A Appendix

A.1 Derivations

In the following subsections, we provide theoretical derivations. In Section A.1.1, we discuss the asymptotic convergence properties as well as the assumptions of score-matching methods. In Section A.1.2, we elaborate on the formulation of EBFlow (i.e., Eqs. (8) and (9)), and provide a explanation of their interpretation. Finally, in Section A.1.3, we present a theoretical analysis of KL divergence and Fisher divergence, and discuss the underlying mechanism behind the proposed MaP technique.

A.1.1 Asymptotic Convergence Property of Score Matching

In this subsection, we provide a formal description of the *consistency* property of score matching. The description follows [16] and the notations are replaced with those used in this paper. The regularity conditions for $p(\cdot\,;\theta)$ are defined in Assumptions A.1$\sim$A.7. In the following paragraph, the parameter space is defined as Θ. In addition, $s(\boldsymbol{x};\theta) \triangleq \frac{\partial}{\partial \boldsymbol{x}} \log p(\boldsymbol{x};\theta) = -\frac{\partial}{\partial \boldsymbol{x}} E(\boldsymbol{x};\theta)$ represents the score function. $\hat{\mathcal{L}}_{\mathrm{SM}}(\theta) \triangleq \frac{1}{N}\sum_{k=1}^{N} f(\mathbf{x}_k;\theta)$ denotes an unbiased estimator of $\mathcal{L}_{\mathrm{SM}}(\theta)$, where $f(\boldsymbol{x};\theta) \triangleq \frac{1}{2}\left\|\frac{\partial}{\partial \boldsymbol{x}} E(\boldsymbol{x};\theta)\right\|^2 - \mathrm{Tr}\left(\frac{\partial^2}{\partial \boldsymbol{x}^2} E(\boldsymbol{x};\theta)\right) = \frac{1}{2}\left\|s(\boldsymbol{x};\theta)\right\|^2 + \mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}} s(\boldsymbol{x};\theta)\right)$ and $\{\mathbf{x}_1,\cdots,\mathbf{x}_N\}$ represents a collection of i.i.d. samples drawn from $p_{\mathbf{x}}$. For notational simplicity, we denote $\partial h(\boldsymbol{x};\theta) \triangleq \frac{\partial}{\partial \boldsymbol{x}} h(\boldsymbol{x};\theta)$ and $\partial_i h_j(\boldsymbol{x};\theta) \triangleq \frac{\partial}{\partial \boldsymbol{x}_i} h_j(\boldsymbol{x};\theta)$, where $h_j(\boldsymbol{x};\theta)$ denotes the j-th element of h.

Assumption A.1. (Positiveness) $p(\boldsymbol{x};\theta) > 0$ and $p_{\mathbf{x}}(\boldsymbol{x}) > 0$, $\forall \theta \in \Theta$, $\forall \boldsymbol{x} \in \mathbb{R}^D$.

Assumption A.2. (Regularity of the score functions) The parameterized score function $s(\boldsymbol{x};\theta)$ and the true score function $\frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{x}}(\boldsymbol{x})$ are both continuous and differentiable. In addition, their expectations $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[s(\boldsymbol{x};\theta)\right]$ and $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{x}}(\boldsymbol{x})\right]$ are finite. (i.e., $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[s(\boldsymbol{x};\theta)\right] < \infty$ and $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{x}}(\boldsymbol{x})\right] < \infty$)

Assumption A.3. (Boundary condition) $\lim_{\|\boldsymbol{x}\|\to\infty} p_{\mathbf{x}}(\boldsymbol{x}) s(\boldsymbol{x};\theta) = 0$, $\forall \theta \in \Theta$.

Assumption A.4. (Compactness) The parameter space Θ is compact.

Assumption A.5. (Identifiability) There exists a set of parameters θ^* such that $p_{\mathbf{x}}(\boldsymbol{x}) = p(\boldsymbol{x};\theta^*)$, where $\theta^* \in \Theta$, $\forall \boldsymbol{x} \in \mathbb{R}^D$.

Assumption A.6. (Uniqueness) $\theta \neq \theta^* \Leftrightarrow p(\boldsymbol{x};\theta) \neq p(\boldsymbol{x};\theta^*)$, where $\theta, \theta^* \in \Theta$, $\boldsymbol{x} \in \mathbb{R}^D$.

Assumption A.7. (Lipschitzness of f) The function f is Lipschitz continuous w.r.t. θ, i.e., $|f(\boldsymbol{x};\theta_1) - f(\boldsymbol{x};\theta_2)| \leq L(\boldsymbol{x})\left\|\theta_1 - \theta_2\right\|_2$, $\forall \theta_1, \theta_2 \in \Theta$, where $L(\boldsymbol{x})$ represents a Lipschitz constant satisfying $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[L(\boldsymbol{x})\right] < \infty$.

Theorem A.8. *(Consistency of a score-matching estimator [16]) The score-matching estimator $\theta_N \triangleq \mathrm{argmin}_{\theta\in\Theta} \hat{\mathcal{L}}_{\mathrm{SM}}$ is consistent, i.e.,*

$$\theta_N \xrightarrow{p} \theta^*, \text{ as } N \to \infty.$$

Assumptions A.1$\sim$A.3 are the conditions that ensure $\frac{\partial}{\partial\theta}\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}}(\boldsymbol{x}) \| p(\boldsymbol{x};\theta)\right] = \frac{\partial}{\partial\theta}\mathcal{L}_{\mathrm{SM}}(\theta)$. Assumptions A.4$\sim$A.7 lead to the uniform convergence property [16] of a score-matching estimator, which gives rise to the *consistency* property. The detailed derivation can be found in Corollary 1 in [16]. In the following Lemma A.9 and Proposition A.10, we examine the sufficient condition for g and $p_{\mathbf{u}}$ to satisfy Assumption A.7.

Lemma A.9. *(Sufficient condition for the Lipschitzness of f) The function $f(\boldsymbol{x};\theta) = \frac{1}{2}\left\|s(\boldsymbol{x};\theta)\right\|^2 + \mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}} s(\boldsymbol{x};\theta)\right)$ is Lipschitz continuous if the score function $s(\boldsymbol{x};\theta)$ satisfies the following conditions: $\forall \theta, \theta_1, \theta_2 \in \Theta$, $\forall i \in \{1,\cdots,D\}$,*

$$\left\|s(\boldsymbol{x};\theta)\right\|_2 \leq L_1(\boldsymbol{x}),$$
$$\left\|s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\right\|_2 \leq L_2(\boldsymbol{x})\left\|\theta_1 - \theta_2\right\|_2,$$
$$\left\|\partial_i s(\boldsymbol{x};\theta_1) - \partial_i s(\boldsymbol{x};\theta_2)\right\|_2 \leq L_3(\boldsymbol{x})\left\|\theta_1 - \theta_2\right\|_2,$$

where L_1, L_2, and L_3 are Lipschitz constants satisfying $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[L_1(\boldsymbol{x})\right] < \infty$, $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[L_2(\boldsymbol{x})\right] < \infty$, and $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[L_3(\boldsymbol{x})\right] < \infty$.

Proof. The Lipschitzness of f can be guaranteed by ensuring the Lipschitzness of $\|s(\boldsymbol{x};\theta)\|_2^2$ and $\mathrm{Tr}\left(\partial s(\boldsymbol{x};\theta)\right)$.

Step 1. (Lipschitzness of $\|s(\boldsymbol{x};\theta)\|_2^2$)

$$
\begin{aligned}
&\left|\|s(\boldsymbol{x};\theta_1)\|_2^2 - \|s(\boldsymbol{x};\theta_2)\|_2^2\right| \\
&= \left|s(\boldsymbol{x};\theta_1)^T s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)^T s(\boldsymbol{x};\theta_2)\right| \\
&= \left|\left(s(\boldsymbol{x};\theta_1)^T s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_1)^T s(\boldsymbol{x};\theta_2)\right) + \left(s(\boldsymbol{x};\theta_1)^T s(\boldsymbol{x};\theta_2) - s(\boldsymbol{x};\theta_2)^T s(\boldsymbol{x};\theta_2)\right)\right| \\
&= \left|s(\boldsymbol{x};\theta_1)^T \left(s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\right) + s(\boldsymbol{x};\theta_2)^T \left(s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\right)\right| \\
&\overset{(i)}{\leq} \left|s(\boldsymbol{x};\theta_1)^T \left(s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\right)\right| + \left|s(\boldsymbol{x};\theta_2)^T \left(s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\right)\right| \\
&\overset{(ii)}{\leq} \|s(\boldsymbol{x};\theta_1)\|_2 \|s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\|_2 + \|s(\boldsymbol{x};\theta_2)\|_2 \|s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\|_2 \\
&\overset{(iii)}{\leq} L_1(\boldsymbol{x}) \|s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\|_2 + L_1(\boldsymbol{x}) \|s(\boldsymbol{x};\theta_1) - s(\boldsymbol{x};\theta_2)\|_2 \\
&\overset{(iii)}{\leq} 2L_1(\boldsymbol{x})L_2(\boldsymbol{x}) \|\theta_1 - \theta_2\|_2\,,
\end{aligned}
$$

where (i) is based on triangle inequality, (ii) is due to Cauchy–Schwarz inequality, and (iii) follows from the listed assumptions.

