DiP: Taming Diffusion Models in Pixel Space

Supplementary Material

A. Why Patch Detailer Head: A Theoretical Perspective

In this section, we try to provide a simplified theoretical analysis to further elucidate why we need local detail refinement in enhancing generation quality. From a general insight, we argue that DiT primarily focuses on the layout and arrangement of the dominant elements in the image, or in other words, the low-frequency signals of the global data. Consequently, it is less effective to learn local details and high-frequency signals. Through the refinement structure, we directly inject all signals from the global data into the learning process, which substantially improves the fine-grained processing of these high-frequency details.

Specifically, we follow the flow matching description of the diffusion process. Given an initial data sample $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x}_0) \in \mathbb{R}^d$ as the input, a Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I}_d)$, and $t \in [0, 1]$, let

$$\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\epsilon. \tag{1}$$

In this paper, since all inputs are partitioned into patches of equal size, we first define the patch-level input as follows.

Definition A.1 (Patch-level Input)**.** *For each input $\mathbf{x}_0 \in \mathbb{R}^d$, we define the patch-level input as $\left\{\mathbf{x}_0^{(s)}\right\}_{s=1}^N$, where $\mathbf{x}_0^{(s)} \in \mathbb{R}^p$ and $\mathbf{x}_0 = \left[\left(\mathbf{x}_0^{(1)}\right)^\top, \cdots, \left(\mathbf{x}_0^{(N)}\right)^\top\right]^\top$, $Np = d$.*

It is natural to represent the patch-level input by a series of selection matrices $\left\{\mathbf{P}^{(s)}\right\}_{s=1}^N$. For each s, $\mathbf{P}^{(s)} \in \mathbb{R}^{p\times d}$ satisfies $\mathbf{P}^{(s)}\left(\mathbf{P}^{(s)}\right)^\top = \mathbf{I}_p$ and $\mathbf{P}^{(s)}\mathbf{x}_0 = \mathbf{x}_0^{(s)}$. The flow-based models try to minimize a loss function defined as $\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|f(\mathbf{x}_t, t) - (\epsilon - \mathbf{x}_0)\|^2\right]$. Assuming that each patch is independent of one another, a patch-level predictor $f(\cdot, t)$ tries to estimate the patch-level objective field $\hat{v}^{(s)} = f\left(\mathbf{x}_t^{(s)}, t\right)$ for each patch-level noised input $\mathbf{x}_t^{(s)} = (1-t)\mathbf{x}_0^{(s)} + t\epsilon^{(s)}$, where $\epsilon^{(s)} = \mathbf{P}^{(s)}\epsilon \sim \mathcal{N}(0, \mathbf{I}_p)$. For the given $\mathcal{L}_{\text{FM}}$, the optimal predictor is the conditional expectation

$$\hat{v}^{(s),*} = \mathbb{E}\left[\epsilon^{(s)} - \mathbf{x}_0^{(s)} \,\middle|\, \mathbf{x}_t^{(s)}\right].$$

However, in true generation tasks, each patch is not independent of others, because, for natural images, the boundaries between adjacent patches are typically continuous and smoothly varying (e.g., there is little difference between one patch of sky and another). The correlation between patches only weakens when an abrupt transition occurs in the image's elements, such as at the boundary between sky and grass. Moreover, DiT's attention-based structure allows a single patch to access partial information from all other patches. Although this information may be coarse, this remains a complex, coupled structure. Therefore, for a DiT model, the estimate of $\hat{v}^{(s)}$ is not only based on $\mathbf{x}_t^{(s)}$ but also some other information from $\left\{\mathbf{x}_t^{(l)}\right\}_{l\neq s}$. Thus, we define the effective information below.

Definition A.2 (Effective Information)**.** *For a patch-level noised input $\left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N$, we define $\text{EI}^{(s)}\left(f; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right)$ to represent the effective information used for a generation model f to estimate the patch-level vector field $\hat{v}^{(s)}$ for any $s \in [N]$.*

Assuming that each patch is independent of one another, the patch-level estimate $\hat{v}^{(s)} = f\left(\mathbf{x}_t^{(s)}, t\right)$ only uses $\mathbf{x}_t^{(s)}$ for prediction, which means $\text{EI}^{(s)}\left(f; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right) = \left\{\mathbf{x}_t^{(s)}\right\}$. Thus the optimal predictor can be more generally formulated as $\hat{v}^{(s),*} = \mathbb{E}\left[\epsilon^{(s)} - \mathbf{x}_0^{(s)} \,\middle|\, \text{EI}^{(s)}\left(f; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right)\right]$. For attention-based generation models, we cannot accurately obtain the effective information due to the complex coupling structure. However, based on some standard assumptions on the initial data distribution and some empirical observations, we can still give a brief formulation for the effective information.

