Dimensionality-Varying Diffusion Process
– *Supplementary Material* –

Overview

This supplementary material is organized as follows. First, we will give proofs and derivations in Sec. A. In Sec. B, we will give more details on the implementation of DVDP. In Sec. C, we will show that our DVDP can be further accelerated with DDIM sampling method [5]. Then, we will demonstrate the compatibility with latent diffusion models [4] in Sec. D. Finally, additional comparisons with subspace diffusion [2] will be given in Sec. E.

A. Proofs and Derivations

A.1. Details on the Forward Transition Kernel

Here we prove that the marginal distributions of each forward ADP given by Eqs. (5, 6) can be derived from the forward transition kernel defined by Eq. (7). The proof uses the following basic property of Gaussians

$$\mathbf{z}_1 \sim \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\Sigma}_1),\ \mathbf{z}_2|\mathbf{z}_1 \sim \mathcal{N}(\boldsymbol{A}\mathbf{z}_1; \boldsymbol{\Sigma}_2) \Rightarrow \mathbf{z}_2 \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\mu}; \boldsymbol{A}\boldsymbol{\Sigma}_1\boldsymbol{A}^T + \boldsymbol{\Sigma}_2). \tag{S1}$$

As a prerequisite, we first re-write the Gaussian transition kernel given by Eq. (7) as

$$\mathbf{x}_t^k|\mathbf{x}_{t-1}^k \sim \mathcal{N}(\boldsymbol{U}_k\boldsymbol{\Lambda}_{k,t}\boldsymbol{U}_k^T\mathbf{x}_{t-1}^k; \boldsymbol{U}_k\boldsymbol{L}_{k,t}^2\boldsymbol{U}_k^T),\ 1 \le t \le T,\ 0 \le k \le K, \tag{S2}$$

where

$$\begin{aligned} \boldsymbol{\Lambda}_{k,t} &= \bar{\boldsymbol{\Lambda}}_{k,t-1}^{-1}\bar{\boldsymbol{\Lambda}}_{k,t}, \\ \boldsymbol{L}_{k,t} &= (\bar{\boldsymbol{L}}_{k,t}^2 - \boldsymbol{\Lambda}_{k,t}^2\bar{\boldsymbol{L}}_{k,t-1}^2)^{1/2}. \end{aligned} \tag{S3}$$

The marginal distributions given by Eqs. (5, 6) can also be re-written as

$$\mathbf{x}_t^k|\mathbf{x}_0^k \sim \mathcal{N}(\boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t}\boldsymbol{U}_k^T\mathbf{x}_0^k; \boldsymbol{U}_k\bar{\boldsymbol{L}}_{k,t}^2\boldsymbol{U}_k^T),\ 1 \le t \le T,\ 0 \le k \le K. \tag{S4}$$

With Eqs. (S1) and (S2), we can prove Eq. (S4) by induction:

1. For $t = 1$, $\mathbf{x}_1^k|\mathbf{x}_0^k \sim \mathcal{N}(\boldsymbol{U}_k\boldsymbol{\Lambda}_{k,1}\boldsymbol{U}_k^T\mathbf{x}_0^k; \boldsymbol{U}_k\boldsymbol{L}_{k,1}^2\boldsymbol{U}_k^T)$ is directly defined by Eq. (S2). It satisfies Eq. (S4) since $\boldsymbol{\Lambda}_{k,1} = \bar{\boldsymbol{\Lambda}}_{k,1}$ and $\boldsymbol{L}_{k,1} = \bar{\boldsymbol{L}}_{k,1}$.
2. Suppose $\mathbf{x}_t^k|\mathbf{x}_0^k$ satisfies Eq. (S4). With the definition of $\mathbf{x}_{t+1}^k|\mathbf{x}_t^k$ given by Eq. (S2) and the property Eq. (S1), $\mathbf{x}_{t+1}^k|\mathbf{x}_0^k$ can be derived as
$$\begin{aligned} \mathbf{x}_{t+1}^k|\mathbf{x}_0^k \sim& \mathcal{N}(\boldsymbol{U}_k\boldsymbol{\Lambda}_{k,t+1}\bar{\boldsymbol{\Lambda}}_{k,t}\boldsymbol{U}_k^T\mathbf{x}_0^k; \boldsymbol{U}_k(\boldsymbol{\Lambda}_{k,t+1}^2\bar{\boldsymbol{L}}_{k,t}^2 + \boldsymbol{L}_{k,t+1}^2)\boldsymbol{U}_k^T) \\ =& \mathcal{N}(\boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t+1}\boldsymbol{U}_k^T\mathbf{x}_0^k; \boldsymbol{U}_k\bar{\boldsymbol{L}}_{k,t+1}^2\boldsymbol{U}_k^T), \end{aligned} \tag{S5}$$
where the equality is due to $\bar{\boldsymbol{\Lambda}}_{k,t+1} = \boldsymbol{\Lambda}_{k,t+1}\bar{\boldsymbol{\Lambda}}_{k,t}$ and $\bar{\boldsymbol{L}}_{k,t+1}^2 = \boldsymbol{\Lambda}_{k,t+1}^2\bar{\boldsymbol{L}}_{k,t}^2 + \boldsymbol{L}_{k,t+1}^2$ derived from Eq. (S3). Thus, $\mathbf{x}_{t+1}^k|\mathbf{x}_0^k$ also satisfies Eq. (S4).

Thus, the proof is completed.

A.2. Derivation of $q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k, \mathbf{x}_0^k)$

Here we derive $q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k, \mathbf{x}_0^k)$ from the marginal distribution given by Eqs. (5, 6) and the forward transition kernel given by Eq. (7).

