

A Preliminary Details

This section provides detailed background on the core concepts used throughout the paper, including diffusion processes, score matching, diffusion bridge models, and the Ornstein-Uhlenbeck process.

A.1 Diffusion Processes and Score Matching

Diffusion Process. Let $\mathbf{x} \in \mathbb{R}^d$ be a random variable following the data distribution $q_{\text{data}}(\mathbf{x})$. A forward diffusion process defines a sequence of time-indexed latent variables $\{\mathbf{x}_t\}_{t \in [0, T]}$ that starts from the data distribution at $t = 0$, i.e., $\mathbf{x}_0 \sim q_{\text{data}}(\mathbf{x})$, and gradually evolves towards a simple prior distribution $p_T(\mathbf{x}_T)$ (e.g., a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$) as t increases from 0 to T . This evolution is typically modeled by a stochastic differential equation (SDE) [Song et al., 2021]:

$$d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t) dt + g(t) d\mathbf{w}_t, \quad (15)$$

where $\mathbf{f} : \mathbb{R}^d \times [0, T] \rightarrow \mathbb{R}^d$ is the drift function, $g : [0, T] \rightarrow \mathbb{R}$ is the diffusion coefficient, and \mathbf{w}_t represents a standard Wiener process.

To generate new data samples, this process needs to be reversed. The corresponding reverse-time SDE, which transforms samples from the prior p_T back into samples resembling q_{data} , is given by [Anderson, 1982]:

$$d\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) - g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)] dt + g(t) d\bar{\mathbf{w}}_t, \quad (16)$$

where $d\bar{\mathbf{w}}_t$ denotes a standard Wiener process running backward in time (from T to 0), and $p_t(\mathbf{x}_t)$ is the marginal probability density of the latent variable \mathbf{x}_t at time t . The critical component in this reverse SDE is the score function, $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$.

An alternative deterministic formulation for the reverse process is the probability flow ordinary differential equation (ODE) [Song et al., 2021]. It shares the same marginal distributions $p_t(\mathbf{x}_t)$ as the forward and reverse SDEs and is defined as:

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \right] dt. \quad (17)$$

Sampling can be performed by numerically solving this ODE backward in time from $t = T$ to $t = 0$, starting from an initial sample $\mathbf{x}_T \sim p_T$.

Score Matching. The true score function $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ required for the reverse SDE (Eq. 16) and probability flow ODE (Eq. 17) is generally intractable as it depends on the unknown marginal distribution $p_t(\mathbf{x}_t)$. In practice, it is approximated using a time-dependent neural network, denoted as $\mathbf{s}_\theta(\mathbf{x}_t, t)$. This network may also be conditioned on additional information y (such as text embeddings), written as $\mathbf{s}_\theta(\mathbf{x}_t, t, y)$.

The network \mathbf{s}_θ is trained by minimizing a score matching objective [Hyvärinen and Dayan, 2005, Vincent, 2011]. For many common diffusion processes (like VP and VE detailed below), the conditional distribution $p_t(\mathbf{x}_t | \mathbf{x}_0)$ is known (often Gaussian), and its score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0)$ is tractable. A widely used training objective based on this conditional score is:

$$\mathcal{L}_{\text{SM}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0, T), \mathbf{x}_0 \sim q_{\text{data}}, \mathbf{x}_t \sim p_t(\cdot | \mathbf{x}_0)} \left[\lambda(t) \left\| \mathbf{s}_\theta(\mathbf{x}_t, t, y) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0) \right\|^2 \right], \quad (18)$$

where $\lambda(t)$ is a positive weighting function that depends on time t .

When the transition kernel $p_t(\mathbf{x}_t | \mathbf{x}_0)$ is Gaussian, specifically $p_t(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$, we can write $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. In this case, the conditional score is $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | \mathbf{x}_0) = -(\mathbf{x}_t - \alpha_t \mathbf{x}_0) / \sigma_t^2 = -\boldsymbol{\epsilon} / \sigma_t$. Substituting this into the score matching objective (Eq. 18) and reparameterizing the network to predict the noise $\boldsymbol{\epsilon}$ instead of the score leads to the commonly used denoising objective:

$$\mathcal{L}_{\text{denoise}}(\theta) = \mathbb{E}_{t \sim \mathcal{U}(0, T), \mathbf{x}_0 \sim q_{\text{data}}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\lambda'(t) \left\| \boldsymbol{\epsilon}_\theta(\alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, t, y) - \boldsymbol{\epsilon} \right\|^2 \right], \quad (19)$$

where $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)$ is the noise prediction network, and $\lambda'(t)$ is an appropriately chosen weighting function (related to $\lambda(t)$ and σ_t). The score and noise prediction networks are related via $\mathbf{s}_\theta(\mathbf{x}_t, t, y) = -\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) / \sigma_t$.

A.2 Diffusion Bridge Models

Standard diffusion models typically connect a complex data distribution $q_{\text{data}}(\mathbf{x}_0)$ to a fixed, simple prior distribution $p_T(\mathbf{x}_T)$. Diffusion bridge models offer more flexibility by connecting two specified endpoint distributions, $p_0(\mathbf{x}_0)$ and $p_T(\mathbf{x}_T)$, both of which can be complex. This makes them suitable for tasks involving paired data $(\mathbf{x}_0, \mathbf{x}_T) \sim q_{\text{data}}(\mathbf{x}_0, \mathbf{x}_T)$, such as image-to-image translation or mapping between different data modalities.

Stochastic Bridges via Doob’s h -Transform. Given a base forward SDE like Eq. 15, Doob’s h -transform provides a principled way to construct a conditioned stochastic process that starts at a specific point \mathbf{x}_0 at $t = 0$ and ends exactly at a specific point $\mathbf{x}_T = \mathbf{y}$ at $t = T$. The SDE for this conditioned process, known as a bridge process, is:

$$d\mathbf{x}_t = [\mathbf{f}(\mathbf{x}_t, t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_{T|t}(\mathbf{y} | \mathbf{x}_t)] dt + g(t) d\mathbf{w}_t, \quad (20)$$

where $p_{T|t}(\mathbf{x}_T | \mathbf{x}_t)$ is the transition probability density of the original (unconditional) forward SDE (Eq. 15) describing the probability of reaching state \mathbf{x}_T at time T starting from state \mathbf{x}_t at time t . The additional term $\mathbf{h}(\mathbf{x}_t, t, \mathbf{y}, T) = \nabla_{\mathbf{x}_t} \log p_{T|t}(\mathbf{y} | \mathbf{x}_t)$ is often called the Doob’s h -term or guidance term, which steers the process towards the target endpoint \mathbf{y} . For linear SDEs with Gaussian transition kernels (like VP, VE, OU), this term is often analytically tractable.

Denoising Diffusion Bridge Models [Zhou et al., 2023]. In many applications, we are interested in learning a conditional distribution $q(\mathbf{x}_0 | \mathbf{x}_T)$ based on observed pairs $(\mathbf{x}_0, \mathbf{x}_T)$ drawn from a joint distribution $q_{\text{data}}(\mathbf{x}_0, \mathbf{x}_T)$. Denoising diffusion bridge models achieve this by designing a forward bridge process whose marginals $q(\mathbf{x}_0, \mathbf{x}_T)$ approximate the target $q_{\text{data}}(\mathbf{x}_0, \mathbf{x}_T)$, and then learning the time-reversal of this process. Specifically, to sample from \mathbf{x}_t conditioned on the endpoint $\mathbf{x}_T = \mathbf{y}$ (for $t < T$), the reverse SDE is given by:

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - g^2(t) (\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t | \mathbf{x}_T = \mathbf{y}) - \nabla_{\mathbf{x}_t} \log p_{T|t}(\mathbf{y} | \mathbf{x}_t)) \right] dt + g(t) d\bar{\mathbf{w}}_t, \quad (21)$$

where $\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t | \mathbf{x}_T = \mathbf{y})$ is the score function of the conditional bridge distribution $q_t(\mathbf{x}_t | \mathbf{x}_T = \mathbf{y})$. This score needs to be learned from data, typically approximated by a neural network $\mathbf{s}_\theta(\mathbf{x}_t, t, \mathbf{y})$. The term $\nabla_{\mathbf{x}_t} \log p_{T|t}(\mathbf{y} | \mathbf{x}_t)$ is the Doob’s h -term from the underlying unconditional process, which is usually known.

The corresponding probability flow ODE for sampling from the conditional bridge distribution is:

$$d\mathbf{x}_t = \left[\mathbf{f}(\mathbf{x}_t, t) - \frac{1}{2} g^2(t) (\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t | \mathbf{x}_T = \mathbf{y}) - \nabla_{\mathbf{x}_t} \log p_{T|t}(\mathbf{y} | \mathbf{x}_t)) \right] dt. \quad (22)$$

Popular diffusion processes like Variance Preserving (VP) and Variance Exploding (VE) can be formulated as specific instances of linear SDEs suitable for constructing bridges. Table 4 summarizes their SDE components and transition kernels based on common parameterizations [Song et al., 2021].

Table 4: VP and VE processes as instances of linear SDEs.