Assumption A.3 (Data Distribution)**.** *For the initial data distribution, we assume that* $p_{\text{data}} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$*,* $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_d]$*, and* $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_d\}$*.*

Assumption A.4 (Eigenvalue Decay)**.** *There exists* $\alpha > 1$ *such that for any* $i \in [d]$*, the eigenvalues of* $\mathbf{\Sigma}$ *satisfies* $\lambda_i \asymp i^{-a}$*.*

Assumption A.3 and A.4 characterize the data distribution as a Gaussian distribution with a covariance of a series of fast-decay eigenvalues. The eigenvalue decay of covariance characterizes the differences in high- and low-frequency signals of the image information. This is consistent with the empirical observation that DiT can effectively learn low-frequency signals but has difficulty capturing high-frequency signals. Given $b > 0$, we can decompose the input $\mathbf{x}_0$ into low- and high-frequency components as

$$\mathbf{x}_0 = \mu + \mathbf{x}_{0,\text{low}} + \mathbf{x}_{0,\text{high}}, \tag{2}$$

where $\mathbf{x}_{0,\text{low}} \sim \mathcal{N}(0, \mathbf{\Sigma}_{\text{low}})$ and $\mathbf{x}_{0,\text{high}} \sim \mathcal{N}(0, \mathbf{\Sigma}_{\text{high}})$. $\mathbf{\Sigma}_{\text{low}}$ satisfies $\mathbf{\Sigma}_{\text{low}} = \mathbf{U}_r\mathbf{\Lambda}_r\mathbf{U}_r^\top = \sum_{i=1}^r \lambda_i \mathbf{u}_i\mathbf{u}_i^\top$ where $\lambda_r > b$ and $\lambda_{r+1} \le b$, and $\mathbf{\Sigma}_{\text{high}} = \mathbf{\Sigma} - \mathbf{\Sigma}_{\text{low}}$. Thus we can decompose the patch-level noised input $\mathbf{x}_t^{(s)}$ as

$$\mathbf{x}_t^{(s)} = \underbrace{(1-t)\mathbf{P}^{(s)}\mu}_{\text{Mean}^{(s)}} + \underbrace{(1-t)\mathbf{P}^{(s)}\mathbf{x}_{0,\text{low}}}_{\text{Low}^{(s)}} + \underbrace{(1-t)\mathbf{P}^{(s)}\mathbf{x}_{0,\text{high}}}_{\text{High}^{(s)}} + \underbrace{t\mathbf{P}^{(s)}\epsilon}_{\text{Noise}^{(s)}} \tag{3}$$

We can assume that for the DiT model, the effective information is composed of the local patch itself and the low-frequency signals of other patches as below.

Assumption A.5 (EI of DiT)**.** *Given DiT as the predictor, there exists* $\beta > 0$ *such that for any* $s \in [N]$*, the effective information to estimate the patch-level vector* $\hat{v}^{(s)}$ *satisfies*

$$\text{EI}^{(s)}\left(\text{DiT}; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right) = \left\{\mathbf{x}_t^{(s)}\right\} \cup \left\{\mathbf{x}_{t,\text{low}}^{(l)}\right\}_{l \neq s}, \tag{4}$$

where

$$\mathbf{x}_{t,\text{low}}^{(l)} = \text{Mean}^{(l)} + \text{Low}^{(l)} + \text{Noise}^{(l)} \tag{5}$$

for all $l \neq s$*.*

Our refinement structure directly injects all signals from the initial data $\mathbf{x}_0$ for prediction, which means that for DiP, the effective information satisfies

$$\text{EI}^{(s)}\left(\text{DiP}; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right) = \text{EI}^{(s)}\left(\text{DiT}; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right) \cup \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N = \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N. \tag{6}$$

We define $\hat{v}_{\text{DiT}}^{(s)} = \mathbb{E}\left[\epsilon^{(s)} - \mathbf{x}_0^{(s)} \,\middle|\, \text{EI}^{(s)}\left(\text{DiT}; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right)\right]$ and $\hat{v}_{\text{DiP}}^{(s)} = \mathbb{E}\left[\epsilon^{(s)} - \mathbf{x}_0^{(s)} \,\middle|\, \text{EI}^{(s)}\left(\text{DiP}; \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^N\right)\right]$ as the general near-optimal estimate of DiT and DiP, respectively. Then we obtain the main results below.