By the Bayes' theorem, $q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k, \mathbf{x}_0^k) \propto q(\mathbf{x}_{t-1}^k|\mathbf{x}_0^k)q(\mathbf{x}_t^k|\mathbf{x}_{t-1}^k, \mathbf{x}_0^k) = q(\mathbf{x}_{t-1}^k|\mathbf{x}_0^k)q(\mathbf{x}_t^k|\mathbf{x}_{t-1}^k)$, where the equality holds because of the Markovian property of $\mathbf{x}_0^k \to \mathbf{x}_1^k \cdots \to \mathbf{x}_T^k$. With $q(\mathbf{x}_{t-1}^k|\mathbf{x}_0^k)$ and $q(\mathbf{x}_t^k|\mathbf{x}_{t-1}^k)$ given by Eqs. (5) to (7), we have

$$
\begin{aligned}
\log q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k, \mathbf{x}_0^k) &= \log q(\mathbf{x}_{t-1}^k|\mathbf{x}_0^k) + \log q(\mathbf{x}_t^k|\mathbf{x}_{t-1}^k) + C_1 \\
&= -\frac{1}{2}(\mathbf{x}_{t-1}^k - \boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t-1}\boldsymbol{U}_k^T\mathbf{x}_0^k)^T\boldsymbol{U}_k\bar{\boldsymbol{L}}_{k,t-1}^{-2}\boldsymbol{U}_k^T(\mathbf{x}_{t-1}^k - \boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t-1}\boldsymbol{U}_k^T\mathbf{x}_0^k) \\
&\quad -\frac{1}{2}(\mathbf{x}_t^k - \boldsymbol{U}_k\boldsymbol{\Lambda}_{k,t}\boldsymbol{U}_k^T\mathbf{x}_{t-1}^k)^T\boldsymbol{U}_k\boldsymbol{L}_{k,t}^{-2}\boldsymbol{U}_k^T(\mathbf{x}_t^k - \boldsymbol{U}_k\boldsymbol{\Lambda}_{k,t}\boldsymbol{U}_k^T\mathbf{x}_{t-1}^k) + C_2 \\
&= -\frac{1}{2}\big[{\mathbf{x}_{t-1}^k}^T\boldsymbol{U}_k(\bar{\boldsymbol{L}}_{k,t-1}^{-2} + \boldsymbol{\Lambda}_{k,t}^2\boldsymbol{L}_{k,t}^{-2})\boldsymbol{U}_k^T\mathbf{x}_{t-1}^k \\
&\quad - 2(\boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t-1}\bar{\boldsymbol{L}}_{k,t-1}^{-2}\boldsymbol{U}_k^T\mathbf{x}_0^k + \boldsymbol{U}_k\boldsymbol{\Lambda}_{k,t}\boldsymbol{L}_{k,t}^{-2}\boldsymbol{U}_k^T\mathbf{x}_t^k)^T\mathbf{x}_{t-1}^k\big] + C_3 \\
&= -\frac{1}{2}\big[{\mathbf{x}_{t-1}^k}^T\boldsymbol{U}_k\boldsymbol{L}_{k,t}^{-2}\bar{\boldsymbol{L}}_{k,t-1}^{-2}\bar{\boldsymbol{L}}_{k,t}^{2}\boldsymbol{U}_k^T\mathbf{x}_{t-1}^k \\
&\quad - 2(\boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t-1}\bar{\boldsymbol{L}}_{k,t-1}^{-2}\boldsymbol{U}_k^T\mathbf{x}_0^k + \boldsymbol{U}_k\boldsymbol{\Lambda}_{k,t}\boldsymbol{L}_{k,t}^{-2}\boldsymbol{U}_k^T\mathbf{x}_t^k)^T\mathbf{x}_{t-1}^k\big] + C_3 \\
&= -\frac{1}{2}(\mathbf{x}_{t-1}^k - \tilde{\boldsymbol{\mu}}_{k,t})^T\tilde{\boldsymbol{\Sigma}}_{k,t}^{-2}(\mathbf{x}_{t-1}^k - \tilde{\boldsymbol{\mu}}_{k,t}) + C_4,
\end{aligned}
\tag{S6}
$$

where C_1, C_2, C_3 and C_4 are constants that do not depend on $\mathbf{x}_{t-1}^k$, and

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}}_{k,t} =& \tilde{\boldsymbol{\mu}}_{k,t}(\mathbf{x}_t^k, \mathbf{x}_0^k) = \boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t-1}\boldsymbol{L}_{k,t}^2\bar{\boldsymbol{L}}_{k,t}^{-2}\boldsymbol{U}_k^T\mathbf{x}_0^k + \boldsymbol{U}_k\boldsymbol{\Lambda}_{k,t}\bar{\boldsymbol{L}}_{k,t-1}^2\bar{\boldsymbol{L}}_{k,t}^{-2}\boldsymbol{U}_k^T\mathbf{x}_t^k, \\
\tilde{\boldsymbol{\Sigma}}_{k,t} =& \boldsymbol{U}_k\boldsymbol{L}_{k,t}^2\bar{\boldsymbol{L}}_{k,t-1}^2\bar{\boldsymbol{L}}_{k,t}^{-2}\boldsymbol{U}_k^T.
\end{aligned}
\tag{S7}
$$

Thus, $q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k, \mathbf{x}_0^k) = \mathcal{N}(\mathbf{x}_{t-1}^k; \tilde{\boldsymbol{\mu}}_{k,t}, \tilde{\boldsymbol{\Sigma}}_{k,t})$.