Step 2. (Lipschitzness of $\mathrm{Tr}\left(\partial s(\boldsymbol{x};\theta)\right)$)

$$
\begin{aligned}
\left|\mathrm{Tr}\left(\partial s(\boldsymbol{x};\theta_1)\right) - \mathrm{Tr}\left(\partial s(\boldsymbol{x};\theta_2)\right)\right| &= \left|\mathrm{Tr}\left(\partial s(\boldsymbol{x};\theta_1) - \partial s(\boldsymbol{x};\theta_2)\right)\right| \\
&\overset{(i)}{\leq} D \|\partial s(\boldsymbol{x};\theta_1) - \partial s(\boldsymbol{x};\theta_2)\|_2 \\
&\overset{(ii)}{\leq} D\sqrt{\sum_i \|\partial_i s(\boldsymbol{x};\theta_1) - \partial_i s(\theta_2)\|_2^2} \\
&\overset{(iii)}{\leq} D\sqrt{DL_3^2(\boldsymbol{x})\|\theta_1 - \theta_2\|_2^2} \\
&= D\sqrt{D}L_3(\boldsymbol{x})\|\theta_1 - \theta_2\|_2
\end{aligned}
$$

where (i) holds by Von Neumann's trace inequality. (ii) is due to the property $\|A\|_2 \leq \sqrt{\sum_i \|\boldsymbol{a}_i\|_2^2}$, where $\boldsymbol{a}_i$ is the column vector of A. (iii) holds by the listed assumptions.

Based on Steps 1 and 2, the Lipschitzness of f is guaranteed, since

$$
\begin{aligned}
|f(\boldsymbol{x};\theta_1) - f(\boldsymbol{x};\theta_2)| &= \left|\frac{1}{2}\|s(\boldsymbol{x};\theta_1)\|^2 + \mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}}s(\boldsymbol{x};\theta_1)\right) - \frac{1}{2}\|s(\boldsymbol{x};\theta_2)\|^2 - \mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}}s(\boldsymbol{x};\theta_2)\right)\right| \\
&= \left|\frac{1}{2}\|s(\boldsymbol{x};\theta_1)\|^2 - \frac{1}{2}\|s(\boldsymbol{x};\theta_2)\|^2 + \mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}}s(\boldsymbol{x};\theta_1)\right) - \mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}}s(\boldsymbol{x};\theta_2)\right)\right| \\
&\leq \frac{1}{2}\left|\|s(\boldsymbol{x};\theta_1)\|^2 - \|s(\boldsymbol{x};\theta_2)\|^2\right| + \left|\mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}}s(\boldsymbol{x};\theta_1)\right) - \mathrm{Tr}\left(\frac{\partial}{\partial \boldsymbol{x}}s(\boldsymbol{x};\theta_2)\right)\right| \\
&\leq L_1(\boldsymbol{x})L_2(\boldsymbol{x})\|\theta_1 - \theta_2\|_2 + D\sqrt{D}L_3(\boldsymbol{x})\|\theta_1 - \theta_2\|_2 \\
&= \left(L_1(\boldsymbol{x})L_2(\boldsymbol{x}) + D\sqrt{D}L_3(\boldsymbol{x})\right)\|\theta_1 - \theta_2\|_2\,.
\end{aligned}
$$

□

Proposition A.10. *(Sufficient condition for the Lipschitzness of f) The function f is Lipschitz continuous if $g(\boldsymbol{x};\theta)$ has bounded first, second, and third-order derivatives, i.e., $\forall i,j \in \{1,\cdots,D\}$, $\forall\theta \in \Theta$.*

$$\|\mathbf{J}_g(\boldsymbol{x};\theta)\|_2 \leq l_1(\boldsymbol{x}), \|\partial_i \mathbf{J}_g(\boldsymbol{x};\theta)\|_2 \leq l_2(\boldsymbol{x}), \|\partial_i\partial_j \mathbf{J}_g(\boldsymbol{x};\theta)\|_2 \leq l_3(\boldsymbol{x}),$$

and smooth enough on Θ, i.e., $\theta_1,\theta_2 \in \Theta$:

$$\|g(\boldsymbol{x};\theta_1) - g(\boldsymbol{x};\theta_2)\|_2 \leq r_0(\boldsymbol{x})\|\theta_1 - \theta_2\|_2\,,$$

$$\|\mathbf{J}_g(\boldsymbol{x};\theta_1)-\mathbf{J}_g(\boldsymbol{x};\theta_2)\|_2 \le r_1(\boldsymbol{x})\,\|\theta_1-\theta_2\|_2\,,$$
$$\|\partial_i\mathbf{J}_g(\boldsymbol{x};\theta_1)-\partial_i\mathbf{J}_g(\boldsymbol{x};\theta_2)\|_2 \le r_2(\boldsymbol{x})\,\|\theta_1-\theta_2\|_2\,.$$
$$\|\partial_i\partial_j\mathbf{J}_g(\boldsymbol{x};\theta_1)-\partial_i\partial_j\mathbf{J}_g(\boldsymbol{x};\theta_2)\|_2 \le r_3(\boldsymbol{x})\,\|\theta_1-\theta_2\|_2\,.$$

In addition, it satisfies the following conditions:

$$\left\|\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\right\|_2 \le l_1^{'}(\boldsymbol{x}),\ \left\|\partial_i\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\right\|_2 \le l_2^{'}(\boldsymbol{x}),$$
$$\left\|\mathbf{J}_g^{-1}(\boldsymbol{x};\theta_1)-\mathbf{J}_g^{-1}(\boldsymbol{x};\theta_2)\right\|_2 \le r_1^{'}(\boldsymbol{x})\,\|\theta_1-\theta_2\|_2\,,$$
$$\left\|\partial_i\mathbf{J}_g^{-1}(\boldsymbol{x};\theta_1)-\partial_i\mathbf{J}_g^{-1}(\boldsymbol{x};\theta_2)\right\|_2 \le r_2^{'}(\boldsymbol{x})\,\|\theta_1-\theta_2\|_2\,,$$

where $\mathbf{J}_g^{-1}$ represents the inverse matrix of $\mathbf{J}_g$. Furthermore, the prior distribution $p_{\mathbf{u}}$ satisfies:

$$\|s_{\mathbf{u}}(\boldsymbol{u})\| \le t_1,\ \|\partial_i s_{\mathbf{u}}(\boldsymbol{u})\| \le t_2$$
$$\|s_{\mathbf{u}}(\boldsymbol{u}_1)-s_{\mathbf{u}}(\boldsymbol{u}_2)\|_2 \le t_3\,\|\boldsymbol{u}_1-\boldsymbol{u}_2\|_2\,,$$
$$\|\partial_i s_{\mathbf{u}}(\boldsymbol{u}_1)-\partial_i s_{\mathbf{u}}(\boldsymbol{u}_2)\|_2 \le t_4\,\|\boldsymbol{u}_1-\boldsymbol{u}_2\|_2\,,$$

where $s_{\mathbf{u}}(\boldsymbol{u}) \triangleq \frac{\partial}{\partial \boldsymbol{u}}\log p_{\mathbf{u}}(\boldsymbol{u})$ is the score function of $p_{\mathbf{u}}$. The Lipschitz constants listed above (i.e., $l_1 \sim l_3$, $r_0 \sim r_3$, $l_1^{'} \sim l_2^{'}$, and $r_1^{'} \sim r_2^{'}$) have finite expectations.

Proof. We show that the sufficient conditions stated in Lemma A.9 can be satisfied using the conditions listed above.

Step 1. (Sufficient condition of $\|s(\boldsymbol{x};\theta)\|_2 \le L_1(\boldsymbol{x})$)

Since $\|s(\boldsymbol{x};\theta)\|_2 = \left\|\frac{\partial}{\partial \boldsymbol{x}}\log p_{\mathbf{u}}(g(\boldsymbol{x};\theta)) + \frac{\partial}{\partial \boldsymbol{x}}\log|\det \mathbf{J}_g(\boldsymbol{x};\theta)|\right\|_2 \le \left\|\frac{\partial}{\partial \boldsymbol{x}}\log p_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right\|_2 + \left\|\frac{\partial}{\partial \boldsymbol{x}}\log|\det \mathbf{J}_g(\boldsymbol{x};\theta)|\right\|_2$, we first demonstrate that $\left\|\frac{\partial}{\partial \boldsymbol{x}}\log p_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right\|_2$ and $\left\|\frac{\partial}{\partial \boldsymbol{x}}\log|\det \mathbf{J}_g(\boldsymbol{x};\theta)|\right\|_2$ are both bounded.

