A. Supplementary Experiments

A.1. RNA-sequence

Figure 1: The performance comparisons among different models on D2 and D7 of Tdh and Gsn.

Table 1: The Wasserstein error of different models on Supplementary RNA-sequence data sets.

<table>
<thead>
<tr>
<th>Data</th>
<th>Task</th>
<th>Dimension</th>
<th>NN</th>
<th>LEGEND</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Tdh</td>
<td>D2</td>
<td>10</td>
<td>16.28</td>
<td>5.75</td>
<td><strong>2.15</strong></td>
</tr>
<tr>
<td></td>
<td>D7</td>
<td>10</td>
<td>28.19</td>
<td>22.49</td>
<td><strong>1.03</strong></td>
</tr>
<tr>
<td>RNA-Gsn</td>
<td>D2</td>
<td>10</td>
<td>34.94</td>
<td>10.77</td>
<td><strong>3.31</strong></td>
</tr>
<tr>
<td></td>
<td>D7</td>
<td>10</td>
<td>15.74</td>
<td>10.42</td>
<td><strong>2.07</strong></td>
</tr>
</tbody>
</table>

A.2. Daily Trading Volume

Figure 2: (a) to (d): TSLA stock. (e) to (h): GOOGL stock. We predictions of traded volume in next 100 days, RM(yellow) fails to capture the regularities of traded volume in time series, kalman filter based model(green) fails to capture noise information and make reasonable predictions, our model(blue) is able to seize the movements of traded volume and yield better predictions.
Table 2: The Mean absolute percentage error (MAPE) of different models on Daily Trading Volume data sets.

<table>
<thead>
<tr>
<th>Stock</th>
<th>Time</th>
<th>RM</th>
<th>KF</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPM</td>
<td>14:35</td>
<td>0.52</td>
<td>0.28</td>
<td><strong>0.01</strong></td>
</tr>
<tr>
<td></td>
<td>15:15</td>
<td>0.54</td>
<td>0.36</td>
<td><strong>0.04</strong></td>
</tr>
<tr>
<td></td>
<td>15:35</td>
<td>0.51</td>
<td>0.42</td>
<td><strong>0.06</strong></td>
</tr>
<tr>
<td></td>
<td>16:15</td>
<td>0.52</td>
<td>0.49</td>
<td><strong>0.12</strong></td>
</tr>
<tr>
<td>TSLA</td>
<td>14:35</td>
<td>0.53</td>
<td>0.31</td>
<td><strong>0.02</strong></td>
</tr>
<tr>
<td></td>
<td>15:15</td>
<td>0.55</td>
<td>0.36</td>
<td><strong>0.03</strong></td>
</tr>
<tr>
<td></td>
<td>15:35</td>
<td>0.53</td>
<td>0.39</td>
<td><strong>0.08</strong></td>
</tr>
<tr>
<td></td>
<td>16:15</td>
<td>0.52</td>
<td>0.38</td>
<td><strong>0.14</strong></td>
</tr>
<tr>
<td>GOOGL</td>
<td>14:35</td>
<td>0.49</td>
<td>0.35</td>
<td><strong>0.01</strong></td>
</tr>
<tr>
<td></td>
<td>15:15</td>
<td>0.51</td>
<td>0.38</td>
<td><strong>0.03</strong></td>
</tr>
<tr>
<td></td>
<td>15:35</td>
<td>0.53</td>
<td>0.44</td>
<td><strong>0.05</strong></td>
</tr>
<tr>
<td></td>
<td>16:15</td>
<td>0.51</td>
<td>0.42</td>
<td><strong>0.11</strong></td>
</tr>
</tbody>
</table>

B. Definition of $G$ in Synthetic-2

Synthetic-2 (Nonlinear, converging to mixed-Gaussian):

$$\hat{x}_0 \sim N(0, \Sigma_0), \quad \hat{x}_{t+\Delta t} = \hat{x}_t - G\Delta t + \sigma \sqrt{\Delta t} N(0, 1)$$

$$G_{11} = \frac{1}{\sigma_1 N_1 + N_2} \left( \hat{x}_{t1}^1 - \mu_{11} \right) + \frac{1}{\sigma_2 N_1 + N_2} \left( \hat{x}_{t2}^1 - \mu_{21} \right)$$

$$G_{22} = \frac{1}{\sigma_1 N_1 + N_2} \left( \hat{x}_{t1}^2 - \mu_{12} \right) + \frac{1}{\sigma_2 N_1 + N_2} \left( \hat{x}_{t2}^2 - \mu_{22} \right)$$

$$N_1 = \frac{1}{\sqrt{2\pi\sigma_1}} \exp\left(-\frac{(\hat{x}_{t1}^1 - \mu_{11})^2}{2\sigma_1^2} - \frac{(\hat{x}_{t1}^1 - \mu_{12})^2}{2\sigma_2^2}\right)$$

$$N_2 = \frac{1}{\sqrt{2\pi\sigma_2}} \exp\left(-\frac{(\hat{x}_{t2}^1 - \mu_{12})^2}{2\sigma_1^2} - \frac{(\hat{x}_{t2}^2 - \mu_{22})^2}{2\sigma_2^2}\right)$$