A.3. Derivation of the Loss Function

Here we derive Eq. (12) from the variational bound on negative log-likelihood

$$
\begin{aligned}
\mathbb{E}_q[-\log p_\theta(\mathbf{x}_0^0)] \le& \mathbb{E}_q\left[-\log\frac{p_\theta(\mathbf{x}_{0:T_1}^0, \mathbf{x}_{T_1:T_2}^1, \cdots, \mathbf{x}_{T_K:T}^K)}{q(\mathbf{x}_{1:T_1}^0, \mathbf{x}_{T_1:T_2}^1, \cdots, \mathbf{x}_{T_K:T}^K|\mathbf{x}_0^0)}\right] \\
=& \mathbb{E}_q\left[-\log\frac{p_\theta(\mathbf{x}_0^0|\mathbf{x}_1^0)p_\theta(\mathbf{x}_{1:T_1}^0, \mathbf{x}_{T_1:T_2}^1, \cdots, \mathbf{x}_{T_K:T-1}^K|\mathbf{x}_T^K)p_\theta(\mathbf{x}_T^K)}{q(\mathbf{x}_{1:T_1}^0, \mathbf{x}_{T_1:T_2}^1, \cdots, \mathbf{x}_{T_K:T-1}^K|\mathbf{x}_T^K, \mathbf{x}_0^0)q(\mathbf{x}_T^K|\mathbf{x}_0^0)}\right] \\
=& \mathbb{E}_q\Bigg[-\log p_\theta(\mathbf{x}_0^0|\mathbf{x}_1^0) - \sum_{k=0}^{K}\sum_{\substack{t=T_k+1\\t>1}}^{T_{k+1}}\log\frac{p_\theta(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k)}{q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k, \mathbf{x}_0^k)} \\
&\quad - \sum_{k=1}^{K}\log\frac{p_\theta(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k)}{q(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k, \mathbf{x}_0^0)} - \log\frac{p_\theta(\mathbf{x}_T^K)}{q(\mathbf{x}_T^K|\mathbf{x}_0^0)}\Bigg] \\
=& \mathbb{E}_q\Bigg[\underbrace{-\log p_\theta(\mathbf{x}_0^0|\mathbf{x}_1^0)}_{L_0} + \sum_{k=0}^{K}\sum_{\substack{t=T_k+1\\t>1}}^{T_{k+1}}\underbrace{\mathrm{D_{KL}}(q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k, \mathbf{x}_0^k)\|p_\theta(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k))}_{L_{t-1}} \\
&\quad + \sum_{k=1}^{K}\underbrace{\mathrm{D_{KL}}(q(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k, \mathbf{x}_0^{k-1})\|p_\theta(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k))}_{L_k^{\text{down}}} + \underbrace{\mathrm{D_{KL}}(q(\mathbf{x}_T^K|\mathbf{x}_0^0)\|p_\theta(\mathbf{x}_T^K))}_{L_T}\Bigg],
\end{aligned}
\tag{S8}
$$

where L_0, L_T and L_{t-1} for $t = 2, 3, \cdots, T$ are similar with the definitions in DDPM [1], and L_k^{down} is a new term which can be viewed as the loss at the dimensionality turning point T_k. As defined in Eq. (11), $p_\theta(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k)$ has no learnable parameters, so we do not optimize L_k^{down}.

As for L_{t-1}, it is the KL divergence of two Gaussians and can be calculated as

$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2}\|\boldsymbol{\Sigma}_t^{-1/2}\boldsymbol{U}_k^T(\tilde{\boldsymbol{\mu}}_{k,t}(\mathbf{x}_t^k,\mathbf{x}_0^k) - \boldsymbol{\mu}_\theta(\mathbf{x}_t^k,t))\|^2\right] + C, \tag{S9}$$

where C is a constant that does not depend on θ, k is determined by t such that $T_k < t \leq T_{k+1}$, $\tilde{\boldsymbol{\mu}}_{k,t}(\mathbf{x}_t^k,\mathbf{x}_0^k)$ is the mean of $q(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k,\mathbf{x}_0^k)$ given by Eq. (S7), and $\boldsymbol{\mu}_\theta$ is the mean of $p_\theta(\mathbf{x}_{t-1}^k|\mathbf{x}_t^k)$ given by Eq. (10).

With Eqs. (5, 6), L_{t-1} can be represented by reparameterization trick as

$$\begin{aligned} L_{t-1} =& \mathbb{E}_{\mathbf{x}_0^k,\epsilon^k}\left[\frac{1}{2}\Big\|\boldsymbol{\Sigma}_t^{-1/2}\boldsymbol{U}_k^T(\tilde{\boldsymbol{\mu}}_{k,t}(\mathbf{x}_t^k(\mathbf{x}_0^k,\epsilon^k), \boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t}^{-1}\boldsymbol{U}_k^T\mathbf{x}_t^k(\mathbf{x}_0^k,\epsilon^k) - \boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,t}^{-1}\bar{\boldsymbol{L}}_{k,t}\boldsymbol{U}_k^T\boldsymbol{\epsilon}^k)\right. \\ & \left. - \boldsymbol{\mu}_\theta(\mathbf{x}_t^k(\mathbf{x}_0^k,\epsilon^k),t))\Big\|^2\right] + C \\ =& \mathbb{E}_{\mathbf{x}_0^k,\epsilon^k}\left[\Big\|\boldsymbol{W}_t(\epsilon^k - \epsilon_\theta(\mathbf{x}_t^k(\mathbf{x}_0^k,\epsilon^k),t))\Big\|^2\right] + C, \end{aligned} \tag{S10}$$

where the final equality is obtained by plugging Eq. (10) and Eq. (S7) into it, and $\boldsymbol{W}_t = \frac{1}{\sqrt{2}}\boldsymbol{\Sigma}_t^{-1/2}\boldsymbol{\Lambda}_{k,t}^{-1}\boldsymbol{L}_{k,t}^2\bar{\boldsymbol{L}}_{k,t}^{-1}\boldsymbol{U}_k^T$.

Finally, by setting $\boldsymbol{W}_t = \boldsymbol{I}$, we can obtain Eq. (12), similar with the simplified training objective in DDPM [1].