(1.1) $\left\|\frac{\partial}{\partial \boldsymbol{x}}\log p_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right\|_2$ is bounded:

$$\left\|\frac{\partial}{\partial \boldsymbol{x}}\log p_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right\|_2 = \left\|\left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T\mathbf{J}_g(\boldsymbol{x};\theta)\right\|_2 \le \|s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\|_2\,\|\mathbf{J}_g(\boldsymbol{x};\theta)\|_2 \le t_1 l_1(\boldsymbol{x}).$$

(1.2) $\left\|\frac{\partial}{\partial \boldsymbol{x}}\log|\det \mathbf{J}_g(\boldsymbol{x};\theta)|\right\|$ is bounded:

$$\begin{aligned}
\left\|\frac{\partial}{\partial \boldsymbol{x}}\log|\det \mathbf{J}_g(\boldsymbol{x};\theta)|\right\| &= \left\||\det \mathbf{J}_g(\boldsymbol{x};\theta)|^{-1}\frac{\partial}{\partial \boldsymbol{x}}|\det \mathbf{J}_g(\boldsymbol{x};\theta)|\right\| \\
&= \left\|(\det \mathbf{J}_g(\boldsymbol{x};\theta))^{-1}\frac{\partial}{\partial \boldsymbol{x}}\det \mathbf{J}_g(\boldsymbol{x};\theta)\right\| \\
&\overset{(i)}{=} \left\|(\det \mathbf{J}_g(\boldsymbol{x};\theta))^{-1}\det \mathbf{J}_g(\boldsymbol{x};\theta)\boldsymbol{v}(\boldsymbol{x};\theta)\right\| \\
&= \|\boldsymbol{v}(\boldsymbol{x};\theta)\|\,,
\end{aligned}$$

where (i) is derived using Jacobi's formula, and $\boldsymbol{v}_i(\boldsymbol{x};\theta) = \mathrm{Tr}\left(\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\partial_i\mathbf{J}_g(\boldsymbol{x};\theta)\right)$.

$$\begin{aligned}
\|\boldsymbol{v}(\boldsymbol{x};\theta)\| &= \sqrt{\sum_i\left(\mathrm{Tr}\left(\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\partial_i\mathbf{J}_g(\boldsymbol{x};\theta)\right)\right)^2} \\
&\overset{(i)}{\le} \sqrt{\sum_i D^2\left\|\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\partial_i\mathbf{J}_g(\boldsymbol{x};\theta)\right\|_2^2} \\
&\overset{(ii)}{\le} \sqrt{\sum_i D^2\left\|\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\right\|_2^2\|\partial_i\mathbf{J}_g(\boldsymbol{x};\theta)\|_2^2} \\
&\overset{(iii)}{\le} \sqrt{\sum_i D^2 l_1^{'2}(\boldsymbol{x}) l_2^2(\boldsymbol{x})} \\
&= \sqrt{D^3}\,l_1^{'}(\boldsymbol{x})l_2(\boldsymbol{x}),
\end{aligned}$$

where (i) holds by Von Neumann's trace inequality, (ii) is due to the property of matrix norm, and (iii) is follows from the listed assumptions.

Step 2. (Sufficient condition of the Lipschitzness of $s(\boldsymbol{x};\theta)$)

Since $s(\boldsymbol{x};\theta) = \frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{u}}(g(\boldsymbol{x};\theta)) + \frac{\partial}{\partial \boldsymbol{x}} \log |\det \mathbf{J}_g(\boldsymbol{x};\theta)|$, we demonstrate that $\frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{u}}(g(\boldsymbol{x};\theta))$ and $\frac{\partial}{\partial \boldsymbol{x}} \log |\det \mathbf{J}_g(\boldsymbol{x};\theta)|$ are both Lipschitz continuous on Θ.

(2.1) Lipschitzness of $\frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{u}}(g(\boldsymbol{x};\theta))$:

$$
\begin{aligned}
&\left\| \frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{u}}(g(\boldsymbol{x};\theta_1)) - \frac{\partial}{\partial \boldsymbol{x}} \log p_{\mathbf{u}}(g(\boldsymbol{x};\theta_2)) \right\|_2 \\
&= \left\| \left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta_1))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta_1) - \left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta_2))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta_2) \right\|_2 \\
&\overset{(i)}{\leq} \|s_{\mathbf{u}}(g(\boldsymbol{x};\theta_1))\|_2 \|\mathbf{J}_g(\boldsymbol{x};\theta_1) - \mathbf{J}_g(\boldsymbol{x};\theta_2)\|_2 + \|s_{\mathbf{u}}(g(\boldsymbol{x};\theta_1)) - s_{\mathbf{u}}(g(\boldsymbol{x};\theta_2))\|_2 \|\mathbf{J}_g(\boldsymbol{x};\theta_2)\|_2 \\
&\overset{(ii)}{\leq} t_1 r_1(\boldsymbol{x}) \|\theta_1 - \theta_2\|_2 + t_2 l_1(\boldsymbol{x}) \|g(\boldsymbol{x};\theta_1) - g(\boldsymbol{x};\theta_2)\|_2 \\
&\overset{(ii)}{\leq} t_1 r_1(\boldsymbol{x}) \|\theta_1 - \theta_2\|_2 + t_2 l_1(\boldsymbol{x}) r_0(\boldsymbol{x}) \|\theta_1 - \theta_2\|_2 \\
&= \left(t_1 r_1(\boldsymbol{x}) + t_2 l_1(\boldsymbol{x}) r_0(\boldsymbol{x})\right) \|\theta_1 - \theta_2\|_2 \,,
\end{aligned}
$$

where (i) is obtained using a similar derivation to Step 1 in Lemma A.9, while (ii) follows from the listed assumptions.

(2.2) Lipschitzness of $\frac{\partial}{\partial \boldsymbol{x}} \log |\det \mathbf{J}_g(\boldsymbol{x};\theta)|$:

Let $\mathbf{M}(i,\boldsymbol{x};\theta) \triangleq \mathbf{J}_g^{-1}(\boldsymbol{x};\theta_1)\partial_i \mathbf{J}_g(\boldsymbol{x};\theta)$. We first demonstrate that $\mathbf{M}$ is Lipschitz continuous:

$$
\begin{aligned}
&\|\mathbf{M}(i,\boldsymbol{x};\theta_1) - \mathbf{M}(i,\boldsymbol{x};\theta_2)\|_2 \\
&= \left\| \mathbf{J}_g^{-1}(\boldsymbol{x};\theta_1)\partial_i \mathbf{J}_g(\boldsymbol{x};\theta_1) - \mathbf{J}_g^{-1}(\boldsymbol{x};\theta_2)\partial_i \mathbf{J}_g(\boldsymbol{x};\theta_2) \right\|_2 \\
&\overset{(i)}{\leq} \left\| \mathbf{J}_g^{-1}(\boldsymbol{x};\theta_1) \right\|_2 \|(\partial_i \mathbf{J}_g(\boldsymbol{x};\theta_1) - \partial_i \mathbf{J}_g(\boldsymbol{x};\theta_2))\|_2 + \left\| \mathbf{J}_g^{-1}(\boldsymbol{x};\theta_1) - \mathbf{J}_g^{-1}(\boldsymbol{x};\theta_2) \right\|_2 \|\partial_i \mathbf{J}_g(\boldsymbol{x};\theta_2)\|_2 \\
&\overset{(ii)}{\leq} l_1^{'}(\boldsymbol{x}) r_2(\boldsymbol{x}) \|\theta_1 - \theta_2\|_2 + l_2(\boldsymbol{x}) r_1^{'}(\boldsymbol{x}) \|\theta_1 - \theta_2\|_2 \\
&= \left(l_1^{'}(\boldsymbol{x}) r_2(\boldsymbol{x}) + l_2(\boldsymbol{x}) r_1^{'}(\boldsymbol{x}) \right) \|\theta_1 - \theta_2\|_2 \,,
\end{aligned}
$$

where (i) is obtained by an analogous derivation of the step 1 in Lemma A.9, and (ii) holds by the listed assumption.

The Lipschitzness of $\mathbf{M}$ leads to the Lipschitzness of $\frac{\partial}{\partial \boldsymbol{x}} \log |\det \mathbf{J}_g(\boldsymbol{x};\theta)|$, since:

$$
\begin{aligned}
&\left\| \frac{\partial}{\partial \boldsymbol{x}} \log |\det \mathbf{J}_g(\boldsymbol{x};\theta_1)| - \frac{\partial}{\partial \boldsymbol{x}} \log |\det \mathbf{J}_g(\boldsymbol{x};\theta_2)| \right\|_2 \\
&= \|\boldsymbol{v}(\boldsymbol{x};\theta_1) - \boldsymbol{v}(\boldsymbol{x};\theta_2)\|_2 \\
&= \sqrt{\sum_i \left(\mathrm{Tr}\left(\mathbf{M}(i,\boldsymbol{x};\theta_1)\right) - \mathrm{Tr}\left(\mathbf{M}(i,\boldsymbol{x};\theta_2)\right)\right)^2} \\
&= \sqrt{\sum_i \left(\mathrm{Tr}\left(\mathbf{M}(i,\boldsymbol{x};\theta_1) - \mathbf{M}(i,\boldsymbol{x};\theta_2)\right)\right)^2} \\
&\overset{(i)}{\leq} \sqrt{\sum_i D^2 \|\mathbf{M}(i,\boldsymbol{x};\theta_1) - \mathbf{M}(i,\boldsymbol{x};\theta_2)\|_2^2} \\
&\overset{(ii)}{\leq} \sqrt{\sum_i D^2 \left(l_1^{'}(\boldsymbol{x}) r_2(\boldsymbol{x}) + l_2(\boldsymbol{x}) r_1^{'}(\boldsymbol{x}) \right)^2 \|\theta_1 - \theta_2\|_2^2} \\
&= \sqrt{D^3} \left(l_1^{'}(\boldsymbol{x}) r_2(\boldsymbol{x}) + l_2(\boldsymbol{x}) r_1^{'}(\boldsymbol{x}) \right) \|\theta_1 - \theta_2\|_2 \,,
\end{aligned}
$$

where (i) holds by Von Neumann's trace inequality, (ii) is due to the Lipschitzness of $\mathbf{M}$.

