This document will demonstrate why fine-tuning diffusion at random time steps is feasible and robust in < Universal Rumor Detection on Modality Consistency and External Knowledge >.

The standard diffusion process can be represented as:

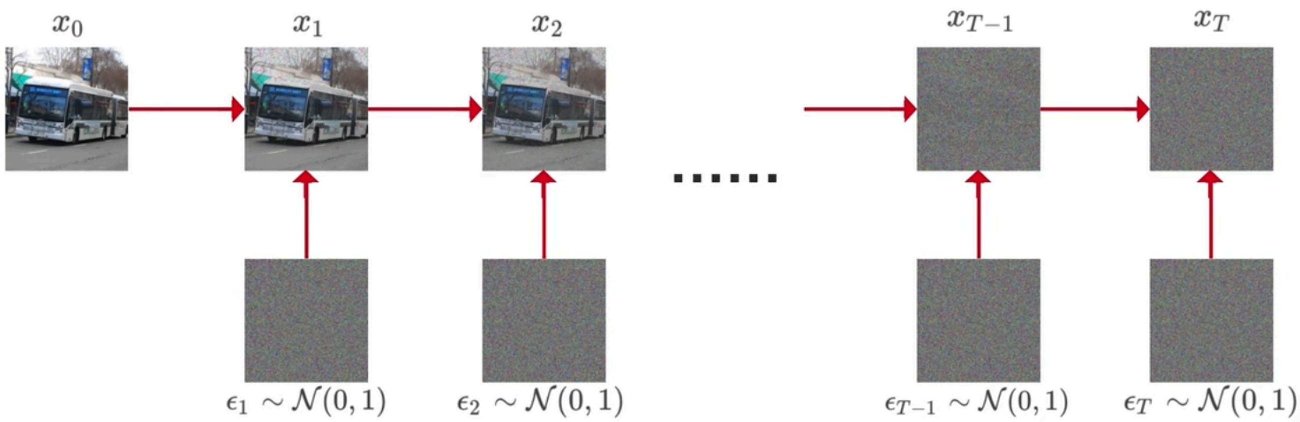
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

where β_t is the noise intensity at time step t, and x_t is the noisy image at time t, where β_t is a hyperparameter, and satisfies $0 < \beta_t < 1$ and $\beta_1 < \beta_2 < ... < \beta_{t-1} < \beta_t$.

If we want to sample a z from a Gaussian distribution $z \sim \mathcal{N}(z; \mu_\theta, \sigma_\theta^2\mathbf{I})$, we can write it as follow:

$$z = \mu_\theta + \sigma_\theta\epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{2}$$

The forward diffusion can be represented in terms of images as:



Based on the above information, we can conclude that:

$$\begin{aligned}
\mathbf{x_1} &= \sqrt{1-\beta_1}\mathbf{x}_0 + \sqrt{\beta_1}\epsilon_1, \epsilon_1 \sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{x_2} &= \sqrt{1-\beta_2}\mathbf{x}_1 + \sqrt{\beta_2}\epsilon_2, \epsilon_2 \sim \mathcal{N}(0, \mathbf{I}) \\
\mathbf{x_3} &= \sqrt{1-\beta_3}\mathbf{x}_2 + \sqrt{\beta_3}\epsilon_3, \epsilon_3 \sim \mathcal{N}(0, \mathbf{I}) \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots
\end{aligned} \tag{3}$$

Thus, Equation (1) can be written as:

$$\mathbf{x_t} = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \tag{4}$$

Where ϵ_t is a random number that is re-sampled at each time t.

Let $\alpha_t = 1 - \beta_t$, we obtain:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon_t}, \epsilon_t \sim \mathcal{N}(0, \mathbf{I}) \tag{5}$$

By continuously iterating, we can derive:

$$\begin{aligned}
\mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon_t} \\
&= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\mathbf{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon_{t-1}}) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon_t} \\
&= \sqrt{\alpha_t}[\sqrt{\alpha_{t-1}}(\sqrt{\alpha_{t-2}}\mathbf{x}_{t-3} + \sqrt{1-\alpha_{t-2}}\boldsymbol{\epsilon_{t-2}}) + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon_{t-1}})] + \sqrt{1-\alpha_t}\boldsymbol{\epsilon_t} \\
&= \ldots\ldots \\
&= \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\ldots\sqrt{\alpha_1}x_0 + \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\ldots\sqrt{\alpha_2}\sqrt{1-\alpha_1}\epsilon_1 + \sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\ldots\sqrt{\alpha_3}\sqrt{1-\alpha_2}\epsilon_2 + \ldots +
\end{aligned} \tag{6}$$