A.4. Proof of Proposition 1

According to the inequality between JSD and *total variation*, we have

$$\text{JSD}(p_1||p_2) \leq \frac{1}{2}\int |p_1(x) - p_2(x)|d\boldsymbol{x}. \tag{S11}$$

The RHS (right-hand side) of Eq. (S11) satisfies

$$\begin{aligned} \frac{1}{2}\int |p_1(\boldsymbol{x}) - p_2(\boldsymbol{x})|d\boldsymbol{x} =& \frac{1}{2}\int |\mathbb{E}_{\mathbf{x}_0\sim p}[p_1(\boldsymbol{x}|\mathbf{x}_0) - p_2(\boldsymbol{x}|\mathbf{x}_0)]|\, d\boldsymbol{x} \\ \leq& \frac{1}{2}\int \mathbb{E}_{\mathbf{x}_0\sim p}\left[|p_1(\boldsymbol{x}|\mathbf{x}_0) - p_2(\boldsymbol{x}|\mathbf{x}_0)|\right] d\boldsymbol{x} \\ =& \frac{1}{2}C_1\int \mathbb{E}_{\mathbf{x}_0\sim p}\left[\Big|\exp\big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{A}_1\mathbf{x}_0)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{A}_1\mathbf{x}_0)\big)\right. \\ & \left. - \exp\big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{A}_2\mathbf{x}_0)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{A}_2\mathbf{x}_0)\big)\Big|\right] d\boldsymbol{x}, \end{aligned} \tag{S12}$$

where $C_1 = (2\pi)^{-d/2}\det(\boldsymbol{\Sigma})^{-1/2}$.

According to the mean value theorem, for each $\mathbf{x}_0$ and $\boldsymbol{x}$, there exists $\theta = \theta(\mathbf{x}_0, \boldsymbol{x}) \in [0,1]$ such that $\boldsymbol{\xi} = \theta(\boldsymbol{x} - \boldsymbol{A}_1\mathbf{x}_0) + (1-\theta)(\boldsymbol{x} - \boldsymbol{A}_2\mathbf{x}_0) = \boldsymbol{x} - [\theta\boldsymbol{A}_1 + (1-\theta)\boldsymbol{A}_2]\mathbf{x}_0$ satisfies

$$\begin{aligned} & \exp\big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{A}_1\mathbf{x}_0)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{A}_1\mathbf{x}_0)\big) - \exp\big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{A}_2\mathbf{x}_0)^T\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{A}_2\mathbf{x}_0)\big) \\ =& \boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{A}_1 - \boldsymbol{A}_2)\mathbf{x}_0\exp\big(-\frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\big) \\ =& F\cdot\exp\big(-\frac{1}{4}\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\big), \end{aligned} \tag{S13}$$

where $F = \boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{A}_1 - \boldsymbol{A}_2)\mathbf{x}_0\exp\big(-\frac{1}{4}\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\big)$, and $|F|$ satisfies the following inequality

$$\begin{aligned} |F| =& \Big|\frac{(\boldsymbol{\Sigma}^{1/2}\boldsymbol{\xi})^T}{\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}\|_2}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}_1 - \boldsymbol{A}_2)\mathbf{x}_0\Big|\cdot\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}\|_2\exp\big(-\frac{1}{4}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\xi}\|_2^2\big) \\ \leq& C_2\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}_1 - \boldsymbol{A}_2)\mathbf{x}_0\|_2 \\ \leq& C_2B\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}_1 - \boldsymbol{A}_2)\|_2, \end{aligned} \tag{S14}$$

where $C_2 = \max_{a\ge 0} ae^{-\frac{1}{4}a^2} = \sqrt{2}e^{-\frac{1}{2}}$, and B is the upper bound of $\|\mathbf{x}_0\|_2$ as assumption.

Combining Eqs. (S12) to (S14), we have

$$\frac{1}{2}\int |p_1(\boldsymbol{x}) - p_2(\boldsymbol{x})| d\boldsymbol{x} \le \frac{1}{2} C_1 C_2 B \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{A}_1 - \boldsymbol{A}_2)\|_2 \int \mathbb{E}_{\mathbf{x}_0\sim p}\Big[\exp\big(-\frac{1}{4}\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\big)\Big] d\boldsymbol{x}, \tag{S15}$$

where $\boldsymbol{\xi} = \boldsymbol{x} - [\theta\boldsymbol{A}_1 + (1-\theta)\boldsymbol{A}_2]\mathbf{x}_0$. Now we only need to prove that the RHS of Eq. (S15) $\le$ the LHS (left-hand side) of Eq. (13).

Let $\mathbf{y} = \boldsymbol{\Sigma}^{-1/2}[\theta\boldsymbol{A}_1 + (1-\theta)\boldsymbol{A}_2]\mathbf{x}_0$, then $\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi} = \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} - \mathbf{y}\|_2^2$, and $\mathbf{y}$ satisfies

$$\begin{aligned}\|\mathbf{y}\|_2 =& \|\boldsymbol{\Sigma}^{-1/2}[\theta\boldsymbol{A}_1 + (1-\theta)\boldsymbol{A}_2]\mathbf{x}_0\|_2 \\ \le& B\|\boldsymbol{\Sigma}^{-1/2}[\theta\boldsymbol{A}_1 + (1-\theta)\boldsymbol{A}_2]\|_2 \\ \le& B\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}_1\|_2,\end{aligned} \tag{S16}$$

where the last inequality is derived from the assumption that $\boldsymbol{A}_1 \succeq \boldsymbol{A}_2 \succeq \mathbf{0}$.