Step 3. (Sufficient condition of the Lipschitzness of $\partial_i s(\boldsymbol{x};\theta)$)

$\partial_i s(\boldsymbol{x};\theta)$ can be decomposed as $\left(\partial_i s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta)$, $\left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T \partial_i \mathbf{J}_g(\boldsymbol{x};\theta)$, and $\partial_i\left[\boldsymbol{v}(\boldsymbol{x};\theta)\right]$ as follows:

$$\begin{aligned}
\partial_i s(\boldsymbol{x};\theta) &= \partial_i\left[\left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta)\right] + \partial_i\left[\boldsymbol{v}(\boldsymbol{x};\theta)\right] \\
&= \left[\left(\partial_i s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta)\right] + \left[\left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T \partial_i \mathbf{J}_g(\boldsymbol{x};\theta)\right] + \partial_i\left[\boldsymbol{v}(\boldsymbol{x};\theta)\right].
\end{aligned}$$

(3.1) The Lipschitzness of $\left(\partial_i s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta)$ and $\left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta))\right)^T \partial_i \mathbf{J}_g(\boldsymbol{x};\theta)$ can be derived using proofs similar to that in Step 2.1:

$$\left\|\left(\partial_i s_{\mathbf{u}}(g(\boldsymbol{x};\theta_1))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta_1) - \left(\partial_i s_{\mathbf{u}}(g(\boldsymbol{x};\theta_2))\right)^T \mathbf{J}_g(\boldsymbol{x};\theta_2)\right\|_2 \leq \left(t_2 r_1(\boldsymbol{x}) + t_4 r_0(\boldsymbol{x}) l_1(\boldsymbol{x})\right)\|\theta_1-\theta_2\|_2,$$

$$\left\|\left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta_1))\right)^T \partial_i \mathbf{J}_g(\boldsymbol{x};\theta_1) - \left(s_{\mathbf{u}}(g(\boldsymbol{x};\theta_2))\right)^T \partial_i \mathbf{J}_g(\boldsymbol{x};\theta_2)\right\|_2 \leq \left(t_1 r_2(\boldsymbol{x}) + t_3 r_0(\boldsymbol{x}) l_2(\boldsymbol{x})\right)\|\theta_1-\theta_2\|_2.$$

(3.2) Lipschitzness of $\partial_i\left[\boldsymbol{v}(\boldsymbol{x};\theta)\right]$:

Let $\partial_i\left[\boldsymbol{v}_j(\boldsymbol{x};\theta)\right] \triangleq \partial_i \mathrm{Tr}\left(\mathbf{M}(j,\boldsymbol{x};\theta)\right) = \mathrm{Tr}\left(\partial_i \mathbf{M}(j,\boldsymbol{x};\theta)\right)$. We first show that $\partial_i \mathbf{M}(j,\boldsymbol{x};\theta)$ can be decomposed as:

$$\partial_i \mathbf{M}(j,\boldsymbol{x};\theta) = \partial_i\left(\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\partial_j \mathbf{J}_g(\boldsymbol{x};\theta)\right) = \left(\partial_i \mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\partial_j \mathbf{J}_g(\boldsymbol{x};\theta)\right) + \left(\mathbf{J}_g^{-1}(\boldsymbol{x};\theta)\partial_i\partial_j \mathbf{J}_g(\boldsymbol{x};\theta)\right)$$

The Lipschitz constant of $\partial_i \mathbf{M}$ equals to $\left(l_2'(\boldsymbol{x}) r_2(\boldsymbol{x}) + l_2(\boldsymbol{x}) r_2'(\boldsymbol{x})\right) + \left(l_1'(\boldsymbol{x}) r_3(\boldsymbol{x}) + l_3(\boldsymbol{x}) r_1'(\boldsymbol{x})\right)$ based on a similar derivation as in Step 3.1. The Lipschitzness of $\partial_i \mathbf{M}(j,\boldsymbol{x};\theta)$ leads to the Lipschitzness of $\partial_i\left[\boldsymbol{v}(\boldsymbol{x};\theta)\right]$:

$$\begin{aligned}
&\left\|\partial_i\left[\boldsymbol{v}(\boldsymbol{x};\theta_1)\right] - \partial_i\left[\boldsymbol{v}(\boldsymbol{x};\theta_2)\right]\right\|_2 \\
&= \sqrt{\sum_j \left(\mathrm{Tr}\left(\partial_i \mathbf{M}(j,\boldsymbol{x};\theta_1)\right) - \mathrm{Tr}\left(\partial_i \mathbf{M}(j,\boldsymbol{x};\theta_2)\right)\right)^2} \\
&= \sqrt{\sum_j \mathrm{Tr}\left(\partial_i \mathbf{M}(j,\boldsymbol{x};\theta_1) - \partial_i \mathbf{M}(j,\boldsymbol{x};\theta_2)\right)^2} \\
&\overset{(i)}{\leq} \sqrt{\sum_j D^2\left\|\partial_i \mathbf{M}(j,\boldsymbol{x};\theta_1) - \partial_i \mathbf{M}(j,\boldsymbol{x};\theta_2)\right\|_2^2} \\
&\overset{(ii)}{\leq} \sqrt{\sum_j D^2\left(l_2'(\boldsymbol{x}) r_2(\boldsymbol{x}) + l_2(\boldsymbol{x}) r_2'(\boldsymbol{x}) + l_1'(\boldsymbol{x}) r_3(\boldsymbol{x}) + l_3(\boldsymbol{x}) r_1'(\boldsymbol{x})\right)^2 \|\theta_1-\theta_2)\|_2^2} \\
&= \sqrt{D^3}\left(l_2'(\boldsymbol{x}) r_2(\boldsymbol{x}) + l_2(\boldsymbol{x}) r_2'(\boldsymbol{x}) + l_1'(\boldsymbol{x}) r_3(\boldsymbol{x}) + l_3(\boldsymbol{x}) r_1'(\boldsymbol{x})\right)\|\theta_1-\theta_2)\|_2
\end{aligned}$$

where (i) holds by Von Neumann's trace inequality, (ii) is due to the Lipschitzness of $\partial_i \mathbf{M}$. $\square$

A.1.2 Derivation of Eqs. (8) and (9)

Energy-based models are formulated based on the observation that any continuous pdf $p(\boldsymbol{x};\theta)$ can be expressed as a Boltzmann distribution $\exp\left(-E(\boldsymbol{x};\theta)\right) Z^{-1}(\theta)$ [13], where the energy function $E(\cdot\,;\theta)$ can be modeled as any scalar-valued continuous function. In EBFlow, the energy function $E(\boldsymbol{x};\theta)$ is selected as $-\log(p_{\mathbf{u}}\left(g(\boldsymbol{x};\theta)\right)\prod_{g_i\in\mathcal{S}_n}|\det(\mathbf{J}_{g_i}(\boldsymbol{x}_{i-1};\theta))|)$ according to Eq. (9). This suggests that the normalizing constant $Z(\theta) = \int \exp\left(-E(\boldsymbol{x};\theta)\right) d\boldsymbol{x}$ is equal to $(\prod_{g_i\in\mathcal{S}_l}|\det(\mathbf{J}_{g_i}(\theta))|)^{-1}$ according to Lemma A.11.

