proposed Method: LABridge

3.1 Motivation: Structured Priors and Directed Diffusion

A primary challenge lies in **directly using text as latent representations**. Text data is fundamentally different from typical latent variables. It is inherently **infinite and unstructured**; the space of possible sentences and meanings is vast and does not easily map to predefined, discrete categories or a simple, low-dimensional manifold. Converting raw text directly into a structured latent code suitable for generative models is non-trivial.

Instead, we must first represent text in a suitable format, typically through powerful pretrained embedding models (like CLIP, T5) that capture semantic meaning in high-dimensional vectors (E_y). However, simply using these embeddings as conditioning signals for a standard diffusion process (mapping image I to noise $\boldsymbol{\epsilon}$) still relies on the model implicitly learning the complex relationship between the text embedding space and the image manifold during the denoising process initiated from a generic prior.

This motivates the need to **represent text in a continuous latent space that is explicitly aligned with the latent space of images**. Rather than mapping images to generic noise, we propose mapping images $\boldsymbol{x}_0$ to text-conditioned priors $\boldsymbol{\mu}_T(y)$ that live in the similar latent space. This requires:

1. A mechanism to map text embeddings E_y to target latent priors $\boldsymbol{\mu}_T(y)$.
2. Ensuring these priors $\boldsymbol{\mu}_T(y)$ are semantically consistent (similar texts map to nearby priors) and aligned with corresponding image latents.
3. A generative process that efficiently connects these aligned endpoint distributions.

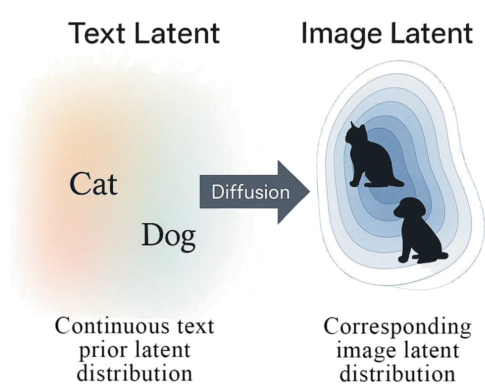

Figure 2: Conceptual illustration: Alignment and Continuity of Text-Image Latent Spaces

https://doi.org/10.52202/085713-1719

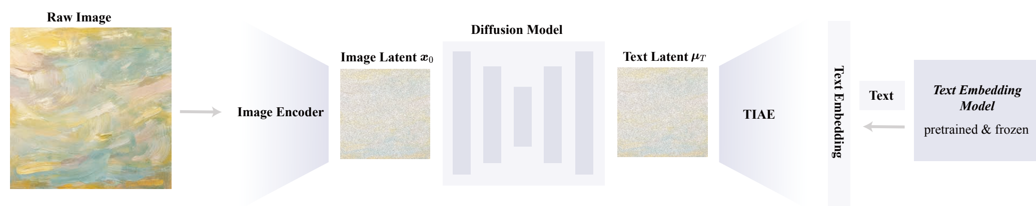

Figure 3: Overview of the LABridge. An image I is encoded to $\boldsymbol{x}_0$ by $\mathcal{E}_{\text{VAE}}$. The corresponding text y is embedded to E_y by $\mathcal{E}_{\text{Emb}}$ and then mapped to a target prior mean $\boldsymbol{\mu}_T(y)$ by the Text-Vision Alignment Encoder $\mathcal{E}_{\text{TE}}$ (TIAE). An OU diffusion process learns the stochastic path between $\boldsymbol{x}_0$ and the distribution $\mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2\boldsymbol{I})$. During inference, sampling starts from $\boldsymbol{x}_T$ and follows the learned reverse process dynamics towards $\boldsymbol{x}_0$, which is then decoded to an image $\hat{I}$ by $\mathcal{D}_{\text{VAE}}$.

3.2 LABridge Framework Overview

As depicted in Fig. 3, LABridge integrates three key components:

1. **Pretrained Image Autoencoder:** We utilize a frozen Variational Autoencoder (VAE) with encoder $\mathcal{E}_{\text{VAE}}$ and decoder $\mathcal{D}_{\text{VAE}}$. The encoder maps an input image I to a latent representation $\boldsymbol{x}_0 = \mathcal{E}_{\text{VAE}}(I)$ in $\mathbb{R}^d$, effectively moving the diffusion process to a compressed latent space. The decoder reconstructs the image $\hat{I} = \mathcal{D}_{\text{VAE}}(\boldsymbol{x}_0)$.
2. **Text-Vision Alignment Encoder (TIAE):** This novel component, denoted $\mathcal{E}_{\text{TE}}$, is responsible for bridging the semantic gap between text and vision. It takes text embeddings E_y (obtained from a frozen pretrained text model $\mathcal{E}_{\text{Emb}}$, e.g., CLIP [Radford et al., 2021]) as input and outputs a target prior mean $\boldsymbol{\mu}_T(y) = \mathcal{E}_{\text{TE}}(E_y) \in \mathbb{R}^d$. The TIAE is specifically trained to ensure that $\boldsymbol{\mu}_T(y)$ is both semantically meaningful (reflecting the content of y) and aligned with the corresponding image latents $\boldsymbol{x}_0$ in the VAE's latent space.
3. **Mean-Conditioned OU Diffusion Process:** Instead of a standard diffusion process mapping to $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, we employ an OU diffusion process [Zhou et al., 2023] to model the stochastic transition between the image latent $\boldsymbol{x}_0$ and the text-conditioned prior distribution $\mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2\boldsymbol{I})$. The process is parameterized by a score network $\boldsymbol{s}_\theta(\boldsymbol{x}_t, t, y)$ or, equivalently, a noise prediction network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, y)$. The OU process inherently incorporates mean reversion towards $\boldsymbol{\mu}_T(y)$, promoting stability and directed sampling.

3.3 TIAE Architecture and Training Objective

Architecture. To ensure seamless integration and potentially leverage existing performant architectures, we adopt a structure similar to modern diffusion models for the TIAE $\mathcal{E}_{\text{TE}}$, specifically using blocks inspired by DiT [Peebles and Xie, 2023] which handle conditional inputs effectively. It takes the text embedding E_y as input and outputs the prior mean $\boldsymbol{\mu}_T(y)$.

Training Objective. The TIAE is trained to produce priors $\boldsymbol{\mu}_T(y)$ that are: (a) aligned with corresponding image latents $\boldsymbol{x}_0$, and (b) preserve the semantic structure of the text embeddings E_y. We use a composite loss:

(a) **Latent Alignment Loss ($\mathcal{L}_{\text{align}}$):** Encourages the TIAE output $\boldsymbol{\mu}_T(y)$ to be close to the VAE latent $\boldsymbol{x}_0$ for corresponding image-text pairs (I, y).

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(\boldsymbol{x}_0, y)\sim q_{\text{data}}} \left[\|\boldsymbol{\mu}_T(y) - \boldsymbol{x}_0\|_2^2\right], \tag{9}$$

where $\boldsymbol{x}_0 = \mathcal{E}_{\text{VAE}}(I)$ and $\boldsymbol{\mu}_T(y) = \mathcal{E}_{\text{TE}}(E_y)$. This loss is visualized in Fig. 4.

(b) **Semantic Consistency Loss ($\mathcal{L}_{\text{sem}}$):** Ensures that the distances between latent priors reflect the distances between text embeddings (using cosine similarity).

$$\mathcal{L}_{\text{sem}} = \mathbb{E}_{y_i, y_j} \left[\left(\text{sim}_{\cos}(\boldsymbol{\mu}_T(y_i), \boldsymbol{\mu}_T(y_j)) - \text{sim}_{\cos}(E_{y_i}, E_{y_j})\right)^2\right]. \tag{10}$$

This encourages the structure of the text embedding space to be preserved in the prior space, as illustrated conceptually in Fig. 5.

 https://doi.org/10.52202/085713-1719

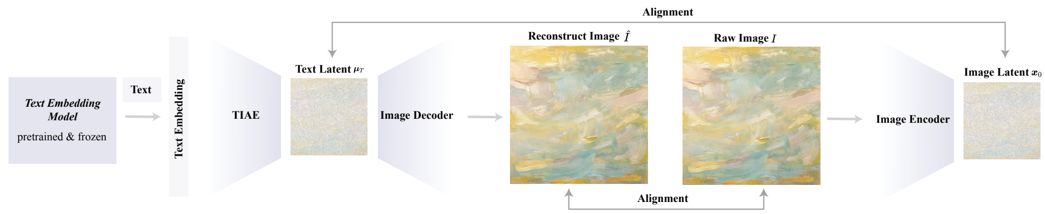

Figure 4: An overview of the TIAE training process. The input text is embedded using a pretrained model and then processed by our TIAE to produce a latent representation $\boldsymbol{\mu}_T$. The alignment loss encourages this $\boldsymbol{\mu}_T$ to be close to the image latent $\boldsymbol{x}_0$ obtained from the corresponding raw image I via the Image Encoder. An Image Decoder can reconstruct $\hat{I}$ from $\boldsymbol{\mu}_T$ (for $\mathcal{L}_{rec}$) or from $\boldsymbol{x}_0$.