Theorem A.6. *Assume that Assumption A.3, A.4 and A.5 hold. Consider using DiT and DiP for the diffusion generation task as the predictor, respectively. The general near-optimal estimate* $\hat{v}_{\text{DiT}}^{(s)}$ *and* $\hat{v}_{\text{DiP}}^{(s)}$ *satisfy*

$$\hat{v}_{\text{DiT}}^{(s)} = \mathbf{P}^{(s)}\hat{\mathbf{B}}\hat{\mathbf{M}}\left(\mathbf{x}_t - (1-t)\mu\right) - \mathbf{P}^{(s)}\mu, \tag{7}$$

and

$$\hat{v}_{\text{DiP}}^{(s)} = \mathbf{P}^{(s)}\mathbf{A}\mathbf{M}\left(\mathbf{x}_t - (1-t)\mu\right) - \mathbf{P}^{(s)}\mu, \tag{8}$$

respectively, where

$$\begin{aligned}
&\hat{\mathbf{M}} = \left[(1-t)^2\mathbf{\Sigma}_{\text{low}} + t^2\mathbf{I}_d + (1-t)^2\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{\Sigma}_{\text{high}}\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\right]^{-1}, \\
&\hat{\mathbf{B}} = t\mathbf{I}_d - (1-t)\mathbf{B}, \quad \mathbf{B} = \mathbf{\Sigma}_{\text{low}} + \mathbf{\Sigma}_{\text{high}}\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}, \\
&\mathbf{M} = \left[(1-t)^2\mathbf{\Sigma} + t^2\mathbf{I}_d\right]^{-1}, \\
&\mathbf{A} = t\mathbf{I}_d - (1-t)\mathbf{\Sigma}.
\end{aligned} \tag{9}$$

The denoising operator $\mathbf{P}^{(s)}\hat{\mathbf{B}}\hat{\mathbf{M}}$ *and* $\mathbf{P}^{(s)}\mathbf{A}\mathbf{M}$ *satisfies*

$$\mathbf{P}^{(s)}\hat{\mathbf{B}}\hat{\mathbf{M}} \asymp \sum_{i=1}^{r} \frac{t-(1-t)\lambda_i}{(1-t)^2\lambda_i + t^2}\mathbf{v}_i\mathbf{u}_i^\top + \sum_{i=r+1}^{d} \frac{\lambda_i}{t}\mathbf{v}_i\mathbf{u}_i^\top + \mathcal{I}_1 + \mathcal{I}_2, \tag{10}$$

and

$$\mathbf{P}^{(s)}\mathbf{A}\mathbf{M} = \sum_{i=1}^{d} \frac{t-(1-t)\lambda_i}{(1-t)^2\lambda_i + t^2}\mathbf{v}_i\mathbf{u}_i^\top, \tag{11}$$

respectively, where $[\mathbf{v}_1, \cdots, \mathbf{v}_d] = \mathbf{P}^{(s)}[\mathbf{u}_1, \cdots, \mathbf{u}_d]$, $\mathcal{I}_1 = -\sum_{i=r+1}^{d}\sum_{j=1}^{r} \frac{(1-t)\lambda_i}{(1-t)^2\lambda_j + t^2}\left[\mathbf{u}_i^\top \left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{u}_j\right]\mathbf{v}_i\mathbf{u}_j^\top$ *and* $\mathcal{I}_2 = -\sum_{i=r+1}^{d}\sum_{j=r+1}^{d} \frac{(1-t)\lambda_i}{t^2}\left[\mathbf{u}_i^\top \left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{u}_j\right]\mathbf{v}_i\mathbf{u}_j^\top$.