Lemma A.11.

$$\left(\prod_{g_i\in\mathcal{S}_l}|\det(\mathbf{J}_{g_i}(\theta))|\right)^{-1} = \int_{\boldsymbol{x}\in\mathbb{R}^D} p_{\mathbf{u}}\left(g(\boldsymbol{x};\theta)\right)\prod_{g_i\in\mathcal{S}_n}|\det(\mathbf{J}_{g_i}(\boldsymbol{x}_{i-1};\theta))|\, d\boldsymbol{x}. \tag{A1}$$

Proof.

$$
\begin{aligned}
1 &= \int_{\boldsymbol{x}\in\mathbb{R}^D} p(\boldsymbol{x};\theta)d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}\in\mathbb{R}^D} p_{\mathbf{u}}\left(g(\boldsymbol{x};\theta)\right) \prod_{g_i\in\mathcal{S}_n} \left|\det(\mathbf{J}_{g_j}(\boldsymbol{x}_{i-1};\theta))\right| \prod_{g_i\in\mathcal{S}_l} \left|\det(\mathbf{J}_{g_i}(\theta))\right| d\boldsymbol{x} \\
&= \prod_{g_i\in\mathcal{S}_l} \left|\det(\mathbf{J}_{g_i}(\theta))\right| \int_{\boldsymbol{x}\in\mathbb{R}^D} p_{\mathbf{u}}\left(g(\boldsymbol{x};\theta)\right) \prod_{g_i\in\mathcal{S}_n} \left|\det(\mathbf{J}_{g_i}(\boldsymbol{x}_{i-1};\theta))\right| d\boldsymbol{x}
\end{aligned}
$$

By multiplying $\left(\prod_{g_i\in\mathcal{S}_l} \left|\det(\mathbf{J}_{g_i}(\theta))\right|\right)^{-1}$ to both sides of the equation, we arrive at the conclusion:

$$
\left(\prod_{g_i\in\mathcal{S}_l} \left|\det(\mathbf{J}_{g_i}(\theta))\right|\right)^{-1} = \int_{\boldsymbol{x}\in\mathbb{R}^D} p_{\mathbf{u}}\left(g(\boldsymbol{x};\theta)\right) \prod_{g_i\in\mathcal{S}_n} \left|\det(\mathbf{J}_{g_i}(\boldsymbol{x}_{i-1};\theta))\right| d\boldsymbol{x}.
$$

□

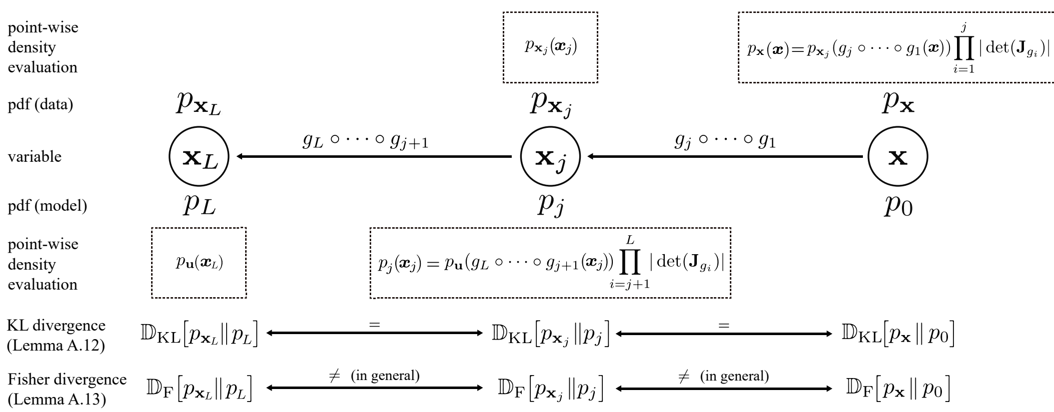

Figure A1: An illustration of the relationship between the variables discussed in Proposition 4.1, Lemma A.12, and Lemma A.13. $\mathbf{x}$ represents a random vector sampled from the data distribution $p_{\mathbf{x}}$. $\{g_i\}_{i=1}^{L}$ is a series of transformations. $\mathbf{x}_j \triangleq g_j \circ \cdots \circ g_1(\mathbf{x})$, and $p_{\mathbf{x}_j}$ is its pdf. $p_j(\boldsymbol{x}_j) = p_{\mathbf{u}}(g_L \circ \cdots \circ g_{j+1}(\boldsymbol{x}_j)) \prod_{i=j+1}^{L} |\det(\mathbf{J}_{g_i})|$, where $p_{\mathbf{u}}$ is a prior distribution. The properties of KL divergence and Fisher divergence presented in the last two rows are derived in Lemmas A.12 and A.13.

A.1.3 Theoretical Analyses of KL Divergence and Fisher Divergence

In this section, we provide formal derivations for Proposition 4.1, Lemma A.12, and Lemma A.13. To ensure a clear presentation, we provide a visualization of the relationship between the variables used in the subsequent derivations in Fig. A1.

Lemma A.12. *Let $p_{\mathbf{x}_j}$ be the pdf of the latent variable of $\mathbf{x}_j \triangleq g_j \circ \cdots \circ g_1(\mathbf{x})$ indexed by j. In addition, let $p_j(\cdot)$ be a pdf modeled as $p_{\mathbf{u}}(g_L \circ \cdots \circ g_{j+1}(\cdot)) \prod_{i=j+1}^{L} |\det(\mathbf{J}_{g_i})|$, where $j \in \{0, \cdots, L-1\}$. It follows that:*

$$
\mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}_j} \| p_j\right] = \mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}} \| p_0\right], \forall j \in \{1, \cdots, L-1\}. \tag{A2}
$$

Proof. The equivalence $\mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}}\|p_0\right]=\mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}_j}\|p_j\right]$ holds for any $j\in\{1,\cdots,L-1\}$ since:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}}\|p_0\right]\\
&=\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\log\left(\frac{p_{\mathbf{x}}(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right)\right]\\
&=\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\log\left(\frac{p_{\mathbf{x}_j}(g_j\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=1}^{j}|\det(\mathbf{J}_{g_i})|}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right]\\
&=\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\log\left(\frac{p_{\mathbf{x}_j}(g_j\circ\cdots\circ g_1(\boldsymbol{x}))}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right]\\
&\overset{(i)}{=}\mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_{j+1}(\boldsymbol{x}_j))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right]\\
&=\mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}_j}\|p_j\right],
\end{aligned}$$

where (i) is due to the property that $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}[f\circ g_j\circ\cdots\circ g_1(\boldsymbol{x})]=\mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}[f(\boldsymbol{x}_j)]$ for a given function f. Therefore, $\mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}_j}\|p_j\right]=\mathbb{D}_{\mathrm{KL}}\left[p_{\mathbf{x}}\|p_0\right]$, $\forall j\in\{1,\cdots,L-1\}$. □

Lemma A.13. *Let $p_{\mathbf{x}_j}$ be the pdf of the latent variable of $\mathbf{x}_j\triangleq g_j\circ\cdots\circ g_1(\mathbf{x})$ indexed by j. In addition, let $p_j(\cdot)$ be a pdf modeled as $p_{\mathbf{u}}(g_L\circ\cdots\circ g_{j+1}(\cdot))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|$, where $j\in\{0,\cdots,L-1\}$. It follows that:*

$$\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}}\|p_0\right]=\mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial\boldsymbol{x}_j}\log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\right)\prod_{i=1}^{j}\mathbf{J}_{g_i}\right\|^2\right],\forall j\in\{1,\cdots,L-1\}. \quad \text{(A3)}$$

Proof. Based on the definition, the Fisher divergence between $p_{\mathbf{x}}$ and p_0 is written as:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}}\|p_0\right]\\
&=\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\frac{1}{2}\left\|\frac{\partial}{\partial\boldsymbol{x}}\log\left(\frac{p_{\mathbf{x}}(\boldsymbol{x})}{p_0(\boldsymbol{x})}\right)\right\|^2\right]\\
&=\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\frac{1}{2}\left\|\frac{\partial}{\partial\boldsymbol{x}}\log\left(\frac{p_{\mathbf{x}_j}(g_j\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=1}^{j}|\det(\mathbf{J}_{g_i})|}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right\|^2\right]\\
&=\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\frac{1}{2}\left\|\frac{\partial}{\partial\boldsymbol{x}}\log\left(\frac{p_{\mathbf{x}_j}(g_j\circ\cdots\circ g_1(\boldsymbol{x}))}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right\|^2\right]\\
&=\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial g_j\circ\cdots\circ g_1(\boldsymbol{x})}\log\left(\frac{p_{\mathbf{x}_j}(g_j\circ\cdots\circ g_1(\boldsymbol{x}))}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right)\frac{\partial g_j\circ\cdots\circ g_1(\boldsymbol{x})}{\partial\boldsymbol{x}}\right\|^2\right]\\
&\overset{(i)}{=}\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial g_j\circ\cdots\circ g_1(\boldsymbol{x})}\log\left(\frac{p_{\mathbf{x}_j}(g_j\circ\cdots\circ g_1(\boldsymbol{x}))}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_1(\boldsymbol{x}))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right)\prod_{i=1}^{j}\mathbf{J}_{g_i}\right\|^2\right]\\
&\overset{(ii)}{=}\mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial\boldsymbol{x}_j}\log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_{\mathbf{u}}(g_L\circ\cdots\circ g_{j+1}(\boldsymbol{x}_j))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|}\right)\right)\prod_{i=1}^{j}\mathbf{J}_{g_i}\right\|^2\right],\\
&=\mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial\boldsymbol{x}_j}\log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\right)\prod_{i=1}^{j}\mathbf{J}_{g_i}\right\|^2\right],
\end{aligned}$$

where (i) is due to the chain rule, and (ii) is because $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}[f\circ g_j\circ\cdots\circ g_1(\boldsymbol{x})]=\mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}[f(\boldsymbol{x}_j)]$ for a given function f. □

Remark A.14. Lemma A.13 implies that $\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}_j} \| p_j\right] \neq \mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}} \| p_0\right]$ in general, as the latter contains an additional multiplier $\prod_{i=1}^{j} \mathbf{J}_{g_i}$ as shown below:

$$\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}} \| p_0\right] = \mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial \boldsymbol{x}_j} \log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\right)\prod_{i=1}^{j}\mathbf{J}_{g_i}\right\|^2\right],$$

$$\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}_j} \| p_j\right] = \mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial \boldsymbol{x}_j} \log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\right)\right\|^2\right].$$