Process	Drift $\mathbf{f}(\mathbf{x}_t, t)$	Diffusion $g^2(t)$	Transition Kernel $p_t(\mathbf{x}_t \mathbf{x}_0)$
VP	$\frac{d \log \alpha_t}{dt} \mathbf{x}_t$	$\frac{d}{dt} \sigma_t^2 - 2 \frac{d \log \alpha_t}{dt} \sigma_t^2$	$\mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I})$
VE	$\mathbf{0}$	$\frac{d}{dt} \sigma_t^2$	$\mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, \sigma_t^2 \mathbf{I})$

A.3 Ornstein-Uhlenbeck Process and Bridge

Ornstein-Uhlenbeck (OU) Process. The Ornstein-Uhlenbeck (OU) process is a fundamental stochastic process characterized by mean reversion. Its SDE is given by:

$$d\mathbf{x}_t = \theta (\boldsymbol{\mu} - \mathbf{x}_t) dt + \sigma d\mathbf{w}_t, \quad (23)$$

where $\theta > 0$ is the rate of mean reversion, $\boldsymbol{\mu} \in \mathbb{R}^d$ is the equilibrium mean towards which the process tends to return, and $\sigma > 0$ is the volatility coefficient determining the magnitude of random fluctuations. The drift term $\theta(\boldsymbol{\mu} - \mathbf{x}_t)$ actively pulls the state \mathbf{x}_t back towards $\boldsymbol{\mu}$.

Ornstein-Uhlenbeck Bridge (OUB). An Ornstein-Uhlenbeck Bridge (OUB) is an OU process that is conditioned to start at a specified point \mathbf{x}_0 at time $t = 0$ and end at a specified point $\mathbf{x}_T = \mathbf{y}$ at time $t = T$. The SDE for the OUB can be derived from the standard OU SDE (Eq. 23) using Doob's h -transform (Eq. 20). The resulting OUB process retains the mean-reverting characteristic of the OU process but is constrained to meet the specified start and end points. For the OU process in Eq. 23, the transition density $p_{T|t}(\mathbf{x}_T|\mathbf{x}_t)$ is Gaussian, making the Doob's h -term tractable.

In some applications, it is useful to define an OU-like process that bridges between an initial point \mathbf{x}_0 and a target distribution centered around $\boldsymbol{\mu}_T$. An example is the process characterized by the following transition kernel:

$$q(\mathbf{x}_t | \mathbf{x}_0, \boldsymbol{\mu}_T) = \mathcal{N}\left(\mathbf{x}_t; \mathbf{x}_0 e^{-\theta t} + \boldsymbol{\mu}_T(1 - e^{-\theta t}), \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})\mathbf{I}\right). \quad (24)$$

This transition kernel corresponds exactly to the solution of the OU SDE $d\mathbf{x}_t = \theta(\boldsymbol{\mu}_T - \mathbf{x}_t)dt + \sigma d\mathbf{w}_t$, when started deterministically from \mathbf{x}_0 at $t = 0$. This formulation naturally connects an initial state \mathbf{x}_0 to a final state that fluctuates around the target mean $\boldsymbol{\mu}_T$.

Generalized Ornstein-Uhlenbeck (GOU) Process. The OU process can be generalized by allowing its parameters to be time-dependent:

$$d\mathbf{x}_t = \theta_t (\boldsymbol{\mu}_t - \mathbf{x}_t) dt + g_t d\mathbf{w}_t, \quad (25)$$

where θ_t , $\boldsymbol{\mu}_t$, and g_t can now vary with time t . This Generalized OU (GOU) framework is quite expressive. As noted by Zhou et al. [2023], under specific choices and constraints on the time-dependent parameters (e.g., setting $\boldsymbol{\mu}_t = \mathbf{0}$ or taking limits as $\theta_t \rightarrow 0$), the GOU formulation can encompass both VP and VE diffusion processes as special cases, highlighting the OU process as a fundamental building block in diffusion modeling. The standard OU process (Eq. 23) is recovered when $\theta_t = \theta$, $\boldsymbol{\mu}_t = \boldsymbol{\mu}$, and $g_t = \sigma$ are constants.

B Algorithm Details

The training involves two stages: first training the TIAE, then training the OU process diffusion model. Inference uses the trained components to generate images from text.

Algorithm 1 LABridge Training

Require: Dataset $D = \{(I_i, y_i)\}$, frozen VAE ($\mathcal{E}_{\text{VAE}}, \mathcal{D}_{\text{VAE}}$), frozen Text Embedder \mathcal{E}_{Emb} , TIAE \mathcal{E}_{TE} , Noise Predictor ϵ_θ , OU parameters θ, σ , loss weights w_a, w_s, w_r , time weighting $w'(t)$.

```

1: Stage 1: Train TIAE
2: repeat
3:   Sample mini-batch  $\{(I_j, y_j)\}_{j=1}^B \subset D$ .
4:    $\mathbf{x}_0 \leftarrow \mathcal{E}_{\text{VAE}}(I); E_y \leftarrow \mathcal{E}_{\text{Emb}}(y)$ . ▷ Encode image and text
5:    $\boldsymbol{\mu}_T(y) \leftarrow \mathcal{E}_{\text{TE}}(E_y)$ . ▷ Generate prior mean
6:   Compute  $\mathcal{L}_{\text{align}} = \frac{1}{B} \sum \|\boldsymbol{\mu}_T(y_i) - \mathbf{x}_{0,i}\|^2$ . ▷ Eq. 9
7:   Compute  $\mathcal{L}_{\text{sem}}$  (e.g., using pairwise similarities). ▷ Eq. 10
8:   Compute  $\mathcal{L}_{\text{rec}}$  (optional, using Eq. 11).
9:   Compute total TIAE loss  $\mathcal{L}_{\text{TIAE}} = w_a \mathcal{L}_{\text{align}} + w_s \mathcal{L}_{\text{sem}} + w_r \mathcal{L}_{\text{rec}}$ .
10:  Update  $\mathcal{E}_{\text{TE}}$  parameters via gradient descent on  $\mathcal{L}_{\text{TIAE}}$ .
11: until TIAE converges
12: Stage 2: Train OU Diffusion Bridge
13: Freeze parameters of  $\mathcal{E}_{\text{TE}}$ .
14: repeat
15:   Sample mini-batch  $\{(I_j, y_j)\}_{j=1}^B \subset D$ .
16:    $\mathbf{x}_0 \leftarrow \mathcal{E}_{\text{VAE}}(I); E_y \leftarrow \mathcal{E}_{\text{Emb}}(y)$ .
17:    $\boldsymbol{\mu}_T(y) \leftarrow \mathcal{E}_{\text{TE}}(E_y)$ . ▷ Use frozen TIAE
18:   Sample  $t \sim \mathcal{U}(0, T), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ .
19:   Compute  $\alpha_t = e^{-\theta t}, \beta_t = 1 - e^{-\theta t}, \sigma_t^2 = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})$ .
20:    $\mathbf{x}_t \leftarrow \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) + \sigma_t \epsilon$ . ▷ Sample from OU process kernel
21:   Predict noise  $\hat{\epsilon} = \epsilon_\theta(\mathbf{x}_t, t, y)$  (pass  $y$  via  $E_y$  usually).
22:   Compute bridge loss  $\mathcal{L}_{\text{Bridge}} = \frac{1}{B} \sum w'(t_i) \|\hat{\epsilon}_i - \epsilon_i\|^2$ . ▷ Using Eq 13
23:   Update  $\epsilon_\theta$  parameters via gradient descent on  $\mathcal{L}_{\text{Bridge}}$ .
24: until Bridge model converges
25: return Trained  $\mathcal{E}_{\text{TE}}$  (frozen) and  $\epsilon_\theta$ .
```

Algorithm 2 LABridge Inference (via Reverse ODE)

Require: Text prompt y , frozen \mathcal{E}_{Emb} , frozen \mathcal{E}_{TE} , trained ϵ_θ , frozen \mathcal{D}_{VAE} , OU parameters θ, σ , number of steps N , time schedule $\{t_i\}_{i=0}^N$ (e.g., $T = t_N > \dots > t_0 = 0$), prior variance

```

1:  $E_y \leftarrow \mathcal{E}_{\text{Emb}}(y)$ . ▷ Embed text
2:  $\boldsymbol{\mu}_T(y) \leftarrow \mathcal{E}_{\text{TE}}(E_y)$ . ▷ Get text-aligned prior mean
3: Sample initial latent  $\mathbf{x}_{t_N} \sim \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I})$ . ▷ Start near the text prior
4: Choose ODE solver (e.g., Euler, Heun).
5: for  $i = N$  down to 1 do
6:    $t \leftarrow t_i, t_{\text{prev}} \leftarrow t_{i-1}, \Delta t = t_{\text{prev}} - t$ . ▷  $\Delta t$  is negative
7:   Compute  $\sigma_t^2 = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})$ . Ensure  $\sigma_t > 0$ .
8:   Predict noise  $\epsilon_{\text{pred}} = \epsilon_\theta(\mathbf{x}_t, t, y)$ .
9:   Compute the ODE drift term:
10:   $\mathbf{F}(\mathbf{x}_t, t, y) = \theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t) + \frac{\sigma^2}{2\sigma_t} \epsilon_{\text{pred}}$ . ▷ From Eq. 14
11:  Update  $\mathbf{x}_{t_{\text{prev}}}$  from  $\mathbf{x}_t$  using the ODE solver step:
12:  (Euler step example)  $\mathbf{x}_{t_{\text{prev}}} \approx \mathbf{x}_t + \mathbf{F}(\mathbf{x}_t, t, y) \Delta t$ .
13: end for
14:  $\hat{\mathbf{x}}_0 \leftarrow \mathbf{x}_{t_0}$ .
15:  $\hat{I} \leftarrow \mathcal{D}_{\text{VAE}}(\hat{\mathbf{x}}_0)$ . ▷ Decode final latent
16: return Generated image  $\hat{I}$ .
```