Proof. See Appendix B. □

Remark A.7. *Theorem A.6 illustrates that our designed refinement mechanism exhibits a strong adaptive correction capability for high-frequency signals within the image. Specifically, (7) and (8) align with our objective to estimate the conditional expectation of* $\epsilon - \mathbf{x}_0$ *through DiT and DiP. The first term of (7) and (8) represents the estimate of noise, while the second term means the estimate of original data. We focus on the first term regarded as the "denoising process", with respective "denoising operator" (10) and (11) serve as a global representation characterization, performing denoising on different frequency-domain components derived from the original image.*

- *When using only DiT for estimation, Equation (10) indicates that DiT can achieve a good adaptive fit for low-frequency signals (*$i \le r$*, dominant signals). However, for high-frequency components in the image, relying solely on DiT may not provide sufficient representational capacity for these components. Specifically,*
  - *The first term in (10) corresponds to the denoising process applied to the low-frequency signals. Although all these belong to the low-frequency regime, those with larger values of* λ_i *(*$i \le r$*) contain a stronger proportion of the original image content relative to noise. Consequently, as* λ_i *decreases, the denoising operator adaptively selects a larger correction magnitude. (It can be readily demonstrated that* $\frac{t-(1-t)\lambda_i}{(1-t)^2\lambda_i+t^2}$ *exhibits a monotonic increase as* λ_i *decreases.)*
  - *The remaining three terms correspond to the denoising process applied to the high-frequency components. Among them,* $\mathcal{I}_1$ *represents the consistent influence exerted by the low-frequency signal on the high-frequency ones. Since the high-frequency components are associated with* λ_i *(*$i \ge r+1$*) that are significantly smaller than those of the low-frequency regime (Assumption A.4), this term can generally be regarded as* $o(\mathbf{1})$*. The second term and* $\mathcal{I}_2$*—particularly the second term—point to the following fact: when* $t \to 1$ *in the early stage of denoising, the magnitude of* λ_i *is much smaller than* t*, leading the model to apply only negligible corrections to the high-frequency components. This weakens the model's ability to learn from this portion of the signal. Conversely, as* $t \to 0$ *in the late stage of denoising, where the model aims to learn a sensitive compensatory mechanism to capture more fine-grained details from the original image,* t *becomes much smaller than* λ_i*, causing the model's corrections to high-frequency components to lose stability. As a result, these corrections may introduce inconsistencies with the previously learned representations, potentially affecting fine details of the final output.*
- *In contrast, when using DiP for estimation, Equation (11) demonstrates that DiP can provide a robust adaptive correction for all signals. This is attributed to our refinement mechanism, which, at a low computational cost, enhances the effective information during the denoising process. Particularly for high-frequency signals, the refinement provides a powerful supplement to the information that DiT struggles to capture, which aligns with both our intuition and experimental results.*

B. Proof of Theorem A.6

Proof. Based on (6), we have $\hat{v}_{\text{DiP}}^{(s)} = \mathbb{E}\left[\epsilon^{(s)} - \mathbf{x}_0^{(s)} \,\middle|\, \left\{\mathbf{x}_t^{(s)}\right\}_{s=1}^{N}\right] = \mathbb{E}\left[\epsilon^{(s)} \mid \mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_0^{(s)} \,\middle|\, \mathbf{x}_t\right]$. We first obtain the following statistics to obtain the first term $\mathbb{E}\left[\epsilon^{(s)} \mid \mathbf{x}_t\right]$.
Expectations:

$$\mathbb{E}\left[\epsilon^{(s)}\right] = 0, \;\; \mathbb{E}\left[\mathbf{x}_t\right] = (1-t)\mu, \tag{12}$$

Covariances:

$$\text{Cov}\left(\epsilon^{(s)}, \mathbf{x}_t\right) = \text{Cov}\left(\mathbf{P}^{(s)}\epsilon, (1-t)\mathbf{x}_0 + t\epsilon\right) = t\mathbf{P}^{(s)}\text{Cov}(\epsilon, \epsilon) = t\mathbf{P}^{(s)}, \tag{13}$$

and

$$\mathrm{Cov}(\mathbf{x}_t) = \mathrm{Cov}\left((1-t)\mathbf{x}_0 + t\epsilon\right) = (1-t)^2\mathrm{Cov}(\mathbf{x}_0) + t^2\mathrm{Cov}(\epsilon) = (1-t)^2\mathbf{\Sigma} + t^2\mathbf{I}_d. \tag{14}$$