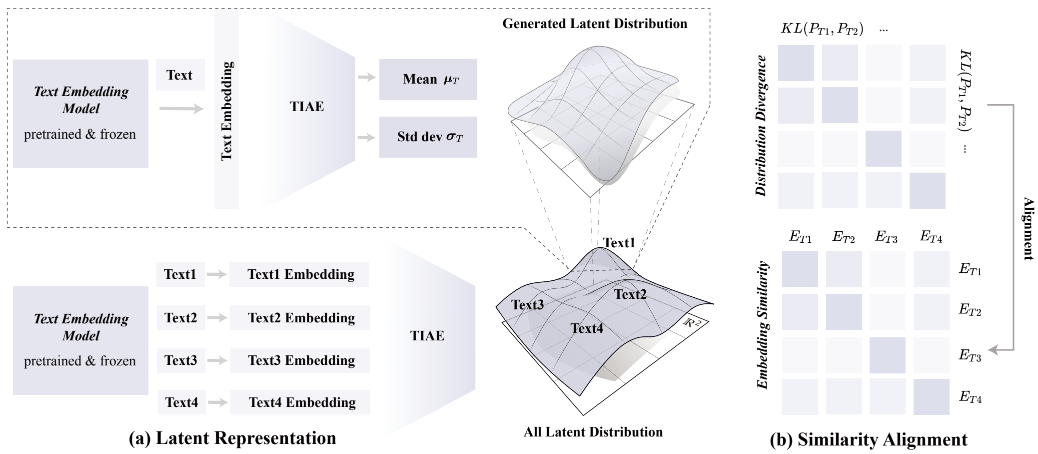

Figure 5: Illustration of the latent space alignment and distribution analysis for TIAE. Text embeddings (left) are encoded into latent distributions centered at $\boldsymbol{\mu}_T$ (center). The semantic consistency loss $\mathcal{L}_{\text{sem}}$ aims to ensure that similar text inputs (e.g., Text1, Text2) produce close latent means/distributions, preserving semantic proximity, as shown by the embedding similarity matrix (right) compared to the latent distribution similarity.

(c) **Reconstruction Loss ($\mathcal{L}_{\text{rec}}$):** To ensure $\boldsymbol{\mu}_T(y)$ can be decoded into meaningful images, we can add a reconstruction term comparing the decoded prior to the original image.

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{(\boldsymbol{x}_0, y) \sim q_{\text{data}}} \left[\| \mathcal{D}_{\text{VAE}}(\boldsymbol{\mu}_T(y)) - I \|_{\text{p}}^{p} \right], \tag{11}$$

The total TIAE loss is $\mathcal{L}_{\text{TIAE}} = w_a \mathcal{L}_{\text{align}} + w_s \mathcal{L}_{\text{sem}} + w_r \mathcal{L}_{\text{rec}}$, with weights w_a, w_s, w_r.

3.4 Integration with OU Diffusion Process

Once the TIAE $\mathcal{E}_{\text{TE}}$ is trained and frozen, we train the diffusion model component. LABridge employs an OU diffusion process defined by the forward stochastic differential equation (SDE):

$$d\boldsymbol{x}_t = \theta(\boldsymbol{\mu}_T(y) - \boldsymbol{x}_t)dt + \sigma d\boldsymbol{w}_t, \quad t \in [0, T], \tag{12}$$

where $\boldsymbol{\mu}_T(y) = \mathcal{E}_{\text{TE}}(\mathcal{E}_{\text{Emb}}(y))$ is the target mean obtained from the frozen TIAE for a given text prompt y. This SDE defines a process that starts near $\boldsymbol{x}_0$ at $t = 0$ and is drawn towards $\boldsymbol{\mu}_T(y)$ as $t \to \infty$. The transition kernel $q(\boldsymbol{x}_t|\boldsymbol{x}_0, y)$ corresponding to this SDE, assuming $\boldsymbol{x}_t$ depends on $\boldsymbol{x}_0$ and the target mean $\boldsymbol{\mu}_T(y)$.