Let $\mathbb{D} = \{\boldsymbol{x} : \|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\|_2 \le r\}$, where $r = 2B\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}_1\|_2$, thus $\|\mathbf{y}\|_2 \le \frac{1}{2}r$ according to Eq. (S16). Then the integration in Eq. (S15) can be split into two regions as

$$\begin{aligned}\int \mathbb{E}_{\mathbf{x}_0\sim p}\Big[\exp\big(-\frac{1}{4}\boldsymbol{\xi}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\big)\Big] d\boldsymbol{x} =& \int_{\mathbb{D}} \mathbb{E}_{\mathbf{x}_0\sim p}\Big[\exp\big(-\frac{1}{4}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} - \mathbf{y}\|_2^2\big)\Big] d\boldsymbol{x} \\ &+ \int_{\mathbb{D}^C} \mathbb{E}_{\mathbf{x}_0\sim p}\Big[\exp\big(-\frac{1}{4}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x} - \mathbf{y}\|_2^2\big)\Big] d\boldsymbol{x} \\ \le& \int_{\mathbb{D}} 1 d\boldsymbol{x} + \int \exp\big(-\frac{1}{16}\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}\|_2^2\big) d\boldsymbol{x} \\ \le& V_d(r)\det(\Sigma)^{1/2} + 2\sqrt{2}(2\pi)^{d/2}\det(\Sigma)^{1/2},\end{aligned} \tag{S17}$$

where $V_d(\cdot)$ is the volume of d-dimensional sphere with respect to the radius.

Combining Eqs. (S11), (S15) and (S17), we can get Proposition 1.

A.5. Proof of Theorem 1

We first prove that $q(\mathbf{x}_{T_k}^{k-1})$ and $p(\mathbf{x}_{T_k}^{k-1})$ defined in Theorem 1 satisfy the conditions claimed in Proposition 1.

$q(\mathbf{x}_{T_k}^{k})$ is the marginal distribution of $q(\mathbf{x}_0^k)q(\mathbf{x}_{T_k}^k|\mathbf{x}_0^k)$ where $q(\mathbf{x}_{T_k}^k|\mathbf{x}_0^k)$ is defined by Eqs. (5, 6). $q(\mathbf{x}_{T_k}^k|\mathbf{x}_0^k)$ can also be expressed as

$$q(\mathbf{x}_{T_k}^k|\mathbf{x}_0^k) = \mathcal{N}(\mathbf{x}_{T_k}^k; \boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,T_k}\boldsymbol{U}_k^T\mathbf{x}_0^k, \boldsymbol{U}_k\bar{\boldsymbol{L}}_{k,T_k}^2\boldsymbol{U}_k^T). \tag{S18}$$

Similarly, $q(\mathbf{x}_{T_k}^{k-1})$ is the marginal distribution of $q(\mathbf{x}_0^{k-1})q(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1})$ where $q(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1})$ can be expressed as

$$q(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1}) = \mathcal{N}(\mathbf{x}_{T_k}^{k-1}; \boldsymbol{U}_{k-1}\bar{\boldsymbol{\Lambda}}_{k-1,T_k}\boldsymbol{U}_{k-1}^T\mathbf{x}_0^{k-1}, \boldsymbol{U}_{k-1}\bar{\boldsymbol{L}}_{k-1,T_k}^2\boldsymbol{U}_{k-1}^T). \tag{S19}$$

By definition, $p(\mathbf{x}_{T_k}^{k-1})$ is the marginal distribution of $q(\mathbf{x}_{T_k}^k)p(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k)$, where $p(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k)$ is defined by Eq. (11) and can be expressed as

$$p(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k) = \mathcal{N}(\mathbf{x}_{T_k}^{k-1}; \mathcal{D}_k^T\mathbf{x}_{T_k}^k, \boldsymbol{U}_{k-1}\Delta\boldsymbol{L}_{k-1}^2\boldsymbol{U}_{k-1}^T). \tag{S20}$$

To transform $p(\mathbf{x}_{T_k}^{k-1})$ into the form in Proposition 1, we construct a Markov chain $\mathbf{x}_0^{k-1} \to \mathbf{x}_{T_k}^k \to \mathbf{x}_{T_k}^{k-1}$, where $\mathbf{x}_0^{k-1} \sim q(\mathbf{x}_0^{k-1})$, $\mathbf{x}_{T_k}^k|\mathbf{x}_0^{k-1} \sim q(\mathbf{x}_{T_k}^k|\mathbf{x}_0^{k-1}) = q(\mathbf{x}_{T_k}^k|\mathcal{D}_k\mathbf{x}_0^{k-1})$, and $\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k \sim p(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k)$. Thus $p(\mathbf{x}_{T_k}^{k-1})$ is also the marginal distribution of the joint distribution $p_{\mathrm{m}}(\mathbf{x}_0^{k-1}, \mathbf{x}_{T_k}^{k-1})$ defined by the Markov chain. This joint distribution can be factorized as $p_{\mathrm{m}}(\mathbf{x}_0^{k-1}, \mathbf{x}_{T_k}^{k-1}) = q(\mathbf{x}_0^{k-1})p_{\mathrm{m}}(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1})$, where $p_{\mathrm{m}}(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1})$ is the marginal distribution of $q(\mathbf{x}_{T_k}^k|\mathbf{x}_0^{k-1})p(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_{T_k}^k)$, and can be derived from Eqs. (S18) and (S20) by using Eq. (S1)

$$\begin{aligned}p_{\mathrm{m}}(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1}) =& \mathcal{N}(\mathbf{x}_{T_k}^{k-1}; \mathcal{D}_k^T\boldsymbol{U}_k\bar{\boldsymbol{\Lambda}}_{k,T_k}\boldsymbol{U}_k^T\mathcal{D}_k\mathbf{x}_0^{k-1},\ \boldsymbol{U}_{k-1}\Delta\boldsymbol{L}_{k-1}^2\boldsymbol{U}_{k-1}^T + \mathcal{D}_k^T\boldsymbol{U}_k\bar{\boldsymbol{L}}_{k,T_k}^2\boldsymbol{U}_k^T\mathcal{D}_k) \\ =& \mathcal{N}(\mathbf{x}_{T_k}^{k-1}; \boldsymbol{U}_{k-1}(\bar{\boldsymbol{\Lambda}}_{k-1,T_k} - \Delta\boldsymbol{\Lambda}_{k-1})\boldsymbol{U}_{k-1}^T\mathbf{x}_0^{k-1},\ \boldsymbol{U}_{k-1}\bar{\boldsymbol{L}}_{k-1,T_k}^2\boldsymbol{U}_{k-1}^T),\end{aligned} \tag{S21}$$