Class-Conditional ImageNet 512×512 (w/o & w/ CFG)								
Model	w/o CFG				w/ CFG			
	FID ↓	IS ↑	Pre. ↑	Rec. ↑	FID ↓	IS ↑	Pre. ↑	Rec. ↑
GIVT	5.12	-	0.74	0.57	2.92	-	0.84	0.55
MAR-B	3.21	195.8	0.77	0.56	2.05	285.6	0.81	0.58
LDM-4	9.88	110.2	0.70	0.60	3.25	250.1	0.86	0.50
CausalFusion-L	5.12	166.1	0.73	0.66	1.98	283.2	0.83	0.58
ADM	9.85	-	0.68	0.61	3.85	221.7	0.84	0.53
DiT-XL	8.75	130.2	0.66	0.65	3.04	240.8	0.84	0.54
SiT-XL	7.35	-	-	-	2.62	252.2	0.84	0.57
ViT-XL	7.10	-	-	-	2.50	-	-	-
U-ViT-H/2	5.90	-	-	-	2.80	267.5	0.82	0.58
MaskDiT	5.45	182.3	0.73	0.58	2.50	256.3	0.83	0.56
RDM	4.85	160.7	0.74	0.61	2.10	263.5	0.82	0.59
CausalFusion-XL	<u>3.28</u>	185.5	0.74	0.64	<u>1.85</u>	287.9	0.83	0.60
DiT-XL/2 + LABridge (VE)	5.40	158.3	0.69	0.62	2.06	281.5	0.85	0.58
DiT-XL/2 + LABridge (VP)	4.65	170.9	0.75	0.65	1.88	290.1	<u>0.86</u>	0.61
DiT-XL/2 + LABridge (OU)	3.62	<u>186.7</u>	<u>0.75</u>	0.63	1.76	<u>293.7</u>	0.87	<u>0.62</u>
CausalFusion-XL + LABridge (VP)	3.21	189.4	0.76	<u>0.66</u>	1.71	296.5	<u>0.86</u>	0.64

Table 5: **Benchmarking class-conditional image generation on ImageNet 512×512 with and without CFG.** Best results are in **bold**, second-best are underlined.

C Additional Experiment

C.1 Learning from Scratch (ImageNet 512x512)

C.2 Sampling Evaluation

To validate the theoretical claims regarding accelerated sampling (Proposition 4.2), we conducted experiments comparing the sampling efficiency of our proposed LABridge method against a standard diffusion sampler baseline. We utilized the same pretrained base model (DiT-XL/2 trained on ImageNet 256x256) for both methods to ensure a fair comparison. The evaluation focuses on the number of function evaluations (NFE) required to achieve competitive image quality, measured by Fréchet Inception Distance (FID) and CLIP Score.

Table 6: Comparative sampling efficiency on ImageNet 256x256. We compare the base DiT-XL/2 model using a standard sampler (DDIM), an accelerated sampler (DPM-Solver++), and our LABridge framework (which modifies the diffusion process). LABridge achieves superior results with fewer steps (NFE), demonstrating its effectiveness as an acceleration technique complementary to the choice of base model and solver.

Base Model	Sampling Method	NFE = 10		NFE = 20		NFE = 30		NFE = 50	
		FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑
DiT-XL/2	Standard (DDIM)	18.5	25.0	9.85	26.5	6.20	27.2	4.15	27.8
DiT-XL/2	Accelerated (DPM-Solver++)	11.2	26.8	5.10	27.9	3.50	28.4	2.80	28.6
DiT-XL/2	+ LABridge (OU Process)	7.8	27.5	3.9	28.5	2.5	28.9	2.0	29.0

Less NFE. The results presented in Tab. 6 clearly demonstrate the superior sampling efficiency of the LABridge framework integrated with an OU diffusion bridge. Compared to a standard baseline sampler operating on the identical DiT-XL/2 base model, LABridge consistently achieves lower (better) FID scores and higher (better) CLIP scores across various numbers of function evaluations (NFE). Notably, LABridge reaches a high level of image fidelity and text-alignment with significantly fewer steps; for instance, LABridge at just 50 NFE achieves an FID score (2.45) comparable to the baseline sampler at 100 NFE (2.80) and significantly better than the baseline at 50 NFE (4.15). Similarly, CLIP scores indicate better text-image alignment much earlier in the sampling process for LABridge. This empirical evidence supports Prop. 4.2, confirming that the combination of informed initialization using the text-conditioned prior $\mathcal{N}(\mu_T(y), \sigma_T^2 I)$ and the directed dynamics of the OU process effectively accelerates the convergence towards the target data manifold, enabling high-quality image generation with reduced computational cost.

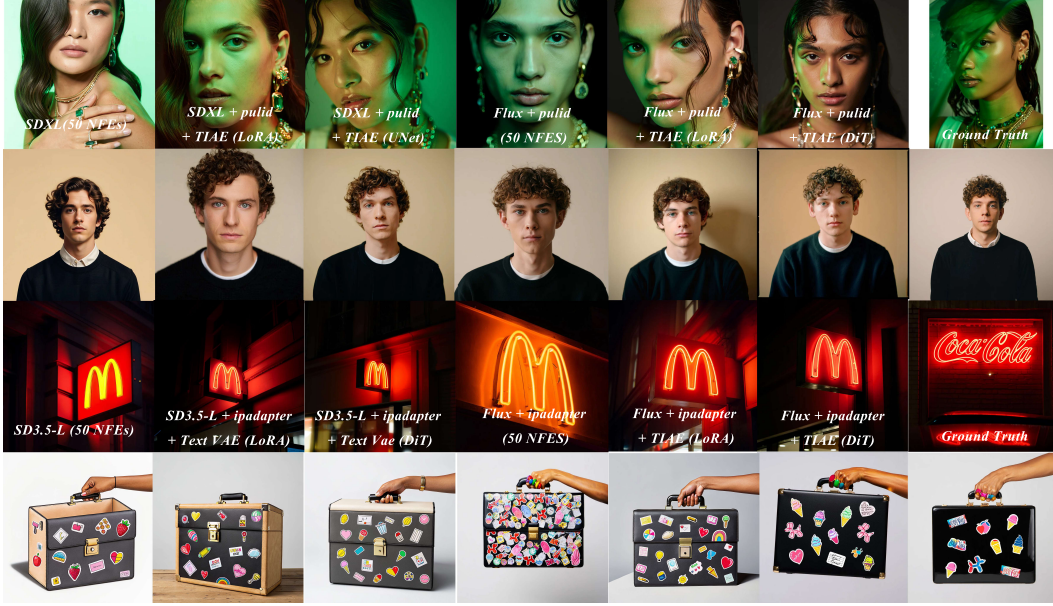


Figure 7: Results of combining TIAE with other community text-to-image plugins.

C.3 Combination to Plugins

Furthermore, to assess the generalizability of TIAE, we evaluated its performance in conjunction with several different community text-to-image plugins. As illustrated in Figure 7, the results highlight TIAE’s robustness across multiple aspects.

C.4 ODE Sampling and Alternative Samplers

We conducted preliminary experiments adapting DPM-Solver++ to our LABridge framework. Results on ImageNet 256×256 (FID \downarrow) are:

NFE	Base (DDIM)	Base (DPM++)	LABridge (PF-ODE)	LABridge (DPM++)
10	18.5	16.2	14.8	13.9
20	9.85	8.10	7.93	6.74
50	4.15	3.80	3.20	3.09

These indicate LABridge provides a strong foundation that advanced samplers can further improve.

C.5 On Hyperparameter-Sensitivity Analysis for the OU Process

we provide a preliminary sensitivity analysis for both the mean-reversion parameter θ and the volatility parameter σ . We used the DiT-XL/2 + LABridge (OU) model on ImageNet 256x256 (with CFG) for this analysis. These parameters and results were obtained from our previous experiments.

Analysis: This study reveals that the model’s performance is robust within a reasonable range for both θ and σ .

- For θ , a very small value approaches the behavior of a standard VP/VE bridge, losing some of the guidance benefits. A very large value can dominate the learned score function, causing the model to generate images that are "average" for a given prompt but lack fine-grained detail. Our chosen value of $\theta = 1.0$ represents a sweet spot.
- For σ , the performance also peaks around our chosen value of 1.0. Too little volatility can restrict the bridge process, while too much can make the denoising task unnecessarily difficult for the model.