The goal is to learn the reverse process to sample $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0|y)$ starting from the prior $p(\boldsymbol{x}_T|y) = \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{\mu}_T(y), \sigma_T^2 \boldsymbol{I})$. The reverse process is governed by the score function $\nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t|y)$. We follow the standard practice in diffusion models and train a neural network $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, y)$ to predict the noise $\boldsymbol{\epsilon}$ that generated $\boldsymbol{x}_t$ from $\boldsymbol{x}_0$ and $\boldsymbol{\mu}_T(y)$ via Eq. 8. The noise prediction objective for the OU process is:

$$\mathcal{L}_{\text{Bridge}} = \mathbb{E}_{t \sim \mathcal{U}(0,T), (\boldsymbol{x}_0, y) \sim q_{\text{data}}, \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \left[w'(t) \left\| \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, y) - \boldsymbol{\epsilon} \right\|_2^2 \right], \tag{13}$$

https://doi.org/10.52202/085713-1719

where $w'(t)$ is a time-dependent weighting function. The reverse probability flow ODE used for sampling is:

$$d\boldsymbol{x}_t = [\theta(\boldsymbol{\mu}_T(y) - \boldsymbol{x}_t) + \frac{\sigma^2}{2\sigma_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, y)]dt. \tag{14}$$

3.5 Training and Inference Algorithms

The overall training procedure for LABridge consists of two sequential stages (details of training procedure are provided in Algo. 1 in Appendix B.):

- **Stage 1: TIAE Training.** Train the Text-Vision Alignment Encoder $\mathcal{E}_{\text{TE}}$ using the composite loss $\mathcal{L}_{\text{TIAE}}$ on paired image-text data, with the VAE and text embedder frozen.
- **Stage 2: OU Process Training.** Freeze the trained $\mathcal{E}_{\text{TE}}$. Train the noise prediction network $\boldsymbol{\epsilon}_\theta$ for the OU process using the denoising objective $\mathcal{L}_{\text{Bridge}}$ (Eq. 13).

The overall inference procedure for LABridge mainly based on Eq. 14 (details of inference procedure are provided in Algo. 2 in Appendix B.).

4 Theoretical Guarantees

We provide theoretical justification for the advantages of LABridge, highlighting improvements in text-vision alignment, sampling efficiency and stability standard diffusion models using fixed priors. Detailed statements and proofs are provided in Sec. D and E.

Theorem 4.1 (Enhanced Text-Vision Alignment)**.** *The TIAE training objective, particularly $\mathcal{L}_{align}$ and $\mathcal{L}_{sem}$, explicitly optimizes the prior mean $\boldsymbol{\mu}_T(y)$ to be (a) close to the corresponding image latent mean $\mathbb{E}[\boldsymbol{x}_0|y]$ and (b) preserve the semantic structure of the text embeddings E_y. The OU process formulation reinforces this alignment during diffusion training and generation. (Ref: Thm. D.3, Prop. D.4 in the appendix).*

Theorem 4.2 (Sampling Acceleration via Informed Initialization and Dynamics)**.** *LABridge accelerates sampling due to two factors:*

(i) ***Reduced Initial Error:*** *Starting the reverse process from $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2\boldsymbol{I})$ provides an initial state closer (in expectation) to the target conditional mean $\mathbb{E}[\boldsymbol{x}_0|y]$ compared to starting from $\mathcal{N}(\boldsymbol{0}, \sigma_T^2\boldsymbol{I})$. (Ref: Thm. D.5 in appendix).*

(ii) ***Directed Drift:*** *The OU reverse dynamics (Eq. 14) include an explicit mean reversion term $\theta(\boldsymbol{\mu}_T(y) - \boldsymbol{x}_t)$ which provides additional drift towards the text-aligned prior mean $\boldsymbol{\mu}_T(y)$, supplementing the learned score/noise term and offering stronger guidance than standard diffusion drifts . (Ref: Thm. D.6 in appendix).*

Theorem 4.3 (Improved Sampling Stability)**.** *The inherent mean-reverting property of the OU process drift term $\theta(\boldsymbol{\mu}_T(y) - \boldsymbol{x}_t)$ enhances the stability of the reverse sampling process, making it less prone to divergence compared to processes with zero or origin-centric drift, especially when score estimates may be imperfect. (Ref: Thm. D.8 in appendix).*

Theorem 4.4 (Tighter Evidence Lower Bound)**.** *Using the text-conditioned prior $p(\boldsymbol{x}_T|y) = \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2\boldsymbol{I})$ results in a smaller expected KL divergence between the forward process endpoint distribution $q(\boldsymbol{x}_T|\boldsymbol{x}_0, y)$ and the prior, compared to using a fixed prior $p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{0}, \sigma_T^2\boldsymbol{I})$, provided $\mathbb{E}[\boldsymbol{x}_0|y]$ varies significantly with y and $\boldsymbol{\mu}_T(y)$ approximates it well. This leads to a tighter ELBO. (Ref: Thm. D.9).*

5 Experiment

We conducted a series of experiments to verify the effectiveness of LABridge under various settings.