Proposition 4.1. *Let $p_{\mathbf{x}_j}$ be the pdf of the latent variable of $\mathbf{x}_j \triangleq g_j \circ \cdots \circ g_1(\mathbf{x})$ indexed by j. In addition, let $p_j(\cdot)$ be a pdf modeled as $p_{\mathbf{u}}(g_L \circ \cdots \circ g_{j+1}(\cdot))\prod_{i=j+1}^{L}|\det(\mathbf{J}_{g_i})|$, where $j \in \{0, \cdots, L-1\}$. It follows that:*

$$\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}_j} \| p_j\right] = 0 \Leftrightarrow \mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}} \| p_0\right] = 0, \forall j \in \{1, \cdots, L-1\}. \tag{A4}$$

Proof. Based on Remark A.14, the following holds:

$$\begin{aligned}
&\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}_j} \| p_j\right] = \mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\frac{1}{2}\left\|\frac{\partial}{\partial \boldsymbol{x}_j} \log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\right\|^2\right] = 0 \\
\overset{(i)}{\Leftrightarrow} &\left\|\frac{\partial}{\partial \boldsymbol{x}_j} \log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\right\|^2 = 0 \\
\overset{(ii)}{\Leftrightarrow} &\left\|\frac{\partial}{\partial \boldsymbol{x}_j} \log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\prod_{i=1}^{j}\mathbf{J}_{g_i}\right\|^2 = 0 \\
\overset{(i)}{\Leftrightarrow} &\mathbb{D}_{\mathrm{F}}\left[p_{\mathbf{x}} \| p_0\right] = \mathbb{E}_{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}\left[\frac{1}{2}\left\|\left(\frac{\partial}{\partial \boldsymbol{x}_j} \log\left(\frac{p_{\mathbf{x}_j}(\boldsymbol{x}_j)}{p_j(\boldsymbol{x}_j)}\right)\right)\prod_{i=1}^{j}\mathbf{J}_{g_i}\right\|^2\right] = 0,
\end{aligned}$$

where (i) and (ii) both result from the positiveness condition presented in Assumption A.1. Specifically, for (i), $p_{\mathbf{x}_j}(\boldsymbol{x}_j) = p_{\mathbf{x}}(g_1^{-1} \circ \cdots \circ g_j^{-1}(\boldsymbol{x}_j))\prod_{i=1}^{j}\left|\det\left(\mathbf{J}_{g_i^{-1}}\right)\right| > 0$, since $p_{\mathbf{x}} > 0$ and $\prod_{i=1}^{j}\left|\det\left(\mathbf{J}_{g_i^{-1}}\right)\right| = \prod_{i=1}^{j}|\det(\mathbf{J}_{g_i})|^{-1} > 0$. Meanwhile (ii) holds since $\prod_{i=1}^{j}|\det(\mathbf{J}_{g_i})| > 0$ and thus all of the singular values of $\prod_{i=1}^{j}\mathbf{J}_{g_i}$ are non-zero. □

A.2 Experimental Setups

In this section, we elaborate on the experimental setups and provide the detailed configurations for the experiments presented in Section 5 of the main manuscript. The code implementation for the experiments is provided in the following repository: `https://github.com/chen-hao-chao/ebflow`. Our code implementation is developed based on [7, 17, 44].

A.2.1 Experimental Setups for the Two-Dimensional Synthetic Datasets

Datasets. In Section 5.1, we present the experimental results on three two-dimensional synthetic datasets: Sine, Swirl, and Checkerboard. The Sine dataset is generated by sampling data points from the set $\{(4w-2, \sin(12w-6)) \,|\, w \in [0,1]\}$. The Swirl dataset is generated by sampling data points from the set $\{(-\pi\sqrt{w}\cos(\pi\sqrt{w}), \pi\sqrt{w}\sin(\pi\sqrt{w})) \,|\, w \in [0,1]\}$. The Checkerboard dataset is generated by sampling data points from the set $\{(4w-2, t-2s+\lfloor 4w-2 \rfloor \bmod 2) \,|\, w \in [0,1],\, t \in [0,1],\, s \in \{0,1\}\}$, where $\lfloor\cdot\rfloor$ is a floor function, and mod represents the modulo operation.

To establish $p_{\mathbf{x}}$ for all three datasets, we smooth a Dirac function using a Gaussian kernel. Specifically, we define the Dirac function as $\hat{p}(\hat{\boldsymbol{x}}) \triangleq \frac{1}{M}\sum_{i=1}^{M}\delta(\|\hat{\boldsymbol{x}} - \hat{\boldsymbol{x}}^{(i)}\|)$, where $\{\hat{\boldsymbol{x}}^{(i)}\}_{i=1}^{M}$ are M uniformly-sampled data points. The data distribution is defined as $p_{\mathbf{x}}(\boldsymbol{x}) \triangleq \int \hat{p}(\hat{\boldsymbol{x}})\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{x}}, \hat{\sigma}^2\boldsymbol{I})d\hat{\boldsymbol{x}} = \frac{1}{M}\sum_{i=1}^{M}\mathcal{N}(\boldsymbol{x}|\hat{\boldsymbol{x}}^{(i)}, \hat{\sigma}^2\boldsymbol{I})$. The closed-form expressions for $p_{\mathbf{x}}(\boldsymbol{x})$ and $\frac{\partial}{\partial \boldsymbol{x}}\log p_{\mathbf{x}}(\boldsymbol{x})$ can be obtained using the derivation in [45]. In the experiments, M is set as $50{,}000$, and $\hat{\sigma}$ is fixed at 0.375 for all three datasets.

Implementation Details. The model architecture of $g(\cdot\,;\theta)$ consists of ten Glow blocks [4]. Each block comprises an actnorm [4] layer, a fully-connected layer, and an affine coupling layer. Table A2 provides the formal definitions of these operations. $p_{\mathbf{u}}(\cdot)$ is implemented as an isotropic Gaussian with zero mean and unit variance. To determine the best hyperparameters, we perform a grid search over the following optimizers, learning rates, and gradient clipping values based on the evaluation results in terms of the KL divergence. The optimizers include Adam [46], AdamW [47], and RMSProp. The learning rate and gradient clipping values are selected from (5e-3, 1e-3, 5e-4, 1e-4) and (None, 2.5, 10.0), respectively. Table A1 summarizes the selected hyperparameters. The optimization processes of Sine and Swirl datasets require 50,000 training iterations for convergence, while that of the Checkerboard dataset requires 100,000 iterations. The batch size is fixed at 5,000 for all setups.

A.2.2 Experimental Setups for the Real-world Datasets

Datasets. The experiments presented in Section 5.2 are performed on the MNIST [19] and CIFAR-10 [37] datasets. The training and test sets of MNIST and CIFAR-10 contain 50,000 and 10,000 images, respectively. The data are smoothed using the uniform dequantization method presented in [1]. The observable parts (i.e., $\boldsymbol{x}_O$) of the images in Fig. 5 are produced using the pre-trained model in [48].

Implementation Details. In Sections 5.2 and 5.4, we adopt three types of model architectures: FC-based [7], CNN-based, and Glow [4] models. The FC-based model contains two fully-connected layers and a smoothed leaky ReLU non-linearity [7] in between, which is identical to [7]. The CNN-based model consists of three convolutional blocks and two squeezing operations [2] between every convolutional block. Each convolutional block contains two convolutional layers and a smoothed leaky ReLU in between. The Glow model adopted in Section 5.4 is composed of 16 Glow blocks. Each of the Glow block consists of an actnorm [4] layer, a convolutional layer, and an affine coupling layer. The squeezing operation is inserted between every eight blocks. The operations used in these models are summarized in Table A2. The smoothness factor α of Smooth Leaky ReLU is set to 0.3 and 0.6 for models trained on MNIST and CIFAR-10, respectively. The scaling and transition functions $s(\cdot\,;\theta)$ and $t(\cdot\,;\theta)$ of the affine coupling layers are convolutional blocks with ReLU activation functions. The prior distribution $p_{\mathbf{u}}(\cdot)$ is implemented as an isotropic Gaussian with zero mean and unit variance. The FC-based and CNN-based models are trained with RMSProp using a learning rate initialized at 1e-4 and a batch size of 100. The Glow model is trained with an Adam optimizer using a learning rate initialized at 1e-4 and a batch size of 100. The gradient clipping value is set to 500 during the training for the Glow model. The learning rate scheduler `MultiStepLR` in `PyTorch` is used for gradually decreasing the learning rates. The hyper-parameters $\{\sigma, \xi\}$ used in DSM and FDSSM are selected based on a grid search over $\{0.05, 0.1, 0.5, 1.0\}$. The selected $\{\sigma, \xi\}$ are $\{1.0, 1.0\}$ and $\{0.1, 0.1\}$ for the MNIST and CIFAR-10 datasets, respectively. The parameter m in EMA is set to 0.999. The algorithms are implemented using `PyTorch` [39]. The gradients w.r.t. $\boldsymbol{x}$ and θ are both calculated using automatic differential tools [40] provided by `PyTorch` [39]. The runtime is evaluated on Tesla V100 NVIDIA GPUs. In the experiments performed on CIFAR-10 and CelebA using score-matching methods, the energy function (i.e., $\mathbb{E}_{p_{\mathbf{x}}(\boldsymbol{x})}\left[E(\boldsymbol{x};\theta)\right]$) is added as a regularization loss with a balancing factor fixed at 0.001 during the optimization processes. The results in Fig. 2 (b) are smoothed with the exponential moving average function used in `Tensorboard` [49], i.e., $w \times d_{i-1} + (1-w) \times d_i$, where w is set to 0.45 and d_i represents the evaluation result at the i-th iteration.