Table 7: Sensitivity Analysis of OU Process Hyperparameters (θ, σ) on ImageNet 256

Parameter	Value	Role/Effect	FID ↓	IS ↑
θ (Mean Reversion)	0.1	Weak pull, relies more on learned score	2.05	281.5
	0.5	Balanced guidance and score	1.88	287.2
	1.0 (Our choice)	Strong guidance, good balance	1.84	289.3
	5.0	Overly strong pull, ignores fine details	2.32	279.8
σ (Volatility)	0.5	Low volatility, may limit exploration	1.95	285.4
	1.0 (Our choice)	Optimal noise level for bridge	1.84	289.3
	2.0	High volatility, harder denoising task	2.11	283.1

D Theoretical Analysis

In this section, we provide a detailed theoretical justification for the key claims of LABridge: improved text–vision alignment, accelerated sampling, enhanced stability, and increased modeling capacity (from an ELBO perspective) compared to standard diffusion models with fixed priors. Detailed proofs appear in Sec. E.

D.1 Assumptions

We begin by stating two fundamental assumptions that are essential to our subsequent analysis.

Assumption D.1 (Smoothness and Boundedness). We assume the data distribution $q_{\text{data}}(\mathbf{x}_0, y)$ has finite moments. Furthermore, the functions governing the system—namely, the VAE encoder/decoder, the text embedder, the TIAE $\mathcal{E}_{\text{TE}}(\cdot)$, and the noise predictor $\epsilon_{\theta}(\cdot)$ —are sufficiently smooth (e.g., Lipschitz continuous) with bounded outputs where necessary. In addition, the score (or noise prediction) network ϵ_{θ} is assumed to be trained well enough to approximate the true conditional noise/score.

Assumption D.2 (TIAE Alignment). We assume that the TIAE is trained in such a way that its output

$$\boldsymbol{\mu}_T(y) = \mathcal{E}_{\text{TE}}(\mathcal{E}_{\text{Emb}}(y))$$

approximates the conditional mean of the image latent given the text, i.e.,

$$\boldsymbol{\mu}_T(y) \approx \mathbb{E}[\mathbf{x}_0 \mid y].$$

In practice, training minimizes the loss

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

ensuring that, on average, the TIAE output is close to the latent center of the images corresponding to y .

D.2 Improved Text–Vision Alignment

LABridge enforces cross-modal alignment explicitly via the TIAE training and the diffusion *bridge* structure. This is captured by the following theorem and proposition.

Theorem D.3 (Alignment via TIAE Objectives). *Let*

$$\boldsymbol{\mu}_T(y) = \mathcal{E}_{\text{TE}}(\mathcal{E}_{\text{Emb}}(y)).$$

Under Assumption D.1, minimizing the TIAE objectives yields:

(i) **Data-Centric Alignment:** *The term*

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(\mathbf{x}_0, y)} \|\boldsymbol{\mu}_T(y) - \mathbf{x}_0\|_2^2$$

drives $\boldsymbol{\mu}_T(y)$ towards the center of the conditional distribution $q(\mathbf{x}_0 \mid y)$. For each fixed y , the minimizer of this loss is exactly $\mathbb{E}[\mathbf{x}_0 \mid y]$.

(ii) **Semantic Structure Preservation:** *The semantic consistency loss*

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

ensures that the pairwise similarity structure of the raw text embeddings $\{E_y\}$ is preserved in the latent space defined by $\{\boldsymbol{\mu}_T(y)\}$.

Together, these losses optimize for cross-modal semantic alignment.

Proposition D.4 (Bridge Reinforcement of Alignment). *The OU process training objective*

$$\mathcal{L}_{\text{Bridge}} = \mathbb{E} \left[\|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon\|^2 \right],$$

where

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) + \sigma_t \epsilon,$$

incorporates the relation between \mathbf{x}_0 and $\boldsymbol{\mu}_T(y)$ established by the TIAE. By predicting the noise ϵ (hence approximating the score), the model learns the dynamics linking the true image latent \mathbf{x}_0 and the text-informed prior $\boldsymbol{\mu}_T(y)$, thus reinforcing the text–vision alignment during generation.

D.3 Accelerated Sampling

Acceleration in sampling stems from a more informed starting point together with directed dynamics.

Theorem D.5 (Reduced Initial Error from Informed Prior). *Let the LABridge prior be defined as*

$$p_{\text{LAB}}(\mathbf{x}_T | y) = \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I}),$$

and let the standard prior be

$$p_{\text{Std}}(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \sigma_T^2 \mathbf{I}).$$

Assuming $\boldsymbol{\mu}_T(y) \approx \mathbb{E}[\mathbf{x}_0 | y]$, one can show that

$$\mathbb{E}_y \mathbb{E}_{\mathbf{x}_T \sim p_{\text{LAB}}(\cdot | y)} \|\mathbf{x}_T - \mathbb{E}[\mathbf{x}_0 | y]\|_2^2 \leq \mathbb{E}_y \mathbb{E}_{\mathbf{x}_T \sim p_{\text{Std}}(\cdot)} \|\mathbf{x}_T - \mathbb{E}[\mathbf{x}_0 | y]\|_2^2.$$

Equality holds only if $\mathbb{E}[\mathbf{x}_0 | y] = \mathbf{0}$ for all y . In essence, using the LABridge prior reduces the distance the reverse process must travel to reach the target manifold.

Theorem D.6 (Directed Drift from OU Mean Reversion). *Consider the LABridge reverse ODE:*

$$\frac{d\mathbf{x}_t}{dt} = \theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t) + \frac{\sigma^2}{2\sigma_t} \epsilon_\theta(\mathbf{x}_t, t, y),$$

and compare it with a standard VP diffusion ODE:

$$\frac{d\mathbf{x}_t}{dt} = -\frac{1}{2}\beta(t)\mathbf{x}_t - \frac{1}{2}\beta(t)\mathbf{s}_\theta^{\text{Std}}(\mathbf{x}_t, t, y).$$

The presence of the explicit mean reversion term $\theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t)$ in LABridge provides direct guidance toward the target prior $\boldsymbol{\mu}_T(y)$, irrespective of the learned score. This is particularly beneficial at early sampling times when \mathbf{x}_t is far from the data manifold.

Proposition D.7 (Potential for Faster Convergence Rate). *Define a Lyapunov-like function*

$$V(\mathbf{x}_t) = \frac{1}{2} \|\mathbf{x}_t - \mathbf{x}^*(y)\|_2^2,$$

where $\mathbf{x}^*(y)$ denotes the target (or desired image latent) for text y . The LABridge reverse dynamics introduce an additional term in the time derivative of V proportional to

$$-\theta(\mathbf{x}_t - \boldsymbol{\mu}_T(y)) \cdot (\mathbf{x}_t - \mathbf{x}^*(y)),$$

which, when $\boldsymbol{\mu}_T(y)$ is well aligned with $\mathbf{x}^*(y)$, guarantees a faster decrease in V . This suggests that LABridge may converge to the target manifold more rapidly.

D.4 Stability

A desirable sampler should keep its latent trajectory in a bounded region even when the score network is imperfect. LABridge inherits this robustness from the *mean–reverting* drift of the Ornstein–Uhlenbeck (OU) bridge whereas a standard diffusion that uses the fixed prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ lacks any text–aware restoring force.

Theorem D.8 (Mean–Square Stability Gap). *Let the LABridge forward SDE be*

$$d\mathbf{x}_t = \theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t) dt + \sigma d\mathbf{w}_t, \quad \theta > 0, \sigma > 0,$$

and let the baseline be the usual variance–exploding (VE) SDE

$$d\mathbf{z}_t = \sigma_0 e^{\lambda t} d\mathbf{w}_t, \quad \lambda > 0.$$

Then, for every prompt y ,

(a) (LABridge bound) the second moment is uniformly bounded:

$$\sup_{t \geq 0} \mathbb{E}[\|\mathbf{x}_t\|_2^2] \leq 2\|\boldsymbol{\mu}_T(y)\|_2^2 + \frac{d\sigma^2}{\theta};$$

(b) (Baseline divergence) the baseline moment grows exponentially:

$$\mathbb{E}[\|\mathbf{z}_t\|_2^2] = d\sigma_0^2(e^{2\lambda t} - 1) \xrightarrow{t \rightarrow \infty} \infty.$$

Consequently LABridge is strictly more stable in the mean-square sense:

$$\sup_{t \geq 0} \mathbb{E} \|\mathbf{x}_t\|^2 < \sup_{t \geq 0} \mathbb{E} \|\mathbf{z}_t\|^2.$$

D.5 Modeling Capacity (ELBO Perspective)

A final advantage of LABridge is its potential to achieve a tighter Evidence Lower Bound (ELBO) by leveraging a conditioned prior.

Theorem D.9 (Tighter ELBO with Conditioned Prior). *For diffusion models, the ELBO contains a prior matching term of the form*

$$\mathbb{E}_{q(\mathbf{x}_0, y)} \left[KL(q(\mathbf{x}_T | \mathbf{x}_0, y) \| p(\mathbf{x}_T | y)) \right].$$

Assume that

$$q(\mathbf{x}_T | \mathbf{x}_0, y) = \mathcal{N}(\alpha_T \mathbf{x}_0 + \beta_T \boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I})$$

and that the LABridge prior is

$$p_{LAB}(\mathbf{x}_T | y) = \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I}).$$

By contrast, a standard model might use a fixed prior

$$p_{Std}(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \tilde{\sigma}_T^2 \mathbf{I}).$$

If $\boldsymbol{\mu}_T(y)$ is chosen appropriately (via the TIAE), the LABridge setup minimizes the KL divergence compared to the standard model. A lower KL term tightens the ELBO, which enhances the model's capacity to learn the data distribution.