Experiment Setup All code was performed on 8 A100 GPUs machine. For the first part, we employ DiT-XL/2 model to learn from scratch on benchmark dataset. We adopt `NV-Embed-v2` [1] as pretrained

[1]https://huggingface.co/nvidia/NV-Embed-v2

 https://doi.org/10.52202/085713-1719

text embedding model, which output 4096 ($1 \times 64 \times 64$) vector. The training utilized `AdamW` optimizer with a learning rate of $1e-5$, $\beta_1 = \beta_2 = 0.9$, weight decay of 0.03, batch size of 16, and run for 200 epochs. All images are preprocessed with center crop and resized (1024×1024). We tune the weighting w_a, w_s and w_r in the range of $[0, 1]$. After a brief sweep, we used $w_a = 1.0$, $w_s = 0.5$, and $w_r = 0.2$.

Dataset. The training data comprises a selection from COYO [Byeon et al., 2022] datasets, following selection criteria in [Lin et al., 2024, Ren et al., 2024]. We evaluated performance on standard benchmarks: COCO [Lin et al., 2014], ImageNet [Deng et al., 2009], MJHQ-30K [Li et al., 2024a].

Evaluation Metrics. We evaluate all models on the COCO validation set [Lin et al., 2014], using two primary metrics: FID [Heusel et al., 2017] and CLIP score [Radford et al., 2021, Hessel et al., 2021]. Specifically, we report FID-10K, where prompts are randomly sampled from the validation set. The generated images for these prompts are then compared to reference images from validation set. In addition to these, we evaluate our models on the GenEval [Ghosh et al., 2024] and DPG-Bench [Hu et al., 2024] benchmarks, both of which are designed to measure text-image alignment. Finally, we report Inception Score [Salimans et al., 2016] and Precision/Recall [Kynkäänniemi et al., 2019] as secondary metrics to provide a more comprehensive evaluation.

Base Models. We implement our framework on five existing popular base models: stable-diffusion-v1-5 (SD15) [Rombach et al., 2022], stable-diffusion-xl-v1.0-base (SDXL) [Rombach et al., 2022] with UNet architecture, and SD-3.5 [Esser et al., 2024] with MM-DiT architecture, PixArt-α [Chen et al., 2023] with DiT architecture.

	w/o CFG				w CFG			
Model	**FID ↓**	**IS ↑**	**Pre. ↑**	**Rec. ↑**	**FID ↓**	**IS ↑**	**Pre. ↑**	**Rec. ↑**
GIVT [Tschannen et al., 2025]	5.67	-	0.75	0.59	3.35	-	0.84	0.53
MAR-B [Li et al., 2024b]	3.48	192.4	0.78	0.58	2.31	281.7	0.82	0.57
LDM-4 [Rombach et al., 2022]	10.56	103.5	0.71	0.62	3.60	247.7	**0.87**	0.48
CausalFusion-L [Deng et al., 2024]	10.56	103.5	0.71	0.62	1.94	264.4	0.82	0.59
ADM [Dhariwal and Nichol, 2021]	10.94	-	0.69	0.63	4.59	186.7	0.82	0.52
DiT-XL [Peebles and Xie, 2023]	9.62	121.5	0.67	<u>0.67</u>	2.27	278.2	0.83	0.57
SiT-XL [Ma et al., 2024]	8.3	-	-	-	2.06	270.3	0.82	0.59
ViT-XL [Hang et al., 2023]	8.10	-	-	-	2.06	-	-	-
U-ViT-H/2 [Bao et al., 2023]	6.58	-	-	-	2.29	263.9	0.82	0.57
MaskDiT [Zheng et al., 2023]	5.69	178.0	0.74	0.60	2.28	276.6	0.80	**0.61**
RDM [Teng et al., 2023]	5.27	153.4	0.75	0.62	1.99	260.4	0.81	0.58
CausalFusion-XL [Deng et al., 2024]	<u>3.61</u>	<u>180.9</u>	0.75	0.66	<u>1.77</u>	282.3	0.82	**0.61**
DiT-XL/2 + LABridge (VE)	5.74	152.1	0.70	0.64	2.13	279.9	0.84	0.57
DiT-XL/2 + LABridge (VP)	4.95	164.4	<u>0.76</u>	<u>0.67</u>	1.95	285.4	0.86	0.60
DiT-XL/2 + LABridge (OU)	3.83	179.2	0.76	0.65	1.84	<u>289.3</u>	0.87	<u>0.62</u>
CausalFusion-XL + LABridge (VP)	**3.49**	**182.4**	**0.77**	**0.68**	**1.69**	**291.3**	<u>0.86</u>	**0.63**

Table 1: **Benchmarking class-conditional image generation on ImageNet 256×256.** The left half is without CFG, the right half is with CFG. Best results are in **bold**, second best are <u>underlined</u>.