where $\Delta\boldsymbol{\Lambda}_{k-1} = \mathrm{diag}(\bar{\lambda}_{k-1,T_k}\boldsymbol{I}_{d_{k-1}}, \boldsymbol{O}_{\bar{d}_k})$.

Thus, $q(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1})$ given by Eq. (S19) and $p_{\rm m}(\mathbf{x}_{T_k}^{k-1}|\mathbf{x}_0^{k-1})$ given by Eq. (S21) satisfy conditions of $p_1(\mathbf{x}|\mathbf{x}_0)$ and $p_2(\mathbf{x}|\mathbf{x}_0)$ claimed in Proposition 1 respectively. And $\|\mathbf{x}_0^{k-1}\|$ satisfies

$$\|\mathbf{x}_0^{k-1}\| = \|\overline{\mathcal{D}}_{k-1}\mathbf{x}_0^0\| \leq \|\mathbf{x}_0^0\| \leq \sqrt{d}. \tag{S22}$$

Finally, substituting all corresponding variables into Eq. (13), we can obtain Eq. (14).

B. Implementation Details

In this section, we will give more details on the implementation of our DVDP. Algorithms 1 and 2 display the complete training and sampling procedures respectively.

Algorithm 1 Training

```
repeat
    k ~ Uniform({0, ···, K})
    t ~ Uniform({T_k + 1, ···, T_{k+1}})
    x_0^0 ~ q(x_0^0)
    ε^k ~ N(0; I_{d̄_k})
    x_0^k ← D̄_k x_0^0
    x_t^k ← U_k Λ̄_{k,t} U_k^T x_0^k + U_k L̄_{k,t} U_k^T ε^k
    Take gradient descent step on
        ∇_θ ||ε^k − ε_θ(x_t^k, t)||^2
until converged
```

Algorithm 2 Sampling

```
x_T^K ~ N(0; I_{d̄_K})
for k = K, ···, 0 do
    for t = T_{k+1}, ···, T_k + 1 do
        ε^k ~ N(0; I_{d̄_k})
        x_{t−1}^k ← U_k Λ̄_{k,t}^{−1}(U_k^T x_t^k − L̄_{k,t} U_k^T ε_θ(x_t^k, t)) + Σ_t ε^k
    if k > 0 then
        ε^{k−1} ~ N(0; I_{d̄_{k−1}})
        x_{T_k}^{k−1} ← D_k^T x_{T_k}^k + U_{k−1} ΔL_{k−1} U_{k−1}^T ε^{k−1}
return x_0^0
```

B.1. Choice of Downsampling Operator $\mathcal{D}_k$

Since an image pixel is usually similar with its neighbours, we can simply choose $\mathcal{D}_k$ to be a 2×2 average-pooling operator for each $k = 1, \cdots, K$ as in subspace diffusion [2] to maintain the main component of an image. Under this choice, the dimensionality will be reduced from $\bar{d}_{k-1}$ to $\bar{d}_k = \frac{1}{4}\bar{d}_{k-1}$ after each downsampling operation, as mentioned in Sec. 5.1.

The above choice needs a simple modification, multiplication by 2 after the average-pooling operation, to ensure that the matrix $\mathcal{D}_k$ satisfies

$$\mathcal{D}_k[\boldsymbol{N}_{k-1}, \boldsymbol{P}_{k-1}] = [\boldsymbol{O}_{\bar{d}_k}, \boldsymbol{U}_k], \tag{S23}$$

where $\boldsymbol{N}_{k-1} \in \mathbb{R}^{\bar{d}_{k-1} \times d_{k-1}}$ and $\boldsymbol{P}_{k-1} \in \mathbb{R}^{\bar{d}_{k-1} \times \bar{d}_k}$ satisfies $\boldsymbol{U}_{k-1} = [\boldsymbol{N}_{k-1}, \boldsymbol{P}_{k-1}]$. Under this condition, the matrix $\mathcal{D}_k \in \mathbb{R}^{\bar{d}_k \times \bar{d}_{k-1}}$ is row-orthogonal, and $\mathcal{D}_k^T$ (*i.e.*, the transpose of $\mathcal{D}_k$) is just the corresponding upsampling operator.

B.2. Attenuation Coefficient $\bar{\lambda}_{k,t}$

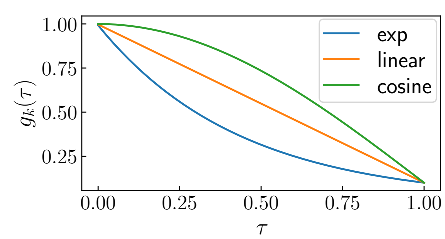

Figure S1. $g_k(\cdot)$ on $[0, 1]$ for $k < K$.