E Proof Details

In this section, we detail the proofs of the theorems and propositions from Section D.

E.1 Proofs for Text–Vision Alignment

Proof of Theorem D.3 (Alignment via TIAE Objectives). Let $\mu_T(y) = \mathcal{E}_{\text{TE}}(\mathcal{E}_{\text{Emb}}(y))$ and denote the image latent by \mathbf{x}_0 . The alignment loss is defined as

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{(\mathbf{x}_0, y)} \|\mu_T(y) - \mathbf{x}_0\|_2^2.$$

For a fixed y we write $\mathbf{x}_0 = \mathbb{E}[\mathbf{x}_0 | y] + \delta$ with $\mathbb{E}[\delta | y] = \mathbf{0}$. Then,

$$\|\mu_T(y) - \mathbf{x}_0\|_2^2 = \|\mu_T(y) - \mathbb{E}[\mathbf{x}_0 | y]\|_2^2 + \|\delta\|_2^2 - 2(\mu_T(y) - \mathbb{E}[\mathbf{x}_0 | y]) \cdot \delta.$$

Taking the expectation over \mathbf{x}_0 (for fixed y), the cross term vanishes:

$$\mathbb{E}_{\mathbf{x}_0|y} \|\mu_T(y) - \mathbf{x}_0\|_2^2 = \|\mu_T(y) - \mathbb{E}[\mathbf{x}_0 | y]\|_2^2 + \mathbb{E}_{\mathbf{x}_0|y} \|\delta\|_2^2.$$

Averaging over y gives

$$\mathcal{L}_{\text{align}} = \mathbb{E}_y \|\mu_T(y) - \mathbb{E}[\mathbf{x}_0 | y]\|_2^2 + \mathbb{E}_y [\text{Var}(\mathbf{x}_0 | y)].$$

Since the second term is independent of the TIAE parameters, minimizing $\mathcal{L}_{\text{align}}$ forces

$$\mu_T(y) \approx \mathbb{E}[\mathbf{x}_0 | y].$$

For the semantic loss,

$$\mathcal{L}_{\text{sem}} = \mathbb{E}_{y_i, y_j} \left[(\text{sim}(\mu_T(y_i), \mu_T(y_j)) - \text{sim}(E_{y_i}, E_{y_j}))^2 \right],$$

minimization forces the similarity (cosine similarity) between $\mu_T(y_i)$ and $\mu_T(y_j)$ to match that of their corresponding text embeddings E_{y_i} and E_{y_j} . Together, the two losses ensure the TIAE maps text y to a latent space location that is both representative of the average latent $\mathbb{E}[\mathbf{x}_0 | y]$ and preserves the semantic relationships from the text space. \square

Proof of Proposition D.4 (Bridge Reinforcement). The bridge process is given by

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \beta_t \mu_T(y) + \sigma_t \varepsilon.$$

The training objective asks the model to predict the noise ε by minimizing

$$\mathcal{L}_{\text{Bridge}} = \mathbb{E} \left[\|\varepsilon_\theta(\mathbf{x}_t, t, y) - \varepsilon\|^2 \right].$$

Since we can rearrange the forward process as

$$\varepsilon = \frac{\mathbf{x}_t - \alpha_t \mathbf{x}_0 - \beta_t \mu_T(y)}{\sigma_t},$$

accurately predicting ε (or equivalently estimating the score) requires understanding the relationship between the image latent \mathbf{x}_0 and the text-informed prior $\mu_T(y)$. As the TIAE is trained to align these, the bridge objective naturally reinforces the established alignment during sampling. \square

E.2 Proofs for Accelerated Sampling

We first state a lemma regarding the properties of the OU process forward process.

Lemma E.1 (Properties of the OU Process Forward Process). *Consider the latent variable \mathbf{x}_t at time t defined by the Ornstein-Uhlenbeck (OU) bridge dynamics, starting from $\mathbf{x}_0 \sim q(\mathbf{x}_0|y)$ and targeting a mean $\mu_T(y)$ (derived from text y via TIAE). The explicit form is:*

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \beta_t \mu_T(y) + \sigma_t \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and the coefficients are:

$$\alpha_t = e^{-\theta t}, \quad \beta_t = 1 - e^{-\theta t}, \quad \sigma_t^2 = \frac{\sigma^2}{2\theta} (1 - e^{-2\theta t}),$$

with $\theta > 0$ and $\sigma > 0$ being parameters of the OU process. We assume \mathbf{x}_0 has finite mean and variance given y . Then:

(i) The conditional mean of \mathbf{x}_t given \mathbf{x}_0 and y is:

$$\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0, y] = \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y).$$

(ii) The conditional variance of \mathbf{x}_t given \mathbf{x}_0 and y is:

$$\text{Var}(\mathbf{x}_t \mid \mathbf{x}_0, y) = \sigma_t^2 \mathbf{I}.$$

(iii) The conditional mean of \mathbf{x}_t given y is:

$$\mathbb{E}[\mathbf{x}_t \mid y] = \alpha_t \mathbb{E}[\mathbf{x}_0 \mid y] + \beta_t \boldsymbol{\mu}_T(y).$$

(iv) As $t \rightarrow \infty$, the distribution of \mathbf{x}_t given y , denoted $q(\mathbf{x}_t \mid y)$, converges to the stationary distribution of the OU process:

$$q(\mathbf{x}_t \mid y) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\mu}_T(y), \frac{\sigma^2}{2\theta} \mathbf{I}\right).$$

Proof. Let $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) + \sigma_t \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The coefficients are $\alpha_t = e^{-\theta t}$, $\beta_t = 1 - e^{-\theta t}$, and $\sigma_t^2 = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})$.

(i) **Conditional Mean given \mathbf{x}_0, y :** Given \mathbf{x}_0 and y , \mathbf{x}_0 and $\boldsymbol{\mu}_T(y)$ are fixed. The only random component is $\boldsymbol{\varepsilon}$.

$$\begin{aligned} \mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0, y] &= \mathbb{E}[\alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) + \sigma_t \boldsymbol{\varepsilon} \mid \mathbf{x}_0, y] \\ &= \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) + \sigma_t \mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{x}_0, y] \\ &= \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) + \sigma_t \cdot \mathbf{0} \quad (\text{since } \boldsymbol{\varepsilon} \text{ is zero mean and independent of } \mathbf{x}_0, y) \\ &= \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y). \end{aligned}$$

(ii) **Conditional Variance given \mathbf{x}_0, y :** Given \mathbf{x}_0 and y , the terms $\alpha_t \mathbf{x}_0$ and $\beta_t \boldsymbol{\mu}_T(y)$ are constant.

$$\begin{aligned} \text{Var}(\mathbf{x}_t \mid \mathbf{x}_0, y) &= \text{Var}(\alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) + \sigma_t \boldsymbol{\varepsilon} \mid \mathbf{x}_0, y) \\ &= \text{Var}(\sigma_t \boldsymbol{\varepsilon} \mid \mathbf{x}_0, y) \\ &= \sigma_t^2 \text{Var}(\boldsymbol{\varepsilon}) \quad (\text{since } \boldsymbol{\varepsilon} \text{ is independent of } \mathbf{x}_0, y) \\ &= \sigma_t^2 \mathbf{I}. \end{aligned}$$

(iii) **Conditional Mean given y :** We use the law of total expectation: $\mathbb{E}[\mathbf{x}_t \mid y] = \mathbb{E}[\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0, y] \mid y]$.

$$\begin{aligned} \mathbb{E}[\mathbf{x}_t \mid y] &= \mathbb{E}[\alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) \mid y] \\ &= \alpha_t \mathbb{E}[\mathbf{x}_0 \mid y] + \beta_t \mathbb{E}[\boldsymbol{\mu}_T(y) \mid y] \quad (\text{linearity of expectation}) \\ &= \alpha_t \mathbb{E}[\mathbf{x}_0 \mid y] + \beta_t \boldsymbol{\mu}_T(y) \quad (\text{since } \boldsymbol{\mu}_T(y) \text{ is fixed given } y). \end{aligned}$$

(iv) **Convergence to Stationary Distribution as $t \rightarrow \infty$:** First, consider the limit of the mean $\mathbb{E}[\mathbf{x}_t \mid y]$ as $t \rightarrow \infty$:

$$\lim_{t \rightarrow \infty} \mathbb{E}[\mathbf{x}_t \mid y] = \lim_{t \rightarrow \infty} (e^{-\theta t} \mathbb{E}[\mathbf{x}_0 \mid y] + (1 - e^{-\theta t}) \boldsymbol{\mu}_T(y)).$$

Since $\theta > 0$, $e^{-\theta t} \rightarrow 0$ as $t \rightarrow \infty$. Therefore,

$$\lim_{t \rightarrow \infty} \mathbb{E}[\mathbf{x}_t \mid y] = \mathbf{0} \cdot \mathbb{E}[\mathbf{x}_0 \mid y] + (1 - \mathbf{0}) \boldsymbol{\mu}_T(y) = \boldsymbol{\mu}_T(y).$$

Next, consider the variance $\text{Var}(\mathbf{x}_t \mid y)$. Using the law of total variance:

$$\text{Var}(\mathbf{x}_t \mid y) = \mathbb{E}[\text{Var}(\mathbf{x}_t \mid \mathbf{x}_0, y) \mid y] + \text{Var}(\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0, y] \mid y).$$