Learning from Scratch Experiments. We evaluated our method (three bridges) on the ImageNet 256×256 dataset with and without classifier-free guidance (CFG) and compared it with various state-of-the-art approaches (ImageNet 512x512 dataset results in Appendix C). The results are summarized in Tab. 2. **Without CFG**: Our method, CausalFusion-XL + LABridge, achieves the best performance among all evaluated models. Compared to the DiT-XL method, DiT-XL/2 + LABridge achieves significant performance improvements across all metrics. This demonstrates the effectiveness of our proposed LABridge in improving the quality and diversity of generated images. **With CFG**: When CFG is applied, our method again outperforms others, achieving the best FID and the highest IS, further highlighting its ability to generate high-quality and consistent text-conditioned images. These results validate that our approach is generalizable across different backbones and significantly enhances both image quality and text-image alignment.

Fine-Tuning Existing Pretrained Models. We conducted extensive experiments with various pretrained models with similar baseline Liu et al. [2024]. The experimental results, summarized in Tab. 2 and Fig. 6 (Appendix C.3), demonstrate that our method consistently outperforms baseline approaches across multiple evaluation metrics and datasets. Notably, our approach excels on the GenEval and DPG benchmarks, which are specifically designed to measure text-image consistency.

https://doi.org/10.52202/085713-1719

Method	COCO-10K FID ↓	COCO-10K CLIP ↑	MJHQ-30K FID ↓	MJHQ-30K CLIP ↑	Text-Alignment GenEval ↑	Text-Alignment DPG ↑
Stable Diffusion V1.5 Comparision						
SD15-Base [Rombach et al., 2022]	15.81±.04	28.03±.08	13.54±.03	28.40±.01	0.48±.04	70.64±.01
SD15 + CF Liu et al. [2024]	14.83±.05	28.52±.07	12.62±.02	28.90±.03	0.54±.04	71.35±.01
SD15 + LABridge (VP, UNet)	**13.82±.04**	**29.01±.06**	**11.63±.03**	**29.42±.02**	**0.57±.04**	**72.42±.02**
Stable Diffusion XL Comparision						
SDXL-Base [Rombach et al., 2022]	**11.68±.04**	28.83±.04	10.55±.01	29.63±.01	0.56±.03	75.52±.02
SDXL + CF (Liu et al. [2024])	12.69±.04	29.33±.03	9.59±.02	30.15±.03	0.62±.03	76.53±.02
SDXL + LABridge(VP, UNet)	12.72±.02	**29.82±.01**	**8.55±.03**	**30.63±.01**	**0.65±.03**	**77.21±.01**
Stable Diffusion 3.5 Medium DiT Comparision						
SD3.5-M [Esser et al., 2024]	9.88±.03	29.91±.04	8.45±.01	30.71±.03	0.62±.03	83.31±.02
SD3.5-M + CF Liu et al. [2024]	8.89±.04	30.43±.02	**7.28±.02**	31.23±.03	0.65±.03	84.31±.02
SD3.5-M + LABridge(VP, DiT)	**7.86±.05**	**30.92±.01**	7.44±.03	**31.72±.02**	**0.67±.03**	**85.31±.01**
Stable Diffusion 3.5 Large DiT Comparision						
SD3.5-L [Esser et al., 2024]	7.33±.03	30.88±.03	5.84±.02	31.41±.02	0.66±.03	84.52±.02
SD3.5-L + CF Liu et al. [2024]	6.33±.04	31.36±.04	4.84±.03	31.89±.02	0.68±.03	85.52±.02
SD3.5-L + LABridge(VP, DiT)	**5.34±.03**	**31.87±.02**	**3.82±.02**	**32.39±.01**	**0.69±.03**	**85.28±.01**
PixArt DiT Comparision						
PixArt-α [Chen et al., 2023]	11.24 ±.02	29.52±.03	9.65±.02	30.01±.04	0.49±.03	75.42±.04
PixArt-α + CF Liu et al. [2024]	10.23±.03	30.02±.04	8.62±.03	30.53±.03	0.51±.03	76.65±.04
PixArt-α + LABridge (VP, DiT)	**9.23±.02**	**30.51±.03**	**7.63±.04**	**31.01±.02**	**0.54±.03**	**77.76±.03**

Table 2: Quantitative comparison of state-of-the-art models across various architectures and benchmarks for different metrics.