Table A1: The hyper-parameters used in the two-dimensional synthetic example in Section 5.1.

Dataset		ML	SML	SSM	DSM	FDSSM
Sine	Optimizer	Adam	AdamW	Adam	Adam	Adam
	Learning Rate	5e-4	5e-4	1e-4	1e-4	1e-4
	Gradient Clip	1.0	None	1.0	1.0	1.0
Swirl	Optimizer	Adam	Adam	Adam	Adam	Adam
	Learning Rate	5e-3	1e-4	1e-4	1e-4	1e-4
	Gradient Clip	None	10.0	10.0	10.0	2.5
Checkerboard	Optimizer	AdamW	AdamW	AdamW	AdamW	Adam
	Learning Rate	1e-4	1e-4	1e-4	1e-4	1e-4
	Gradient Clip	10.0	10.0	10.0	10.0	10.0

Table A2: The components of $g(\cdot\,;\theta)$ used in this paper. In this table, $\boldsymbol{z}$ and $\boldsymbol{y}$ are the output and the input of a layer, respectively. β and γ represent the mean and variance of an actnorm layer. $\boldsymbol{w}$ is a convolutional kernel, and $\boldsymbol{w} \star \boldsymbol{y} \triangleq \hat{\boldsymbol{W}}\boldsymbol{y}$, where $\star$ is a convolutional operator, and $\hat{\boldsymbol{W}}$ is a $D \times D$ matrix. $\boldsymbol{W}$ and $\boldsymbol{b}$ represent the weight and bias in a fully-connected layer. α is a hyper-parameter for adjusting the smoothness of smooth leaky ReLU. In the affine coupling layer, $\boldsymbol{z}$ and $\boldsymbol{y}$ are split into two parts $\{\boldsymbol{z}_a, \boldsymbol{z}_b\}$ and $\{\boldsymbol{y}_a, \boldsymbol{y}_b\}$, respectively. $s(\cdot;\theta)$ and $t(\cdot;\theta)$ are the scaling and transition networks parameterized with θ. $\mathrm{sig}\,(y) = 1/(1+\exp{(-y)})$ represents the sigmoid function. $\dim{(\cdot)}$ represents the dimension of the input vector. $\boldsymbol{y}_{[i]}$ represents the i-th element of vector $\boldsymbol{y}$.

Layer	Function	Log Jacobian Determinant	Set
actnorm [4]	$\boldsymbol{z} = (\boldsymbol{y} - \beta)/\gamma$	$\sum_{i=1}^{D} \log \left\|1/\gamma_{[i]}\right\|$	$\mathcal{S}_l$
convolutional	$\boldsymbol{z} = \boldsymbol{w} \star \boldsymbol{y} + \boldsymbol{b}$	$\log \left\|\det\left(\hat{\boldsymbol{W}}\right)\right\|$	$\mathcal{S}_l$
fully-connected	$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{y} + \boldsymbol{b}$	$\log \left\|\det\left(\boldsymbol{W}\right)\right\|$	$\mathcal{S}_l$
smooth leaky ReLU [7]	$\boldsymbol{z} = \alpha \boldsymbol{y} + (1-\alpha)\log(1+\exp{(\boldsymbol{y})})$	$\sum_{i=1}^{D} \log \left\|\alpha + (1-\alpha)\mathrm{sig}\left(\boldsymbol{y}_{[i]}\right)\right\|$	$\mathcal{S}_n$
affine coupling [4]	$\boldsymbol{z}_a = s(\boldsymbol{y}_b;\theta)\boldsymbol{y}_a + t(\boldsymbol{y}_b;\theta)$, $\boldsymbol{z}_b = \boldsymbol{y}_b$	$\sum_{i=1}^{\dim(\boldsymbol{y}_b)} \log \left\|s(\boldsymbol{y}_b;\theta)_{[i]}\right\|$	$\mathcal{S}_n$

Table A3: The simulation results of Eq. (A6). The error rate is measured by $|d_{\text{true}} - d_{\text{est}}|/|d_{\text{true}}|$, where d_{true} and d_{est} represent the true and estimated Jacobian determinants, respectively.

	D =50	D =100	D =200
Error Rate (M =50)	0.004211	0.099940	0.355314
Error Rate (M =100)	0.003503	0.034608	0.076239
Error Rate (M =200)	**0.002332**	**0.015411**	**0.011175**

Results of the Related Works. The results of the relative gradient [7], SSM [16], and FDSSM [17] methods are directly obtained from their original paper. On the other hand, the results of the DSM method is obtained from [17]. Please note that the reported results of [16] and [17] differ from each other given that they both adopt the NICE [1] model. Specifically, the SSM method achieves NLL$= 3,355$ and NLL$= 6,234$ in [16] and [17], respectively. Moreover, the DSM method achieves NLL$= 4,363$ and NLL$= 3,398$ in [16] and [17], respectively. In Table 4, we report the results with lower NLL.

A.3 Estimating the Jacobian Determinants using Importance Sampling

Importance sampling is a technique used to estimate integrals, which can be employed to approximate the normalizing constant $Z(\theta)$ in an energy-based model. In this method, a pdf q with a simple closed form that can be easily sampled from is selected. The normalizing constant can then be expressed as the following formula:

$$
\begin{aligned}
Z(\theta) = \int_{\boldsymbol{x}\in\mathbb{R}^D} \exp{(-E(\boldsymbol{x};\theta))}\, d\boldsymbol{x} &= \int_{\boldsymbol{x}\in\mathbb{R}^D} q(\boldsymbol{x}) \frac{\exp{(-E(\boldsymbol{x};\theta))}}{q(\boldsymbol{x})} d\boldsymbol{x} \\
&= \mathbb{E}_{q(\boldsymbol{x})}\left[\frac{\exp{(-E(\boldsymbol{x};\theta))}}{q(\boldsymbol{x})}\right] \approx \frac{1}{M}\sum_{j=1}^{M} \frac{\exp\left(-E(\hat{\boldsymbol{x}}^{(j)};\theta)\right)}{q(\hat{\boldsymbol{x}}^{(j)})},
\end{aligned}
\tag{A5}
$$

where $\{\hat{\boldsymbol{x}}^{(j)}\}_{j=1}^{M}$ represents M i.i.d. samples drawn from q. According to Lemma A.11, the Jacobian determinants of the layers in $\mathcal{S}_l$ can be approximated using Eq. (A5) as follows:

$$
\left(\prod_{g_i\in\mathcal{S}_l} |\det(\mathbf{J}_{g_i}(\theta))|\right)^{-1} \approx \frac{1}{M}\sum_{j=1}^{M} \frac{p_{\mathbf{u}}\left(g(\hat{\boldsymbol{x}}^{(j)};\theta)\right)\prod_{g_i\in\mathcal{S}_n}\left|\det(\mathbf{J}_{g_i}(\hat{\boldsymbol{x}}_{i-1}^{(j)};\theta))\right|}{q(\hat{\boldsymbol{x}}^{(j)})}. \tag{A6}
$$

Table A4: An overall comparison between EBFlow, the baseline method, the Relative Gradient method [7], and the methods that utilize specially designed linear layers [8–12, 29]. The notations ✓/ ✗ in row 'Unbiased' represent whether the models are optimized according to an unbiased target. On the other hand, the notations ✓/ ✗ in row 'Unconstrained' represent whether the models can be constructed with arbitrary linear transformations. $^{(\dagger)}$ The approximation errors $o(\xi)$ of FDSSM is controlled by its hyper-parameter ξ. $^{(\ddagger)}$ The error $o(\boldsymbol{W})$ of the Relative Gradient method is determined by the values of a model's weights.

	KL-Divergence-Based				Fisher-Divergence-Based		
	Baseline (ML)	EBFlow (SML)	Relative Grad.	Special Linear	EBFlow (SSM)	EBFlow (DSM)	EBFlow (FDSSM)
Complexity	$\mathcal{O}(D^3L)$	$\mathcal{O}(D^3L)$	$\mathcal{O}(D^2L)$	$\mathcal{O}(D^2L)$	$\mathcal{O}(D^2L)$	$\mathcal{O}(D^2L)$	$\mathcal{O}(D^2L)$
Unbiased	✓	✓	✗$^{(\ddagger)}$	✓	✓	✗	✗$^{(\dagger)}$
Unconstrained	✓	✓	✓	✗	✓	✓	✓

To validate this idea, we provide a simple simulation with $p_{\mathbf{u}} = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $q = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $g(\boldsymbol{x}; \boldsymbol{W}) = \boldsymbol{W}\boldsymbol{x}$, $M = \{50, 100, 200\}$, and $D = \{50, 100, 200\}$ in Table A3. The results show that larger values of M lead to more accurate estimation of the Jacobian determinants. Typically, the choice of q is crucial to the accuracy of importance sampling. To obtain an accurate approximation, one can adopt the technique of annealed importance sampling (AIS) [33] or Reverse AIS Estimator (RAISE) [34], which are commonly-adopted algorithms for effectively estimating $Z(\theta)$.