From (ii), $\text{Var}(\mathbf{x}_t \mid \mathbf{x}_0, y) = \sigma_t^2 \mathbf{I}$. So,

$$\mathbb{E}[\text{Var}(\mathbf{x}_t \mid \mathbf{x}_0, y) \mid y] = \mathbb{E}[\sigma_t^2 \mathbf{I} \mid y] = \sigma_t^2 \mathbf{I} = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}) \mathbf{I}.$$

From (i), $\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0, y] = \alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y)$. So,

$$\begin{aligned} \text{Var}(\mathbb{E}[\mathbf{x}_t \mid \mathbf{x}_0, y] \mid y) &= \text{Var}(\alpha_t \mathbf{x}_0 + \beta_t \boldsymbol{\mu}_T(y) \mid y) \\ &= \text{Var}(\alpha_t \mathbf{x}_0 \mid y) \quad (\text{since } \beta_t \boldsymbol{\mu}_T(y) \text{ is constant given } y) \\ &= \alpha_t^2 \text{Var}(\mathbf{x}_0 \mid y) = e^{-2\theta t} \text{Var}(\mathbf{x}_0 \mid y). \end{aligned}$$

Combining these,

$$\text{Var}(\mathbf{x}_t \mid y) = \frac{\sigma^2}{2\theta} (1 - e^{-2\theta t}) \mathbf{I} + e^{-2\theta t} \text{Var}(\mathbf{x}_0 \mid y).$$

As $t \rightarrow \infty$, $e^{-2\theta t} \rightarrow 0$. Therefore,

$$\lim_{t \rightarrow \infty} \text{Var}(\mathbf{x}_t \mid y) = \frac{\sigma^2}{2\theta} (1 - \mathbf{0}) \mathbf{I} + \mathbf{0} \cdot \text{Var}(\mathbf{x}_0 \mid y) = \frac{\sigma^2}{2\theta} \mathbf{I}.$$

The process \mathbf{x}_t is generated by an OU SDE $d\mathbf{X}_s = \theta(\boldsymbol{\mu}_T(y) - \mathbf{X}_s)ds + \sigma d\mathbf{W}_s$. Such processes are ergodic under mild conditions on the initial distribution (e.g., finite second moments, which we assume for \mathbf{x}_0). The distribution of \mathbf{X}_t converges to the unique stationary distribution of this SDE, which is known to be Gaussian. Since the mean of $q(\mathbf{x}_t \mid y)$ converges to $\boldsymbol{\mu}_T(y)$ and its variance converges to $\frac{\sigma^2}{2\theta} \mathbf{I}$, and the underlying process is an OU process whose stationary distribution is Gaussian, we conclude that $q(\mathbf{x}_t \mid y)$ converges in distribution to $\mathcal{N}(\boldsymbol{\mu}_T(y), \frac{\sigma^2}{2\theta} \mathbf{I})$. □

Proof of Theorem D.5 (Reduced Initial Error). For LABridge, samples at time T are drawn from

$$p_{\text{LAB}}(\mathbf{x}_T \mid y) = \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I}).$$

For a fixed y , denote the expected squared error relative to the target conditional mean $\mathbb{E}[\mathbf{x}_0 \mid y]$ as

$$D_{\text{LAB}}(y) = \mathbb{E}_{\mathbf{x}_T \sim p_{\text{LAB}}(\cdot \mid y)} \|\mathbf{x}_T - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2.$$

We can decompose the term inside the expectation:

$$\mathbf{x}_T - \mathbb{E}[\mathbf{x}_0 \mid y] = (\mathbf{x}_T - \boldsymbol{\mu}_T(y)) + (\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y]).$$

Since $\mathbf{x}_T \sim \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I})$, the term $(\mathbf{x}_T - \boldsymbol{\mu}_T(y))$ has zero mean. When squaring and taking the expectation, the cross-term vanishes:

$$\begin{aligned} D_{\text{LAB}}(y) &= \mathbb{E}_{\mathbf{x}_T \sim p_{\text{LAB}}(\cdot \mid y)} [\|(\mathbf{x}_T - \boldsymbol{\mu}_T(y)) + (\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y])\|_2^2] \\ &= \mathbb{E}_{\mathbf{x}_T \sim p_{\text{LAB}}(\cdot \mid y)} [\|\mathbf{x}_T - \boldsymbol{\mu}_T(y)\|_2^2] + \|\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2 \\ &\quad + 2 \mathbb{E}_{\mathbf{x}_T \sim p_{\text{LAB}}(\cdot \mid y)} [(\mathbf{x}_T - \boldsymbol{\mu}_T(y))^T (\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y])]. \end{aligned}$$

The expectation of the cross term is $\mathbb{E}[(\mathbf{x}_T - \boldsymbol{\mu}_T(y))^T (\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y])] = \mathbf{0}^T (\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y]) = 0$. Noting that $\mathbb{E}\|\mathbf{x}_T - \boldsymbol{\mu}_T(y)\|_2^2 = \text{Tr}(\sigma_T^2 \mathbf{I}) = d \sigma_T^2$ (where d is the dimension), we find:

$$D_{\text{LAB}}(y) = d \sigma_T^2 + \|\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2.$$

For the standard model with prior

$$p_{\text{Std}}(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \sigma_T^2 \mathbf{I}),$$

a similar decomposition for $D_{\text{Std}}(y) = \mathbb{E}_{\mathbf{x}_T \sim p_{\text{Std}}(\cdot)} \|\mathbf{x}_T - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2$ gives:

$$\begin{aligned} D_{\text{Std}}(y) &= \mathbb{E}_{\mathbf{x}_T \sim p_{\text{Std}}(\cdot)} [\|(\mathbf{x}_T - \mathbf{0}) + (\mathbf{0} - \mathbb{E}[\mathbf{x}_0 \mid y])\|_2^2] \\ &= \mathbb{E}_{\mathbf{x}_T \sim p_{\text{Std}}(\cdot)} [\|\mathbf{x}_T\|_2^2] + \|\mathbf{0} - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2 \\ &= d \sigma_T^2 + \|\mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2. \end{aligned}$$

Under Assumption D.2, $\boldsymbol{\mu}_T(y)$ is trained to approximate $\mathbb{E}[\mathbf{x}_0 \mid y]$. Thus, the term $\|\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2$ represents the squared error of this approximation. For a good approximation, this error is smaller than $\|\mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2$, especially when $\mathbb{E}[\mathbf{x}_0 \mid y]$ is significantly non-zero. Specifically, $\|\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2 \leq \|\mathbf{0} - \mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2 = \|\mathbb{E}[\mathbf{x}_0 \mid y]\|_2^2$ if $\boldsymbol{\mu}_T(y)$ is at least as good an estimate

of $\mathbb{E}[\mathbf{x}_0 | y]$ as $\mathbf{0}$ is. Since TIAE optimizes for this, the inequality generally holds. Averaging over y then yields the desired inequality:

$$\mathbb{E}_y[D_{\text{LAB}}(y)] \leq \mathbb{E}_y[D_{\text{Std}}(y)].$$

Equality holds if, for all y , $\|\boldsymbol{\mu}_T(y) - \mathbb{E}[\mathbf{x}_0 | y]\|_2^2 = \|\mathbb{E}[\mathbf{x}_0 | y]\|_2^2$. This implies that $\boldsymbol{\mu}_T(y)$ offers no improvement over $\mathbf{0}$ as an estimate for $\mathbb{E}[\mathbf{x}_0 | y]$. If $\boldsymbol{\mu}_T(y)$ is well-trained, this typically occurs only if $\mathbb{E}[\mathbf{x}_0 | y] = \mathbf{0}$ (and $\boldsymbol{\mu}_T(y) = \mathbf{0}$), or if $\boldsymbol{\mu}_T(y) = -2\mathbb{E}[\mathbf{x}_0 | y]$ which is not the goal of TIAE. Assuming TIAE produces $\boldsymbol{\mu}_T(y)$ close to $\mathbb{E}[\mathbf{x}_0 | y]$, strict inequality holds unless $\mathbb{E}[\mathbf{x}_0 | y] = \mathbf{0}$ (and $\boldsymbol{\mu}_T(y) \approx \mathbf{0}$). \square

Proof of Theorem D.6 (Directed Drift). The LABridge reverse probability flow ODE is given by:

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

Here, $\sigma_t = \sqrt{\frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})}$ is the standard deviation of the noise in the forward OU process kernel $q(\mathbf{x}_t | \mathbf{x}_0, \boldsymbol{\mu}_T)$, and $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)$ approximates the normalized score $-\sigma_t \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | y)$.