C. Error Analysis

In this section, we provide an error analysis of our model. Suppose the hidden dynamics is driven by $g_r(x)$, the dynamics that we learn from data is $g_f(x)$, then original Itô process, Euler processes computed by true $g_r$ and estimated $g_f$ are:

$$dX = g(X)dt + \sigma dW$$

$$x^r_{t+\Delta t} = x^r_t + g_r(x^r_t)\Delta t + \sigma \sqrt{\Delta t} N(0, 1)$$

$$x^f_{t+\Delta t} = x^f_t + g_f(x^f_t)\Delta t + \sigma \sqrt{\Delta t} N(0, 1)$$

where $X$ is the ground truth, $x^r$ is computed by true $g_r$ and $x^f$ is computed by estimated $g_f$. Estimating the error between original Itô process and its Euler form can be very complex, hence we cite the conclusion from (Milstein & Tretyakov, 2013) and focus more on the error between original form and our model.

**Lemma 2.** With the same initial $X_{t_0} = x_{t_0} = x_0$, if there is a global Lipschitz constant $K$ which satisfies:

$$|g(x, t) - g(y, t)| \leq K|x - y|$$

then after $n$ steps, the expectation error between Itô process $x^r_{t_n}$ and Euler forward process $x^f_{t_n}$ is:

$$\mathbb{E}|x^r_{t_n} - x^f_{t_n}| \leq K \left(1 + \mathbb{E}|X_0|^2\right)^{1/2} \Delta t$$
Lemma 2 illustrates that the expectation error between original Itô process and its Euler form is not related to total steps \( n \) but time step \( \Delta t \).

**Proposition 3.** With the same initial \( x_0 \), suppose the generalization error of neural network \( g \) is \( \varepsilon \) and existence of global Lipschitz constant \( K \):

\[
|g(x) - g(y)| \leq K|x - y|
\]

then after \( n \) steps with step size \( \Delta t = T/n \), the expectation error between Itô process \( x_{t_n} \) and approximated forward process \( x'_{t_n} \) is bounded by:

\[
\mathbb{E}[x_{t_n} - x'_{t_n}] \leq \frac{\varepsilon}{K}(e^{KT} - 1) + K(1 + \mathbb{E}|x_0|^2)^{1/2} \Delta t
\]

Proposition 3 implies that besides time step size \( \Delta t \), our expectation error interacts with three factors, generalization error, Lipschitz constant of \( g \) and total time length. In our experiments, we find the best way to decrease the expectation error is reducing the value of \( K \) and \( n \).

**D. Proofs**

**D.1. Proof of Proposition 1**

**Proof.** Suppose \( \hat{x}^{(k)}_{t_m} \) and \( \hat{x}^{(k)}_{t_{m-1}} \) are our observed samples at \( t_m \) and \( t_{m-1} \) respectively, then expectations could be approximated by:

\[
\mathbb{E}_{x \sim \hat{p}(x,t_m)}[f(x)] = \int f(x) \hat{p}(x,t_m)dx = \frac{1}{N} \sum_{k=1}^{N} f(\hat{x}^{(k)}_{t_m})
\]

Then for the second term \( I \) above, it is difficult to calculate directly, but we can use integration by parts to rewrite \( I \) as:

\[
I = \int_{t_{m-1}}^{t_m} \int \left[ \sum_{i=1}^{D} -f(x) \frac{\partial}{\partial x_i}g^i_o(x)p(x,\tau) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} f(x) \frac{\partial^2}{\partial x_i^2}p(x,\tau) \right] dx d\tau
\]

\[
= \int_{t_{m-1}}^{t_m} \int \left[ \sum_{i=1}^{D} g^i_o(x)p(x,\tau) \frac{\partial}{\partial x_i}f(x) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} p(x,\tau) \frac{\partial^2}{\partial x_i^2}f(x) \right] dx d\tau
\]

\[
= \int_{t_{m-1}}^{t_m} \left[ \mathbb{E}_{x \sim p(x,\tau)} \left[ \sum_{i=1}^{D} g^i_o(x) \frac{\partial}{\partial x_i}f(x) \right] + \mathbb{E}_{x \sim p(x,\tau)} \left[ \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2}f(x) \right] \right] d\tau
\]