Where $\epsilon_1 \ldots \epsilon_t$ all follow a standard normal distribution, following the distribution of $\mathcal{N}(0, \mathbf{I})$, possessing the following two properties:

$$\begin{gathered}
c\epsilon \sim \mathcal{N}(0, c^2\mathbf{I}), c \text{ is a constant.} \\
\text{Additivity means that } \mathcal{N}(0, \boldsymbol{\sigma_1^2}\mathbf{I}) + \mathcal{N}(0, \boldsymbol{\sigma_2^2}\mathbf{I}) \sim \mathcal{N}(0, (\boldsymbol{\sigma_1^2} + \boldsymbol{\sigma_2^2})\mathbf{I})
\end{gathered}$$

Based on the above properties, we can conclude that

$$\sqrt{\alpha_t}\sqrt{\alpha_{t-1}}\ldots\sqrt{\alpha_2}\sqrt{1-\alpha_1}\epsilon_1 \sim \mathcal{N}(0, \alpha_t\alpha_{t-1}\ldots\alpha_2(1-\alpha_1)\mathbf{I})$$

Furthermore, from additivity, we can derive that the variance is

$$\begin{aligned}
&\alpha_t\alpha_{t-1}\ldots\alpha_2(1-\alpha_1) + \alpha_t\alpha_{t-1}\ldots\alpha_3(1-\alpha_2) + \ldots + \alpha_t(1-\alpha_{t-1}) + (1-\alpha_t) \\
&= \alpha_t[\alpha_{t-1}\ldots\alpha_2(1-\alpha_1) + \alpha_{t-1}\ldots\alpha_3(1-\alpha_2) + \ldots + (1-\alpha_{t-1}) - 1] + 1 \\
&= \alpha_t[\alpha_{t-1}\ldots\alpha_2(1-\alpha_1) + \alpha_{t-1}\ldots\alpha_3(1-\alpha_2) + \ldots + \alpha_{t-1}(1-\alpha_{t-2}) - \alpha_{t-1}] + 1 \\
&= \alpha_t\alpha_{t-1}[\alpha_{t-2}\ldots\alpha_2(1-\alpha_1) + \alpha_{t-2}\ldots\alpha_3(1-\alpha_2) + \ldots + \alpha_{t-2}(1-\alpha_{t-3}) - \alpha_{t-2}] + 1 \\
&= \alpha_t\alpha_{t-1}\ldots\alpha_3[\alpha_2(1-\alpha_1) + (1-\alpha_2) - 1] + 1 \\
&= 1 - \alpha_t\alpha_{t-1}\ldots\alpha_3\alpha_2\alpha_1 \\
&= 1 - \bar{\alpha}_t
\end{aligned} \tag{7}$$

Where $\bar{\alpha}_t = \alpha_t \alpha_{t-1} ... \alpha_3 \alpha_2 \alpha_1$.
Therefore, we can get

$$\begin{aligned} \mathbf{x}_t &= \sqrt{\alpha_t}\mathbf{x}_{t-1} + \sqrt{1-\alpha_t}\epsilon_{\mathbf{t}} \\ &= \sqrt{\alpha_t}\sqrt{\alpha_{t-1}} ... \sqrt{\alpha_1}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \\ &= \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \end{aligned} \tag{8}$$

The above formula (8) is the derivation proof of formula (8) in our article. Given x_t and $\bar{\alpha}_t$, we only need to predict the distribution of ϵ as accurately as possible to obtain x_0. That is

$$x_0 \approx \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\bar{\alpha}_t} \tag{9}$$

Where ϵ_θ is the predicted noise. So, when we fine-tune the unet, we can complete the fine-tuning by minimizing the gap between the actual error $\epsilon(x_t, t)$ and the predicted error $\epsilon_\theta(x_t, t)$ by MSE(Mean Squared Error).

$$\mathcal{L}_{diffusion} = \mathbb{E}_{\mathbf{x}_0, t, \epsilon}\left[|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)|^2\right] \tag{10}$$