Then we use $\mathbb{E}[Y|X] = \mathbb{E}Y + \mathrm{Cov}(Y,X)\mathrm{Cov}(X,X)^{-1}(X - \mathbb{E}X)$ to obtain that

$$\mathbb{E}\left[\epsilon^{(s)} \,\middle|\, \mathbf{x}_t\right] = t\mathbf{P}^{(s)}\left[(1-t)^2\mathbf{\Sigma} + t^2\mathbf{I}_d\right]^{-1}(\mathbf{x}_t - (1-t)\mu). \tag{15}$$

Similarly, for the second term of $\hat{v}_{\mathrm{DiP}}^{(s)}$ we have

$$\mathrm{Cov}\left(\mathbf{x}_0^{(s)}, \mathbf{x}_t\right) = \mathrm{Cov}\left(\mathbf{P}^{(s)}\mathbf{x}_0, (1-t)\mathbf{x}_0 + t\epsilon\right) = (1-t)\mathbf{P}^{(s)}\mathrm{Cov}(\mathbf{x}_0, \mathbf{x}_0) = (1-t)\mathbf{P}^{(s)}\mathbf{\Sigma}, \tag{16}$$

Then we obtain

$$\mathbb{E}\left[\mathbf{x}_0^{(s)} \,\middle|\, \mathbf{x}_t\right] = \mathbf{P}^{(s)}\mu + (1-t)\mathbf{P}^{(s)}\mathbf{\Sigma}\left[(1-t)^2\mathbf{\Sigma} + t^2\mathbf{I}_d\right]^{-1}(\mathbf{x}_t - (1-t)\mu). \tag{17}$$

Thus we have

$$\begin{aligned}\hat{v}_{\mathrm{DiP}}^{(s)} &= \mathbb{E}\left[\epsilon^{(s)} \,\middle|\, \mathbf{x}_t\right] - \mathbb{E}\left[\mathbf{x}_0^{(s)} \,\middle|\, \mathbf{x}_t\right] \\ &= \mathbf{P}^{(s)}\left[t\mathbf{I}_d - (1-t)\mathbf{\Sigma}\right]\left[(1-t)^2\mathbf{\Sigma} + t^2\mathbf{I}_d\right]^{-1}(\mathbf{x}_t - (1-t)\mu) - \mathbf{P}^{(s)}\mu.\end{aligned} \tag{18}$$

Letting $\mathbf{M} = \left[(1-t)^2\mathbf{\Sigma} + t^2\mathbf{I}_d\right]^{-1}$, $[\mathbf{v}_1, \cdots, \mathbf{v}_d] = \mathbf{P}^{(s)}[\mathbf{u}_1, \cdots, \mathbf{u}_d]$, $\mathbf{A} = t\mathbf{I}_d - (1-t)\mathbf{\Sigma}$, we have

$$\begin{aligned}\mathbf{P}^{(s)}\mathbf{A}\mathbf{M} &= \left(\sum_{j=1}^{d}\mathbf{v}_j\mathbf{u}_j^\top\right)\left(\sum_{i=1}^{d}\frac{t-(1-t)\lambda_i}{(1-t)^2\lambda_i + t^2}\mathbf{u}_i\mathbf{u}_i^\top\right) \\ &= \sum_{i=1}^{d}\frac{t-(1-t)\lambda_i}{(1-t)^2\lambda_i + t^2}\mathbf{v}_i\mathbf{u}_i^\top.\end{aligned} \tag{19}$$

Similarly, based on Assumption A.5, we have $\hat{v}_{\mathrm{DiT}}^{(s)} = \mathbb{E}\left[\epsilon^{(s)} - \mathbf{x}_0^{(s)} \,\middle|\, \left\{\mathbf{x}_t^{(s)}\right\} \cup \left\{\mathbf{x}_{t,\mathrm{low}}^{(l)}\right\}_{l\neq s}\right]$, where $\mathbf{x}_{t,\mathrm{low}}^{(l)} = (1-t)\mathbf{P}^{(l)}\mu + (1-t)\mathbf{P}^{(l)}\mathbf{x}_{0,\mathrm{low}} + t\mathbf{P}^{(l)}\epsilon$. We first use one vector to represent the condition $\left\{\mathbf{x}_t^{(s)}\right\} \cup \left\{\mathbf{x}_{t,\mathrm{low}}^{(l)}\right\}_{l\neq s}$. We try to construct an observation $\hat{\mathbf{x}}_t$ such that at s patch, $\mathbf{P}^{(s)}\hat{\mathbf{x}}_t = \mathbf{x}_t^{(s)}$, and at $l \neq s$ patch $\mathbf{P}^{(l)}\hat{\mathbf{x}}_t = \mathbf{x}_{t,\mathrm{low}}^{(l)}$. The following $\hat{\mathbf{x}}_t$ satisfies the requirement above