\[
\approx \int_{t_{m-1}}^{t_m} \frac{1}{N} \sum_{k=1}^{N} \left( \sum_{i=1}^{D} g^i_o(x^{(k)}) \frac{\partial}{\partial x_i}f(x^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2}f(x^{(k)}) \right) d\tau
\]
To approximate the integral from \( t_{m-1} \) to \( t_m \), we adopt trapezoidal rule, then we could rewrite the expectation in Equation (3) as:

\[
E_{x \sim \hat{p}(x, t_m)}[f(x)] \approx \frac{1}{N} \sum_{k=1}^{N} f(\hat{x}_{t_{m-1}}^{(k)}) + \frac{\Delta t}{2} \left[ \frac{1}{N} \sum_{k=1}^{N} \left( \sum_{i=1}^{D} g_u(\hat{x}_{t_{m-1}}^{(k)}) \frac{\partial}{\partial x_i} f(\hat{x}_{t_{m-1}}^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2} f(\hat{x}_{t_{m-1}}^{(k)}) \right) \right] + \frac{1}{N} \sum_{k=1}^{N} \left( \sum_{i=1}^{D} g_u(\hat{x}_{t_m}^{(k)}) \frac{\partial}{\partial x_i} f(\hat{x}_{t_m}^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2} f(\hat{x}_{t_m}^{(k)}) \right) \]

We subtract (2) by (5) to finish the proof.

\[ (6) \]

D.2. Proof of Proposition 2

**Proof.** Given initial \( \hat{x}_{t_0} \), we generate \( \hat{x}_{t_1}, \hat{x}_{t_2}, \hat{x}_{t_3}, \ldots \hat{x}_{t_n} \) sequentially by Euler-Maruyama scheme. Then the expectations can be rewritten as:

\[
E_{x \sim \hat{p}(x, t_n)}[f(x)] = \int f(x) \hat{p}(x, t_n) dx \approx \frac{1}{N} \sum_{k=1}^{N} f(\hat{x}_{t_n}^{(k)})
\]

\[ (7) \]

which is:

\[
E_{x \sim \hat{p}(x, t_n)}[f(x)] \approx \frac{1}{N} \sum_{k=1}^{N} f(\hat{x}_{t_0}^{(k)}) + \int_{t_0}^{t_1} \frac{1}{N} \sum_{k=1}^{N} \left[ \sum_{i=1}^{D} g_u(\hat{x}_{t_0}^{(k)}) \frac{\partial}{\partial x_i} f(\hat{x}_{t_0}^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2} f(\hat{x}_{t_0}^{(k)}) \right] d\tau + \ldots
\]

\[
E_{x \sim \hat{p}(x, t_n)}[f(x)] \approx \frac{1}{N} \sum_{k=1}^{N} f(\hat{x}_{t_0}^{(k)}) + \frac{\Delta t}{2} \left[ \int_{t_0}^{t_1} \left( \sum_{i=1}^{D} \frac{\partial}{\partial x_i} f(\hat{x}_{t_0}^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2} f(\hat{x}_{t_0}^{(k)}) \right) d\tau \right] + \frac{\Delta t}{2} \left[ \int_{t_1}^{t_2} \left( \sum_{i=1}^{D} \frac{\partial}{\partial x_i} f(\hat{x}_{t_1}^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2} f(\hat{x}_{t_1}^{(k)}) \right) d\tau \right] + \ldots
\]

\[ (8) \]

Finally it comes to:

\[
E_{x \sim \hat{p}(x, t_n)}[f(x)] \approx \frac{1}{N} \sum_{k=1}^{N} f(\hat{x}_{t_0}^{(k)}) + \frac{\Delta t}{2} \left[ \int_{t_0}^{t_n} \left( \sum_{i=1}^{D} \frac{\partial}{\partial x_i} f(\hat{x}_{t_0}^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2} f(\hat{x}_{t_0}^{(k)}) \right) d\tau \right] + \frac{\Delta t}{2} \left[ \int_{t_1}^{t_n} \left( \sum_{i=1}^{D} \frac{\partial}{\partial x_i} f(\hat{x}_{t_1}^{(k)}) + \frac{1}{2} \sigma^2 \sum_{i=1}^{D} \frac{\partial^2}{\partial x_i^2} f(\hat{x}_{t_1}^{(k)}) \right) d\tau \right] + \ldots
\]

\[ (9) \]

We subtract (6) by (9) to finish the proof.