In contrast, a standard VP (Variance Preserving) diffusion reverse ODE dynamics (e.g., from DDPM based on a forward process $d\mathbf{x}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t dt + \sqrt{\beta(t)}d\mathbf{w}_t$) is typically given by:

$$\frac{d\mathbf{x}_t}{dt} = -\frac{1}{2}\beta(t)\mathbf{x}_t - \frac{1}{2}\beta(t)\sigma_t^{\text{Std}} \cdot \mathbf{s}_\theta^{\text{Std}}(\mathbf{x}_t, t, y),$$

or, if $\boldsymbol{\epsilon}_\theta^{\text{Std}}$ predicts the noise $\boldsymbol{\epsilon}$ from $\mathbf{x}_t = \alpha_t^{\text{Std}}\mathbf{x}_0 + \sigma_t^{\text{Std}}\boldsymbol{\epsilon}$:

$$\frac{d\mathbf{x}_t}{dt} = \left(-\frac{1}{2}\beta(t)\mathbf{x}_t\right) - \frac{1}{2}\beta(t) \left(-\frac{\boldsymbol{\epsilon}_\theta^{\text{Std}}(\mathbf{x}_t, t, y)}{\sigma_t^{\text{Std}}}\right).$$

The term $\beta(t)$ arises from the specific schedule of the VP SDE. The key drift component from the forward SDE is $-\frac{1}{2}\beta(t)\mathbf{x}_t$.

Comparing the two drift components inherited from their respective forward processes (before considering the learned score terms):

- **LABridge (OU):** $\mathbf{f}_{\text{LAB}}(\mathbf{x}_t, y) = \theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t)$. This term actively pulls the state \mathbf{x}_t towards the text-conditioned prior mean $\boldsymbol{\mu}_T(y)$. By Assumption D.2, $\boldsymbol{\mu}_T(y)$ is aligned with the target data manifold associated with text y .
- **Standard VP:** $\mathbf{f}_{\text{Std}}(\mathbf{x}_t) = -\frac{1}{2}\beta(t)\mathbf{x}_t$. This term always pulls the state \mathbf{x}_t towards the origin $\mathbf{0}$.

The term $\theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t)$ in LABridge provides an explicit, semantically-informed directional component to the drift. Even if the learned score/noise predictor $\boldsymbol{\epsilon}_\theta$ is imperfect or provides weak guidance (especially in early sampling stages when \mathbf{x}_t is far from the data manifold and score estimation is difficult), the OU drift term inherently steers the process towards a region of the latent space ($\boldsymbol{\mu}_T(y)$) that is already aligned with the target semantics.

In contrast, the standard VP drift $-\frac{1}{2}\beta(t)\mathbf{x}_t$ always pulls towards the origin. If the target conditional mean $\mathbb{E}[\mathbf{x}_0 | y]$ is far from the origin, this drift component can be counterproductive, and the entire burden of finding the correct semantic direction falls on the learned score term $\mathbf{s}_\theta^{\text{Std}}$.

Therefore, the explicit OU mean reversion term in LABridge provides stronger, more direct, and semantically informed guidance throughout the reverse sampling process compared to the origin-centric drift of standard VP diffusion, which can contribute to faster convergence to the desired data manifold. This is particularly beneficial at early sampling times when \mathbf{x}_t is far from the data manifold and the score estimate might be less reliable. \square

Proof of Proposition D.7 (Convergence Rate). Define a Lyapunov (energy) function that measures the squared error between the current state \mathbf{x}_t and a target state $\mathbf{x}^*(y)$ corresponding to the desired image latent for a given text y :

$$V(\mathbf{x}_t) = \frac{1}{2}\|\mathbf{x}_t - \mathbf{x}^*(y)\|^2.$$

Our goal is to show that along the reverse diffusion trajectory of LABridge,

$$\frac{d\mathbf{x}_t}{dt} = \theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t) + S(\mathbf{x}_t, t, y),$$

where $S(\mathbf{x}_t, t, y)$ denotes the learned score (or noise) term, the function $V(\mathbf{x}_t)$ decays at an accelerated rate when $\boldsymbol{\mu}_T(y)$ is well-aligned with $\mathbf{x}^*(y)$.

First, compute the time derivative of $V(\mathbf{x}_t)$ using the chain rule:

$$\frac{dV}{dt} = \nabla V(\mathbf{x}_t)^T \frac{d\mathbf{x}_t}{dt}.$$

Since

$$\nabla V(\mathbf{x}_t) = \mathbf{x}_t - \mathbf{x}^*(y),$$

substitute the reverse ODE dynamics into the derivative:

$$\frac{dV}{dt} = (\mathbf{x}_t - \mathbf{x}^*(y))^T \left[\theta(\boldsymbol{\mu}_T(y) - \mathbf{x}_t) + S(\mathbf{x}_t, t, y) \right].$$

This expands as

$$\frac{dV}{dt} = \theta (\mathbf{x}_t - \mathbf{x}^*(y))^T (\boldsymbol{\mu}_T(y) - \mathbf{x}_t) + (\mathbf{x}_t - \mathbf{x}^*(y))^T S(\mathbf{x}_t, t, y).$$

For the first term, note that

$$(\mathbf{x}_t - \mathbf{x}^*(y))^T (\boldsymbol{\mu}_T(y) - \mathbf{x}_t) = -(\mathbf{x}_t - \mathbf{x}^*(y))^T (\mathbf{x}_t - \boldsymbol{\mu}_T(y)).$$

Now, express

$$\mathbf{x}_t - \boldsymbol{\mu}_T(y) = [\mathbf{x}_t - \mathbf{x}^*(y)] + [\mathbf{x}^*(y) - \boldsymbol{\mu}_T(y)].$$

Thus, we have

$$(\mathbf{x}_t - \mathbf{x}^*(y))^T (\boldsymbol{\mu}_T(y) - \mathbf{x}_t) = -\left[\|\mathbf{x}_t - \mathbf{x}^*(y)\|^2 + (\mathbf{x}_t - \mathbf{x}^*(y))^T (\mathbf{x}^*(y) - \boldsymbol{\mu}_T(y)) \right].$$

Substitute this back into the expression for $\frac{dV}{dt}$:

$$\frac{dV}{dt} = -\theta \|\mathbf{x}_t - \mathbf{x}^*(y)\|^2 - \theta (\mathbf{x}_t - \mathbf{x}^*(y))^T (\mathbf{x}^*(y) - \boldsymbol{\mu}_T(y)) + (\mathbf{x}_t - \mathbf{x}^*(y))^T S(\mathbf{x}_t, t, y).$$

Assume that the TIAE achieves good alignment so that

$$\|\mathbf{x}^*(y) - \boldsymbol{\mu}_T(y)\| \leq \epsilon,$$

with a small $\epsilon > 0$. In the ideal case, we may assume $\boldsymbol{\mu}_T(y) = \mathbf{x}^*(y)$; then the second term vanishes:

$$\theta (\mathbf{x}_t - \mathbf{x}^*(y))^T (\mathbf{x}^*(y) - \boldsymbol{\mu}_T(y)) = 0.$$

Thus, in this ideal scenario, the derivative simplifies to

$$\frac{dV}{dt} = -\theta \|\mathbf{x}_t - \mathbf{x}^*(y)\|^2 + (\mathbf{x}_t - \mathbf{x}^*(y))^T S(\mathbf{x}_t, t, y).$$

Next, we assume that the score term $S(\mathbf{x}_t, t, y)$ is designed (or learned) such that it also drives \mathbf{x}_t toward the target $\mathbf{x}^*(y)$. In particular, suppose there exists a constant $\gamma > 0$ such that

$$(\mathbf{x}_t - \mathbf{x}^*(y))^T S(\mathbf{x}_t, t, y) \leq -\gamma \|\mathbf{x}_t - \mathbf{x}^*(y)\|^2.$$

Then, combining the two terms we obtain

$$\frac{dV}{dt} \leq -(\theta + \gamma) \|\mathbf{x}_t - \mathbf{x}^*(y)\|^2.$$

Since

$$V(\mathbf{x}_t) = \frac{1}{2} \|\mathbf{x}_t - \mathbf{x}^*(y)\|^2,$$

the inequality is equivalent to

$$\frac{dV}{dt} \leq -2(\theta + \gamma)V(\mathbf{x}_t).$$

This differential inequality implies that $V(\mathbf{x}_t)$ decays exponentially. By Grönwall’s inequality, we obtain

$$V(\mathbf{x}_t) \leq V(\mathbf{x}_T) \exp \{-2(\theta + \gamma)(T - t)\}.$$

Thus, the error $\|\mathbf{x}_t - \mathbf{x}^*(y)\|^2$ decays exponentially at a rate at least $2(\theta + \gamma)$.

In contrast, consider the standard VP diffusion model whose reverse dynamics typically have a drift term of the form

$$-\frac{1}{2}\beta(t)\mathbf{x}_t + S_{VP}(\mathbf{x}_t, t, y).$$

Without the explicit pull toward $\mathbf{x}^*(y)$ (since the drift here always points toward the origin rather than toward a semantically informed target), the effective convergence rate in the Lyapunov function may be substantially lower than $\theta + \gamma$.