Figure 6: Qualitative comparison of LABridge incroprate pretrained models against raw models. Please zoom in to check details, lighting, and aesthetic performances. **All methods that do not have an NFE in place default to 50.**

The superior performance on these benchmarks underscores the strong competitiveness of our method in improving alignment between text and generated images. Moreover, to verify the generalizability of TIAE, we tested TIAE in combination with various other community text-to-image plugins. The results are shown in Fig. 7, demonstrating the robustness.

Ablation Studies. We conducted extensive ablation studies to evaluate the effectiveness of the different loss functions used in the LABridge training method, aiming to validate its correctness. The CLIP and FID metrics were tested on COCO-10k and MJHQ-30K, with the results summarized in Tab. 3. These results show that the training process combining all three loss functions consistently delivers the best performance across most backbones and benchmark datasets. In contrast, using a single loss

https://doi.org/10.52202/085713-1719

Module			COCO-10K				MJHQ-30K			
$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{align}}$	CLIP ↑	FID ↓	GenEval ↑	DPG ↑	CLIP ↑	FID ↓	GenEval ↑	DPG ↑
Stable Diffusion V1.5 Ablation										
✓	✓	✓	**29.01** ± .02	**13.82** ± .04	**0.57** ± .02	**72.42** ± .03	**29.42** ± .01	**11.63** ± .05	**0.57** ± .01	**72.42** ± .02
✓		✓	28.90 ± .03	14.01 ± .04	0.55 ± .03	71.85 ± .04	29.25 ± .02	11.90 ± .04	0.55 ± .03	71.54 ± .04
✓	✓		28.56 ± .03	14.32 ± .05	0.54 ± .04	71.35 ± .05	28.97 ± .02	12.04 ± .04	0.54 ± .03	71.35 ± .04
✓			28.12 ± .04	14.85 ± .06	0.53 ± .05	70.65 ± .06	28.45 ± .03	12.38 ± .06	0.53 ± .04	70.65 ± .05
	✓		27.89 ± .05	15.12 ± .07	0.52 ± .06	69.80 ± .07	27.83 ± .04	12.67 ± .05	0.52 ± .05	69.80 ± .06
		✓	27.45 ± .06	15.54 ± .08	0.51 ± .07	69.32 ± .08	27.32 ± .05	13.01 ± .07	0.51 ± .06	69.32 ± .07
Stable Diffusion XL Ablation										
✓	✓	✓	**29.82** ± .03	12.82 ± .02	**0.65** ± .03	**77.21** ± .02	30.63 ± .03	**8.55** ± .02	**0.65** ± .02	**77.21** ± .01
✓		✓	29.61 ± .04	13.09 ± .03	0.63 ± .04	76.75 ± .03	**30.65** ± .04	8.85 ± .03	0.63 ± .04	76.70 ± .03
✓	✓		29.45 ± .04	**12.79** ± .03	0.62 ± .05	76.53 ± .04	30.12 ± .04	8.98 ± .03	0.62 ± .04	76.53 ± .03
✓			29.12 ± .05	13.42 ± .04	0.60 ± .06	75.90 ± .05	29.78 ± .05	9.27 ± .04	0.60 ± .05	75.90 ± .04
	✓		28.78 ± .06	13.83 ± .05	0.58 ± .07	75.30 ± .06	29.32 ± .06	9.64 ± .05	0.58 ± .06	75.30 ± .05
		✓	28.45 ± .07	14.21 ± .06	0.57 ± .08	74.74 ± .07	28.95 ± .07	10.02 ± .06	0.57 ± .07	74.74 ± .06

Table 3: A Evaluation of Performance Following the Ablation Studies of $\mathcal{L}_{\text{rec}}$, $\mathcal{L}_{\text{sem}}$, and $\mathcal{L}_{\text{align}}$ Modules in Stable Diffusion V1.5 and Stable Diffusion XL Models across COCO-10K and MJHQ-30K Datasets.

or a combination of two losses yields slightly inferior results. This highlights the effectiveness of our proposed LABridge approach.

6 Conclusion

We presented LABridge, a novel text-to-image generation framework designed to overcome semantic instability and slow sampling inherent in diffusion models. By employing a TIAE to create structured, text-conditioned prior aligned with image latents, and utilizing an OU diffusion bridge to connect these representations, LABridge establishes explicit text-vision consistency. LABridge offers a robust pathway towards efficient, high-quality conditional image synthesis.