Algorithm 3 Adaptation on Noise Schedule

```
Initialize ᾱ_0, ··· ᾱ_T as in DDPM
for t = 0, ···, T do
    σ̄_t ← sqrt(1/ᾱ_t − 1)
for k = 1, ···, K do
    for t = T_k + 1, ···, T do
        σ̄_t ← σ̄_{T_k} + 2(σ̄_t − σ̄_{T_k})
```

As mentioned in Sec. 4.1, among $\bar{\lambda}_{0,t}, \bar{\lambda}_{1,t}, \cdots, \bar{\lambda}_{K,t}$, only $\bar{\lambda}_{k,t}$ is required to approximate zero at dimensionality turning point T_{k+1} for $k < K$. Thus we only need to decrease $\bar{\lambda}_{k,t}$ when $T_k < t \leq T_{k+1}$. Hence, $\bar{\lambda}_{k,t}$ can be set in the following manner for $k < K$:

$$\bar{\lambda}_{k,t} = \begin{cases} 1, & t \leq T_k \\ \bar{\lambda}_{\min}, & t > T_{k+1} \\ g_k(\frac{t-T_k}{T_{k+1}-T_k}), & T_k < t \leq T_{k+1}, \end{cases} \tag{S24}$$

where $\bar{\lambda}_{\min} \in (0, 1)$ is a shared hyperparameter for $\bar{\lambda}_{0,t}, \bar{\lambda}_{1,t}, \cdots, \bar{\lambda}_{K-1,t}$, and $g_k(\cdot)$ is a continuous decreasing function on the interval $[0, 1]$ such that $g_k(0) = 1$ and $g_k(1) = \bar{\lambda}_{\min}$. For all experiments, we set $\bar{\lambda}_{\min} = 0.01$. As for $k = K$, we set $\bar{\lambda}_{K,t} = 1$ for all t. This schedule means that between two adjacent dimensionality turning points T_k and T_{k+1}, we only attenuate one data component $\mathbf{v}_k^k$. Once we have $\bar{\lambda}_{k,t}$ for each k and t, matrices $\bar{\boldsymbol{\Lambda}}_{k,t}$ and $\boldsymbol{\Lambda}_{k,t}$ are determined.

To determine $g_k(\cdot)$, we try *exp*, *linear* and *cosine*, 3 different functions as shown in Fig. S1, with $T_0 = 300$ and $T_1 = 600$ on CelebA 64. The FIDs are 3.98, 4.25 and 4.31 respectively, while the baseline case, *i.e.*, $\bar{\lambda}_{\min} = 1$ without attenuation, achieves 4.99. Although all three attenuation functions outperform baseline, we choose the best-performing *exp* for all other experiments. Specifically, we set $g_k(\tau) = (\bar{\lambda}_{\min})^\tau$ for $\tau \in [0, 1]$.

B.3. Noise Schedule $\bar{\sigma}_{k,t}$

Recall that $\bar{\sigma}_{k,t}$ denotes the standard deviation of noise component $\mathbf{z}_k^0$ defined in Eq. (5). To determine $\bar{\sigma}_{k,t}$ for all $k = 0, 1, \cdots, K$, it is equivalent to determine the matrix $\bar{\boldsymbol{L}}_{0,t} = \mathrm{diag}(\bar{\sigma}_{0,t}\boldsymbol{I}_{d_0},\ \bar{\sigma}_{1,t}\boldsymbol{I}_{d_1},\ \cdots,\ \bar{\sigma}_{K,t}\boldsymbol{I}_{d_K})$. Note that $\bar{\boldsymbol{L}}_{0,t}$ for all $t = 1, 2, \cdots, T$ can be uniquely determined by $\bar{\boldsymbol{L}}_{0,t} = (\boldsymbol{L}_{0,t}^2 + \boldsymbol{\Lambda}_{0,t}^2\bar{\boldsymbol{L}}_{0,t-1}^2)^{1/2}$, which is derived from the definition of $\boldsymbol{L}_{0,t}$ under Eq. (7), once we know $\boldsymbol{\Lambda}_{0,t},\ \boldsymbol{L}_{0,t}$ for all $t > 0$ (initial $\bar{\boldsymbol{L}}_{0,t} = \boldsymbol{O}_{\bar{d}_0}$). Thus, we can equivalently design $\boldsymbol{L}_{0,t}$ instead of directly setting $\bar{\boldsymbol{L}}_{0,t}$.

For simplicity, we can set $\boldsymbol{L}_{0,t} = \sigma_t\boldsymbol{I}_{\bar{d}_0}$ as a diagonal matrix, which means that the added noise at each forward diffusion step is symmetric, similar with that in DDPM [1]. Rather than setting σ_t directly, we first determine $\bar{\sigma}_t = \sqrt{\sum_{s=1}^t \sigma_s^2}$, then obtain σ_t by $\sigma_t = \sqrt{\bar{\sigma}_t^2 - \bar{\sigma}_{t-1}^2}$. In fact, $\bar{\sigma}_t = \bar{\sigma}_{t,K}$, *i.e.*, the standard deviation of noise component $\mathbf{z}_K^0$ in the smallest subspace $\mathbb{S}_K$, in which no attenuation is applied. Thus, $\bar{\sigma}_t$ determines the signal-to-noise ratio (SNR) when $T_K < t \leq T$, accompanied by data distribution and the choice of $\mathbb{S}_K$. Given our choice of $\mathcal{D}_k$ in Sec. B.1, subspace $\mathbb{S}_K$ contains the main component of an image, but only small parts of a Gaussian noise. We can approximately define the SNR as $2^K/\bar{\sigma}_T$ at timestep T after K times downsampling.

For diffusion models, small SNR at timestep T is a key ingredient for high quality samples, which ensures that the noisy data can be well approximated by a Gaussian noise. Here, we choose to mimic the SNR schedule in DDPM [1], and this requires to adjust the sequence of $\bar{\sigma}_t$ to approximately compensate the factor 2^K in SNR. The noise adaptation algorithm is given in Algorithm 3.