D.3. Proof of Error Analysis

**Proof.** The proof process of Lemma 2 is quite long and out of the scope of this paper, for more details please see first two chapters in reference book (Milstein & Tretyakov, 2013). While for the proof of Proposition 3, with initial \( X \) and first one-step iteration:

\[
\begin{cases}
x_t = x_0 \\
x_0 = x_0
\end{cases}
\]

\[ (10) \]

\[
\begin{cases}
x_t = x_t + g(x_t) \Delta t + \sigma \sqrt{\Delta t} \mathcal{N}(0, 1) \\
x_{t_1} = x_{t_0} + g(x_{t_0}) \Delta t + \sigma \sqrt{\Delta t} \mathcal{N}(0, 1)
\end{cases}
\]

\[ (11) \]
Then we have:

$$\mathbb{E}|x_{t_0}^r - x_{t_0}^f| = \mathbb{E}|x_{t_0} - x_{t_0}| = 0 \quad (12)$$

$$\mathbb{E}|x_{t_1}^r - x_{t_1}^f| = \mathbb{E}|x_{t_0}^r - x_{t_0}^f + g_r(x_{t_0}^r)\Delta t - g_f(x_{t_0}^f)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0, 1) - \sigma\sqrt{\Delta t}\mathcal{N}(0, 1)|$$

$$\leq \mathbb{E}|x_{t_0}^r - x_{t_0}^f| + \mathbb{E}|g_r(x_{t_0}^r) - g_f(x_{t_0}^f)|\Delta t$$

$$= \mathbb{E}|g_r(x_{t_0}^r) - g_f(x_{t_0}^r) + g_f(x_{t_0}^f)|\Delta t$$

$$\leq \varepsilon\Delta t + \mathbb{E}|g_f(x_{t_0}^f)|\Delta t \quad (x_{t_0}^f \in [x_{t_0}^r, x_{t_0}^f])$$

$$\leq \varepsilon\Delta t + K\mathbb{E}|x_{t_0}^r - x_{t_0}^f|\Delta t$$

$$= \varepsilon\Delta t$$ \quad (13)

Follow the pattern we have:

$$\begin{cases} x_{t_2}^r = x_{t_1}^r + g_r(x_{t_1}^r)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0, 1) \\
 x_{t_2}^f = x_{t_1}^f + g_f(x_{t_1}^f)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0, 1) \\
 \vdots \\
 x_{t_n}^r = x_{t_{n-1}}^r + g_r(x_{t_{n-1}}^r)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0, 1) \\
 x_{t_n}^f = x_{t_{n-1}}^f + g_f(x_{t_{n-1}}^f)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0, 1) \quad (14) \end{cases}$$

Which leads to:

$$\mathbb{E}|x_{t_2}^r - x_{t_2}^f| = \mathbb{E}|x_{t_1}^r - x_{t_1}^f + g_r(x_{t_1}^r)\Delta t - g_f(x_{t_1}^f)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(0, 1) - \sigma\sqrt{\Delta t}\mathcal{N}(0, 1)|$$

$$\leq \mathbb{E}|x_{t_1}^r - x_{t_1}^f| + \mathbb{E}|g_r(x_{t_1}^r) - g_f(x_{t_1}^f)|\Delta t$$

$$\leq \varepsilon\Delta t + K\mathbb{E}|x_{t_1}^r - x_{t_1}^f|\Delta t$$

$$\leq (1 + K\Delta t)\varepsilon\Delta t + \varepsilon\Delta t$$ \quad (15)

$$\mathbb{E}|x_{t_n}^r - x_{t_n}^f| \leq \varepsilon\Delta t \sum_{i=0}^{n-1} (1 + K\Delta t)^i \quad (16)$$

Now let $S = \sum_{i=0}^{n-1} (1 + K\Delta t)^i$, then consider followings:

$$S(K\Delta t) = S(1 + K\Delta t) - S$$

$$= \sum_{i=1}^{n} (1 + K\Delta t)^i - \sum_{i=0}^{n-1} (1 + K\Delta t)^i$$

$$= (1 + K\Delta t)^n - 1$$

$$= (1 + K\frac{T}{n})^n - 1$$

$$\leq e^{KT} - 1$$ \quad (18)

Finally we have:

$$\mathbb{E}|x_{t_n}^r - x_{t_n}^f| \leq \frac{\varepsilon}{K}(e^{KT} - 1) \quad (19)$$

$$\mathbb{E}|x_{t_n} - x_{t_n}^f| \leq \frac{\varepsilon}{K}(e^{KT} - 1) + K(1 + E|x_0|^2)^{1/2}\Delta t \quad (20)$$
References