$$\hat{\mathbf{x}}_t = (1-t)\mu + (1-t)\mathbf{x}_{0,\mathrm{low}} + t\epsilon + (1-t)\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\mathbf{x}_{0,\mathrm{high}}. \tag{20}$$

Now we use the same technique to obtain $\mathbb{E}\left[\epsilon^{(s)} \,\middle|\, \hat{\mathbf{x}}_t\right]$. The covariance terms satisfy

$$\mathrm{Cov}(\epsilon^{(s)}, \hat{\mathbf{x}}_t) = \mathrm{Cov}\left(\mathbf{P}^{(s)}\epsilon, t\epsilon\right) = t\mathbf{P}^{(s)}, \tag{21}$$

and

$$\begin{aligned}\mathrm{Cov}(\hat{\mathbf{x}}_t) &= \mathrm{Cov}\left((1-t)\mathbf{x}_{0,\mathrm{low}} + t\epsilon + (1-t)\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\mathbf{x}_{0,\mathrm{high}}\right) \\ &= (1-t)^2\mathrm{Cov}(\mathbf{x}_{0,\mathrm{low}}) + t^2\mathrm{Cov}(\epsilon) + (1-t)^2\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\mathrm{Cov}(\mathbf{x}_{0,\mathrm{high}})\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)} \\ &= (1-t)^2\mathbf{\Sigma}_{\mathrm{low}} + t^2\mathbf{I}_d + (1-t)^2\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\mathbf{\Sigma}_{\mathrm{high}}\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}.\end{aligned} \tag{22}$$

Thus we have

$$\mathbb{E}\left[\epsilon^{(s)} \,\middle|\, \hat{\mathbf{x}}_t\right] = t\mathbf{P}^{(s)}\left[(1-t)^2\mathbf{\Sigma}_{\mathrm{low}} + t^2\mathbf{I}_d + (1-t)^2\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\mathbf{\Sigma}_{\mathrm{high}}\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\right]^{-1}(\hat{\mathbf{x}}_t - (1-t)\mu). \tag{23}$$

For $\mathbb{E}\left[\mathbf{x}_0^{(s)} \,\middle|\, \hat{\mathbf{x}}_t\right]$, we have

$$\begin{aligned}
\operatorname{Cov}\left(\mathbf{x}_0^{(s)}, \hat{\mathbf{x}}_t\right) &= \operatorname{Cov}\left(\mathbf{P}^{(s)}\mathbf{x}_{0,\text{low}} + \mathbf{P}^{(s)}\mathbf{x}_{0,\text{high}}, (1-t)\mu + (1-t)\mathbf{x}_{0,\text{low}} + t\epsilon + (1-t)\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{x}_{0,\text{high}}\right) \\
&= \operatorname{Cov}\left(\mathbf{P}^{(s)}\mathbf{x}_{0,\text{low}}, (1-t)\mathbf{x}_{0,\text{low}}\right) + \operatorname{Cov}\left(\mathbf{P}^{(s)}\mathbf{x}_{0,\text{high}}, \left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{x}_{0,\text{high}}\right) \\
&= (1-t)\mathbf{P}^{(s)}\left[\mathbf{\Sigma}_{\text{low}} + \mathbf{\Sigma}_{\text{high}}\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\right].
\end{aligned} \tag{24}$$