B.4. Simplification of Matrix Multiplication $\boldsymbol{U}_k\boldsymbol{G}_k\boldsymbol{U}_k^T$

With the above choices of attenuation coefficients and noise schedule, all matrix multiplications with the form of $\boldsymbol{U}_k\boldsymbol{G}_k\boldsymbol{U}_k^T$ in the implementation of DVDP can be expressed by downsampling operator $\mathcal{D}_{k+1}$ and upsampling operator $\mathcal{D}_{k+1}^T$, since each diagonal matrix $\boldsymbol{G}_k$ only includes two different elements and can be expressed in the form of $\boldsymbol{G} = \mathrm{diag}(a_k\boldsymbol{I}_{d_k}, b_k\boldsymbol{I}_{\bar{d}_{k+1}})$. Thus, $\boldsymbol{U}_k\boldsymbol{G}_k\boldsymbol{U}_k^T$ can be expressed as

$$\begin{aligned}
\boldsymbol{U}_k\boldsymbol{G}_k\boldsymbol{U}_k^T =& [\boldsymbol{N}_k, \boldsymbol{P}_k]\begin{bmatrix} a_k\boldsymbol{I}_{d_k} & \mathbf{0} \\ \mathbf{0} & b_k\boldsymbol{I}_{\bar{d}_{k+1}} \end{bmatrix}[\boldsymbol{N}_k, \boldsymbol{P}_k]^T \\
=& a_k\boldsymbol{I}_{\bar{d}_k} + [\boldsymbol{N}_k, \boldsymbol{P}_k]\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (b_k - a_k)\boldsymbol{I}_{\bar{d}_{k+1}} \end{bmatrix}[\boldsymbol{N}_k, \boldsymbol{P}_k]^T \\
=& a_k\boldsymbol{I}_{\bar{d}_k} + (b_k - a_k)\boldsymbol{P}_k\boldsymbol{P}_k^T \\
=& a_k\boldsymbol{I}_{\bar{d}_k} + (b_k - a_k)\mathcal{D}_{k+1}^T\mathcal{D}_{k+1},
\end{aligned} \tag{S25}$$

where the last equality can be derived from Eq. (S23).

B.5. Hyperparameters

Hyperparameters for training DVDP models are in Tab. S1. We train all of our models using AdamW [3] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use EMA for all experiments with a decay factor of 0.9999. A single NVIDIA A100 is used for all experiments, except FFHQ 1024 with two A100s.

C. Experiments on DDIM Sampling

To demonstrate that our DVDP is compatible with DDIM [5], an accelerated sampling method, we apply DDIM to our models trained on LSUN Church 256×256 and FFQH 256×256. The results are shown in Tab. S2. In experiment, we find that it is beneficial for DVDP to add noises in some middle steps of sampling, unlike DDIM that sets all inserted noises to zeros. Specifically, we set $\eta_t = 1$ for $t = T_1 - \lfloor T_1/4 \rfloor, \cdots, T_1 + \lceil (T_2 - T_1)/2 \rceil$ and $\eta_t = 0$ otherwise, where $\eta_t \in [0, 1]$ controls the strength of added noise as in DDIM [5] for timestep t. This adaption is marked by * in Tab. S2.

Table S1. Hyperparameters for our DVDP models. *We used 1500K iterations for FFHQ 256, 1200K for LSUN church, and 1600K for both LSUN bedroom and cat.

	CIFAR 32	FFHQ & LSUN 256	FFHQ 1024
Model size	40M	125M	105M
Channels	128	128	128
Depth	2	2	2
Channel multi.	1,2,2,2	1,1,2,2,4,4	1,1,2,2,3,4
Head channels	64	64	64
Attention scale	1/2	1/16	1/16
Dropout	0.1	0	0
Batch size	128	24	8
Iterations	400K	vary by datasets*	1000K
Learning rate	2e-4	1e-4	1e-4

Table S2. Quantitative comparison measured in FID. DDIM* denotes an adapted DDIM sampling method.

Dataset	Church 256×256			FFHQ 256×256		
#Steps	50	100	200	50	100	200
DDIM Baseline	10.44	10.22	10.26	12.32	10.80	10.19
DDIM* Baseline	9.36	8.91	9.03	13.33	10.28	9.06
DDIM* DVDP	**8.52**	**7.33**	**7.32**	**12.01**	**8.39**	**7.04**

Table S3. Combination with latent diffusion model (LDM) on FFHQ 256, where K denotes downsampling times.

Method	FID ↓	Acc ↑
LDM	4.98	-
Comb. ($K = 1$)	**4.55**	1.39×
Comb. ($K = 2$)	5.12	**1.90×**

Table S4. Comparison with subspace diffusion model (SDM) measured in FID, where $*$ indicates the same noise schedule as in conventional diffusion without noise schedule adaptation in Algorithm 3.

Method	CelebA 64	Church 128	Bedroom 128
SDM*	5.26	6.86	5.21
SDM	4.99	6.74	5.42
Ours	**3.98**	**5.62**	**4.88**

D. Combination with Latent Diffusion Models

The latent diffusion model (LDM) [4], which carries out the diffusion process at a latent space instead of the image space, can speed up both training and sampling like our DVDP. However, DVDP is compatible with it and can realize further acceleration. Tab. S3 gives the performance of combining DVDP with LDM on FFHQ 256 with official configuration. The combined models are finetuned from the official model weight.

E. Comparison with Subspace Diffusion

We compare DVDP with subspace diffusion model (SDM) [2] on a few additional datasets with $K = 2$. Note that the two alternatives *share the same* efficiency. For a fair comparison, we train each SDM with a single network like our DVDP, instead of using different models for different stages as in [2]. Tab. S4 suggests that DVDP outperforms SDM on all these datasets.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, pages 6840–6851, 2020. 2, 3, 6

[2] Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. Subspace diffusion generative models. *ArXiv:2205.01490*, 2022. 1, 5, 7

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ArXiv:1711.05101*, 2017. 6

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 1, 7

[5] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *Int. Conf. Learn. Represent.*, 2021. 1, 6