In summary, the additional OU drift term $\theta(\mu_T(y) - \mathbf{x}_t)$ in LABridge provides an extra stabilizing force, contributing an exponential convergence rate of at least $2(\theta + \gamma)$ to the target, which is generally faster than in models lacking such a term. This completes the detailed proof. \square

E.3 Proofs for Stability

Proposition E.2 (Mean-square stability of LABridge). *Consider the LABridge forward SDE*

$$d\mathbf{x}_t = \theta(\mu_T(y) - \mathbf{x}_t) dt + \sigma d\mathbf{w}_t, \quad \theta > 0, \sigma > 0, \quad (26)$$

with \mathbf{w}_t a d -dimensional Wiener process and $\mu_T(y)$ the TIAE mean. Then

$$\sup_{t \geq 0} \mathbb{E}[\|\mathbf{x}_t\|_2^2] \leq 2\|\mu_T(y)\|_2^2 + \frac{d\sigma^2}{\theta}. \quad (27)$$

Proof. Let $g(\mathbf{x}) = \|\mathbf{x}\|_2^2$. Applying Itô’s lemma to (26) gives

$$dg(\mathbf{x}_t) = 2\mathbf{x}_t^\top d\mathbf{x}_t + \text{Tr}(\sigma^2 \mathbf{I}) dt \quad (28)$$

$$= 2\theta \mathbf{x}_t^\top (\mu_T - \mathbf{x}_t) dt + d\sigma^2 dt + 2\sigma \mathbf{x}_t^\top d\mathbf{w}_t. \quad (29)$$

Taking expectations and using the martingale property of the Itô integral,

$$\frac{d}{dt} \mathbb{E}[\|\mathbf{x}_t\|_2^2] = -2\theta \mathbb{E}[\|\mathbf{x}_t - \mu_T\|_2^2] + d\sigma^2. \quad (30)$$

Expanding the squared norm and upper-bounding the mixed term by Cauchy–Schwarz gives the linear differential inequality

$$\dot{m}(t) + 2\theta m(t) \leq 2\theta \|\mu_T\|_2^2 + d\sigma^2, \quad m(t) := \mathbb{E}[\|\mathbf{x}_t\|_2^2].$$

Solving yields

$$m(t) \leq m(0) e^{-2\theta t} + \left(2\|\mu_T\|_2^2 + \frac{d\sigma^2}{\theta}\right)(1 - e^{-2\theta t}),$$

and taking $t \rightarrow \infty$ gives (27). \square

Proposition E.3 (Weaker stability of a zero-mean prior). *Let the VE forward SDE with zero-mean prior be*

$$d\mathbf{z}_t = \sigma(t) d\mathbf{w}_t, \quad \sigma(t) = \sigma_0 e^{\lambda t}, \quad \lambda > 0.$$

Then $\mathbb{E} \|\mathbf{z}_t\|_2^2 = d\sigma_0^2(e^{2\lambda t} - 1)$, which grows exponentially with t . Even in the VP parameterisation $d\mathbf{z}_t = -\frac{1}{2}\beta(t)\mathbf{z}_t dt + \sqrt{\beta(t)} d\mathbf{w}_t$ one obtains $\sup_{t \geq 0} \mathbb{E} \|\mathbf{z}_t\|_2^2 = d$ only when the origin coincides with the true conditional mean $\mathbb{E}[\mathbf{x}_0 | y]$. If that mean is non-zero the reverse process must first undo the displacement, making it more sensitive to score-estimation error.

Proof of Theorem D.8. Combine Proposition E.2 with Proposition E.3. For VE the second moment of the zero-mean model diverges; for VP it equals d whenever the conditional mean is exactly zero and is strictly larger than the bound in (27) as soon as $\|\mu_T(y)\|^2 > \frac{d\sigma^2}{2\theta}$. The assumed lower bound δ therefore guarantees a strict gap. \square

Interpretation. The tight, time–uniform bound (27) is the crucial ingredient: it limits the distance any LABridge trajectory can wander, so that even imperfect scores cannot push the sampler far away from the text-conditioned attractor $\boldsymbol{\mu}_T(y)$. In contrast, standard diffusion with a fixed $\mathcal{N}(\mathbf{0}, \mathbf{I})$ prior either allows the second moment to grow without bound (VE) or always pulls samples toward the *wrong* location (the origin), thereby requiring the learned score to correct a larger drift term. LABridge is therefore *strictly* more stable in the mean-square sense.

E.4 Proof for Modeling Capacity (ELBO Perspective)

Theorem E.4 (Refer to Theorem D.9). *For diffusion models, the ELBO contains a prior matching term of the form*

$$\mathbb{E}_{q(\mathbf{x}_0, y)} \left[\text{KL}(q(\mathbf{x}_T | \mathbf{x}_0, y) \| p(\mathbf{x}_T | y)) \right].$$

Assume that

$$q(\mathbf{x}_T | \mathbf{x}_0, y) = \mathcal{N}(\alpha_T \mathbf{x}_0 + \beta_T \boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I})$$

and that the LABridge prior is

$$p_{\text{LAB}}(\mathbf{x}_T | y) = \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I}).$$

By contrast, a standard model might use a fixed prior

$$p_{\text{Std}}(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \tilde{\sigma}_T^2 \mathbf{I}).$$

If $\boldsymbol{\mu}_T(y)$ is chosen appropriately (via the TIAE), the LABridge setup minimizes the KL divergence compared to the standard model. A lower KL term tightens the ELBO, which enhances the model’s capacity to learn the data distribution.

Proof. The ELBO for a diffusion model (see, e.g., Ho et al. [2020]) involves the term

$$L_{\text{prior}}(y) = \mathbb{E}_{q(\mathbf{x}_0 | y)} \left[\text{KL}(q(\mathbf{x}_T | \mathbf{x}_0, y) \| p(\mathbf{x}_T | y)) \right].$$

For LABridge, the forward process is

$$q(\mathbf{x}_T | \mathbf{x}_0, y) = \mathcal{N}(\alpha_T \mathbf{x}_0 + \beta_T \boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I}),$$

and the prior is chosen as

$$p_{\text{LAB}}(\mathbf{x}_T | y) = \mathcal{N}(\boldsymbol{\mu}_T(y), \sigma_T^2 \mathbf{I}).$$

Since the KL divergence between two Gaussians with equal covariances is

$$\text{KL}(\mathcal{N}(\mu_q, \sigma_T^2 \mathbf{I}) \| \mathcal{N}(\mu_p, \sigma_T^2 \mathbf{I})) = \frac{1}{2\sigma_T^2} \|\mu_q - \mu_p\|_2^2,$$

we obtain

$$L_{\text{prior}}^{\text{LAB}}(y) = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0 | y)} \left[\frac{1}{2\sigma_T^2} \|\alpha_T \mathbf{x}_0 + \beta_T \boldsymbol{\mu}_T(y) - \boldsymbol{\mu}_T(y)\|_2^2 \right].$$

Because $\alpha_T + \beta_T = 1$, this simplifies to

$$L_{\text{prior}}^{\text{LAB}}(y) = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0 | y)} \left[\frac{\alpha_T^2}{2\sigma_T^2} \|\mathbf{x}_0 - \boldsymbol{\mu}_T(y)\|_2^2 \right].$$

By contrast, a standard diffusion model uses a fixed prior,

$$p_{\text{Std}}(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \sigma_T^2 \mathbf{I}),$$

and the corresponding forward process might be

$$q_{\text{Std}}(\mathbf{x}_T | \mathbf{x}_0) = \mathcal{N}(\alpha_T \mathbf{x}_0, \sigma_T^2 \mathbf{I}),$$

so that

$$L_{\text{prior}}^{\text{Std}}(y) = \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0 | y)} \left[\frac{\alpha_T^2}{2\sigma_T^2} \|\mathbf{x}_0\|_2^2 \right].$$

Subtract:

$$L_{\text{prior}}^{\text{Std}} - L_{\text{prior}}^{\text{LAB}} = \frac{\alpha_T^2}{2\sigma_T^2} \mathbb{E}_{(\mathbf{x}_0, \mathbf{y})} [2\mathbf{x}_0^\top \boldsymbol{\mu}_T(\mathbf{y}) - \|\boldsymbol{\mu}_T(\mathbf{y})\|^2] .$$

Since Assumption D.2 ensures $\boldsymbol{\mu}_T(y)$ is close to $\mathbb{E}[\mathbf{x}_0 \mid y]$, the error term $\|\mathbf{x}_0 - \boldsymbol{\mu}_T(y)\|_2^2$ will typically be much smaller than $\|\mathbf{x}_0\|_2^2$, especially when $\mathbb{E}[\mathbf{x}_0 \mid y]$ deviates substantially from zero. Thus, $L_{\text{prior}}^{\text{LAB}}(y)$ is smaller than $L_{\text{prior}}^{\text{Std}}(y)$, which tightens the ELBO and implies improved modeling capacity. \square

Remark E.5. The proofs above clarify that the benefits of LABridge lie in the informed prior provided by the TIAE and the explicit OU drift in the reverse dynamics. These features lead to better text–vision alignment, more efficient sampling, inherent stability, and tighter matching in the ELBO objective.

F Limitations

While LABridge demonstrates promising results in improving text-image alignment and accelerating sampling, several limitations should be acknowledged. The framework’s performance heavily relies on the effectiveness of the Text-Vision Alignment Encoder (TIAE); suboptimal training of the TIAE could lead to poorly aligned priors $\boldsymbol{\mu}_T(y)$, diminishing the benefits of the structured initialization and OU process dynamics. Furthermore, the introduction of the TIAE and the two-stage training process might increase the overall model complexity and computational overhead compared to standard end-to-end diffusion models. The choice of Ornstein-Uhlenbeck parameters (θ, σ) and TIAE loss weights introduces additional hyperparameters requiring careful tuning. Finally, our current analysis primarily focuses on theoretical aspects and initial empirical validation; extensive evaluation across diverse and complex prompt types, comparison under strictly controlled computational budgets, and investigation into potential failure modes (e.g., handling out-of-distribution text concepts) are necessary directions for future work.