References

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, and Adam et al. Letts. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Chaorui Deng, Deyao Zh, Kunchang Li, Shi Guan, and Haoqi Fan. Causal diffusion transformers for generative modeling. *arXiv preprint arXiv:2412.12095*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, and Frederic et al. Boesel. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *arXiv preprint arXiv:2402.04825*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7441–7451, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, and David J et al. Fleet. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu Ella. Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

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

S Kang et al. Distilling diffusion models into conditional gans. *ECCV*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024a.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024b.

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.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025.

https://doi.org/10.52202/085713-1719

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A framework for cross-modality evolution. *arXiv preprint arXiv:2412.15213*, 2024.

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

S Mei et al. Codi: Conditional diffusion distillation for higher-fidelity and faster image generation. *CVPR*, 2024.

T Nguyen et al. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. *CVPR*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *arXiv preprint arXiv:2404.13686*, 2024.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Huiyang Shao, Qianqian Xu, Peisong Wen, Peifeng Gao, Zhiyong Yang, and Qingming Huang. Building bridge across the time: Disruption and restoration of murals in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20259–20269, 2023.

Huiyang Shao, Xin Xia, Yuhong Yang, Yuxi Ren, Xing Wang, and Xuefeng Xiao. Rayflow: Instance-aware diffusion acceleration via adaptive flow trajectories. *arXiv preprint arXiv:2503.07699*, 2025.

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.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pages 11895–11907, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jiayan Teng, Wendi Zheng, Ming Ding, Wenyi Hong, Jianqiao Wangni, Zhuoyi Yang, and Jie Tang. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*, 2023.

Michael Tschannen, Cian Eastwood, and Fabian Mentzer. Givt: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision*, pages 292–309. Springer, 2025.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Jian Jiao, Juntao Li, Jian Guo, Nan Duan, Weizhu Chen, et al. Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36:39957–39974, 2023.

Sirui Xie, Zhisheng Xiao, Diederik P Kingma, Tingbo Hou, Ying Nian Wu, Kevin Murphy, Tim Salimans, Ben Poole, and Ruiqi Gao. Em distillation for one-step diffusion models. *arXiv preprint arXiv:2405.16852*, 2024.

Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*, 2023.

https://doi.org/10.52202/085713-1719

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize the proposed LABridge framework, its components (TIAE, OU Bridge), the claimed benefits (improved alignment, accelerated sampling), and the intention to provide theoretical backing, which are reflected in the subsequent sections (Method, Theoretical Guarantees, Proof Details).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide limitation in Sec. F.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper includes dedicated sections for Theoretical Analysis (Sec.D) and Proof Details (Sec. E). Assumptions are stated (Assum. D.1), and detailed proofs are provided for the theorems and propositions presented (e.g., Thm. D.3, D.5, D.6, D.9, Propositions D.4, D.7, D.8).

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the algorithms (Algo. 1, 2) and specific details (Sec. C) about the experimental setup, such as datasets used, specific hyperparameters (learning rates, batch sizes, parameters), training duration, evaluation metrics implementation, or baseline implementation details necessary to fully reproduce the claimed experimental results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will provide the source code and data after the draft is completed.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Key details about the experimental setup are provided, including specific hyperparameter values (e.g., for TIAE loss weights w_a, w_s, w_r, OU parameters θ, σ, learning rates, optimizer types, batch sizes), data splits for training/testing on the mentioned benchmarks, and specific schedules used.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We conducted extensive experiments to demonstrate the effectiveness of our method.

8. **Experiments compute resources**

https://doi.org/10.52202/085713-1719

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide information about the computational resources used for the experiments.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper appears to be focused on algorithmic contributions and does not inherently conflict with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work primarily focuses on algorithm design and technical aspects, with limited direct societal impact. As such, the paper does not extensively discuss broader societal implications, positive or negative.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This question is not applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cites prior work for components like pretrained VAEs (Stable Diffusion), text encoders (CLIP), and architectural inspiration (MMDiT).

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Documentation requirements for new assets are not applicable.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The research presented in the paper does not involve crowdsourcing experiments or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

https://doi.org/10.52202/085713-1719

Justification: The research presented does not involve human subjects, so IRB approval is not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Large language models (or derived text encoders like CLIP/T5) are used to obtain text embeddings for conditioning, which is standard practice in text-to-image generation. LLMs are not an important, original, or non-standard component of the core novel methodology presented in this paper.