Eq. (A6) can be interpreted as a generalization of the stochastic estimator presented in [50], where the distributions $p_{\mathbf{u}}$ and q are modeled as isotropic Gaussian distributions, and g is restricted as a linear transformation. For the further analysis of this concept, particularly in the context of determinant estimation for matrices, we refer readers to Section I of [50], where a more sophisticated approximation approach and the corresponding experimental findings are provided.

A.4 A Comparison among the Methods Discussed in this Paper

In Sections 2, 3, and 4, we discuss various methods for efficiently training flow-based models. To provide a comprehensive comparison of these methods, we summarize their complexity and characteristics in Table A4.

A.5 The Impacts of the Constraint of Linear Transformations on the Performance of a Flow-based Model

In this section, we examine the impact of the constraints of linear transformations on the performance of a flow-based model. A key distinction between constrained and unconstrained linear layers lies

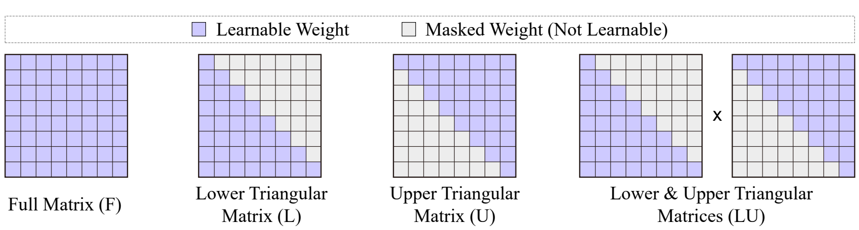

Figure A2: An illustration of the weight matrices in the F, L, U, and LU layers described in Section A.5.

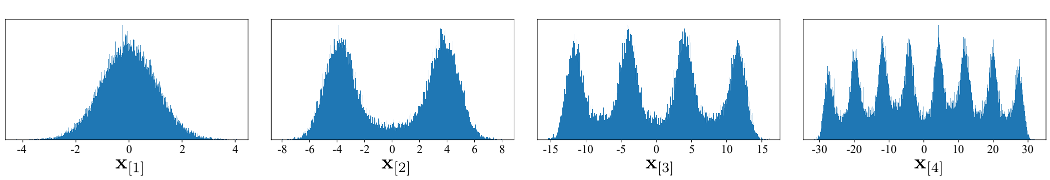

Figure A3: Visualized marginal distributions of $p_{\mathbf{x}_{[i]}}$ for $i = 1, 2, 3$, and 4.

in how they model the correlation between each element in a data vector. Constrained linear transformations, such as those used in the previous works [8–12, 29], impose predetermined correlations that are not learnable during the optimization process. For instance, masked linear layers [8–10] are constructed by masking either the upper or lower triangular weight matrix in a linear layer. In contrast, unconstrained linear layers have weight matrices that are fully learnable, making them more flexible than their constrained counterparts.

To demonstrate the influences of the constraint on the expressiveness of a model, we provide a performance comparison between flow-based models constructed using different types of linear layers. Specifically, we compare the performance of the models constructed using linear layers with full matrices, lower triangular matrices, upper triangular matrices, and matrices that are the multiplication of both lower and upper triangular matrices. These four types of linear layers are hereafter denoted as F, L, U, and LU, respectively, and the differences between them are depicted in Fig. A2. Furthermore, to highlight the performance discrepancy between these models, we construct the target distribution $p_{\mathbf{x}}$ based on an autoregressive relationship of data vector $\mathbf{x}$. Let $\mathbf{x}_{[i]}$ denote the i-th element of $\mathbf{x}$, and $p_{\mathbf{x}_{[i]}}$ represent its associated pdf. $\mathbf{x}_{[i]}$ is constructed based on the following equation:

$$\mathbf{x}_{[i]} = \begin{cases} \mathbf{u}_{[0]} & \text{if } i = 1, \\ \tanh(\mathbf{u}_{[i]} \times s) \times (\mathbf{x}_{[i-1]} + d \times 2^i), & \text{if } i \in \{2, \ldots, D\}, \end{cases} \tag{A7}$$

where $\mathbf{u}$ is sampled from an isotropic Gaussian, and s and d are coefficients controlling the shape and distance between each mode, respectively. In Eq. (A7), the function $\tanh(\cdot)$ can be intuitively viewed as a smoothed variant of the function $2H(\cdot) - 1$, where $H(\cdot)$ represents the Heaviside step function. In this context, the values of $(\mathbf{x}_{[i-1]} + d \times 2^i)$ are multiplied by a value close to either -1 or 1, effectively transforming a positive number to a negative one. Fig. A3 depicts a number of examples of $p_{\mathbf{x}_{[i]}}$ constructed using this method. By employing this approach to design $p_{\mathbf{x}}$, where capturing $p_{\mathbf{x}_{[i]}}$ is presumed to be more challenging than modeling $p_{\mathbf{x}_{[j]}}$ for any $j < i$, we can inspect how the applied constraints impact performance. Inappropriately masking the linear layers, like the U-type layer, is anticipated to result in degraded performance, similar to the *anti-casual* effect explained in [51].

In this experiment, we constructed flow-based models using the smoothed leakyReLU activation and different types of linear layers (i.e., F, L, U, and LU) with a dimensionality of $D = 10$. The models are optimized according to Eq. (2). The performance of these models is evaluated in terms of NLL, and its trends are depicted in Fig. A4. It is observed that the flow-based model built with the F-type layers achieved the lowest NLL, indicating the advantage of using unconstrained weight matrices in linear layers. In addition, there is a noticeable performance discrepancy between models with the L-type and U-type layers, indicating that imposing inappropriate constraints on linear layers may negatively affect the modeling abilities of flow-based models. Furthermore, even when both L-type and U-type layers were adopted, as shown in the red curve in Fig. A4, the performance remains inferior to those using the F-type layers. This experimental evidence suggests that linear layers constructed based on matrix decomposition (e.g., [4, 9]) may not possess the same expressiveness as unconstrained linear layers.

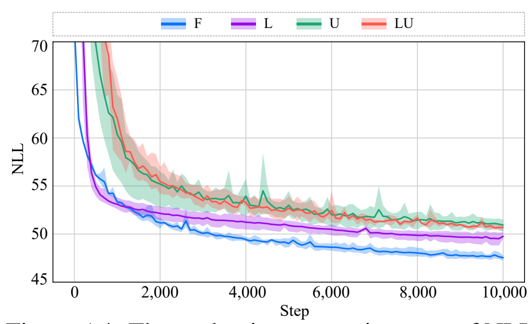

Figure A4: The evaluation curves in terms of NLL of the flow-based models constructed with the F-type, L-type, U-type, and LU-type layers. The curves and shaded area depict the mean and 95% confidence interval of three independent runs.

A.6 Limitations and Discussions

We noticed that score-matching methods sometimes exhibit difficulty in differentiating the weights between individual modes within a multi-modal distribution. This deficiency is illustrated in Fig. A5 (a), where EBFlow fails to accurately capture the density of the Checkerboard dataset. This phenomenon bears resemblance to the *blindness* problem discussed in [52]. While the solution proposed in [52] has the potential to address this issue, their approach is not directly applicable to the flow-based architectures employed in this paper.

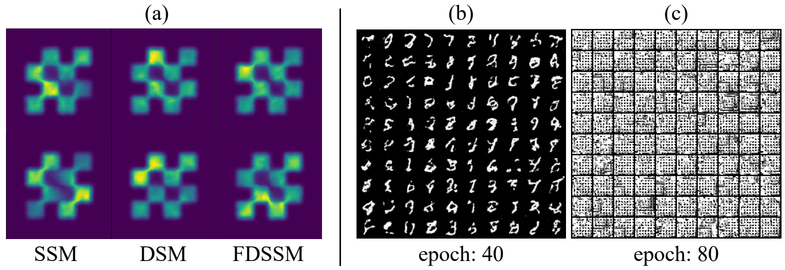

Figure A5: (a) Visualized examples of EBFlow trained with SSM, DSM, and FDSSM on the Checkerboard dataset. (b) The samples generated by the Glow model at the 40-th training epoch. (c) The samples generated by the Glow model at the 80-th training epoch.

In addition, we observed that the sampling quality of EBFlow occasionally experiences a significant reduction during the training iterations. This phenomenon is illustrated in Fig. A5 (b) and (c), where the Glow model trained using our approach demonstrates a decline in performance with extended training periods. The underlying cause of this phenomenon remains unclear, and we consider it a potential avenue for future investigation.