Thus we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathbf{x}_0^{(s)} \,\middle|\, \hat{\mathbf{x}}_t\right] =& \mathbf{P}^{(s)}\mu + (1-t)\mathbf{P}^{(s)}\left[\mathbf{\Sigma}_{\text{low}} + \mathbf{\Sigma}_{\text{high}}\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\right] \times \\
& \left[(1-t)^2\mathbf{\Sigma}_{\text{low}} + t^2\mathbf{I}_d + (1-t)^2\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{\Sigma}_{\text{high}}\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\right]^{-1}\left(\mathbf{x}_t - (1-t)\mu\right).
\end{aligned} \tag{25}$$

Finally we have

$$\begin{aligned}
\hat{v}_{\text{DiT}}^{(s)} &= \mathbb{E}\left[\epsilon^{(s)} \,\middle|\, \hat{\mathbf{x}}_t\right] - \mathbb{E}\left[\mathbf{x}_0^{(s)} \,\middle|\, \hat{\mathbf{x}}_t\right] \\
&= \mathbf{P}^{(s)}\left[t\mathbf{I}_d - (1-t)\mathbf{B}\right]\hat{\mathbf{M}}\left(\mathbf{x}_t - (1-t)\mu\right) - \mathbf{P}^{(s)}\mu,
\end{aligned} \tag{26}$$

where $\mathbf{B} = \mathbf{\Sigma}_{\text{low}} + \mathbf{\Sigma}_{\text{high}}\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}$ and $\hat{\mathbf{M}} = \left[(1-t)^2\mathbf{\Sigma}_{\text{low}} + t^2\mathbf{I}_d + (1-t)^2\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{\Sigma}_{\text{high}}\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\right]^{-1}$.

Letting $\hat{\mathbf{B}} = t\mathbf{I}_d - (1-t)\mathbf{B}$, we have

$$\mathbf{P}^{(s)}\hat{\mathbf{B}}\hat{\mathbf{M}} = \left(\sum_{j=1}^{d}\mathbf{v}_j\mathbf{u}_j^\top\right)\left(\hat{\mathbf{C}}_1 + \hat{\mathbf{C}}_2\right)\left(\hat{\mathbf{D}} + \hat{\mathbf{E}}\right)^{-1}, \tag{27}$$

where

$$\begin{aligned}
\hat{\mathbf{C}}_1 &= \sum_{i=1}^{r}\left(t - (1-t)\lambda_i\right)\mathbf{u}_i\mathbf{u}_i^\top \\
\hat{\mathbf{C}}_2 &= t\sum_{i=r+1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top - (1-t)\sum_{i=r+1}^{d}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}, \\
\hat{\mathbf{D}} &= \sum_{i=1}^{r}((1-t)^2\lambda_i + t^2)\mathbf{u}_i\mathbf{u}_i^\top + \sum_{i=r+1}^{d}t^2\mathbf{u}_i\mathbf{u}_i^\top, \\
\hat{\mathbf{E}} &= (1-t)^2\sum_{i=r+1}^{d}\lambda_i\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}\mathbf{u}_i\mathbf{u}_i^\top\left(\mathbf{P}^{(s)}\right)^\top \mathbf{P}^{(s)}.
\end{aligned} \tag{28}$$

We notice that $\hat{\mathbf{D}}$ is a positive diagonal matrix and $\hat{\mathbf{D}}^{-1}\hat{\mathbf{E}} \asymp o(1)$ because Assumption A.4 shows that $\lambda_p/\lambda_q \asymp (q/p)^a \asymp o(1)$ for any $1 \le q \le r$ and $p \ge r+1$. Thus due to first-order Taylor expansion we have

$$\left(\hat{\mathbf{D}} + \hat{\mathbf{E}}\right)^{-1} = \left(\mathbf{I}_d + \hat{\mathbf{D}}^{-1}\hat{\mathbf{E}}\right)^{-1}\hat{\mathbf{D}}^{-1} \approx \left(\mathbf{I}_d - \hat{\mathbf{D}}^{-1}\hat{\mathbf{E}}\right)\hat{\mathbf{D}}^{-1} = \hat{\mathbf{D}}^{-1} - \hat{\mathbf{D}}^{-1}\hat{\mathbf{E}}\hat{\mathbf{D}}^{-1} \asymp \hat{\mathbf{D}}^{-1}. \tag{29}$$

Therefore we obtain

$$\begin{aligned}
\mathbf{P}^{(s)}\hat{\mathbf{B}}\hat{\mathbf{M}} &\asymp \left(\sum_{j=1}^{d}\mathbf{v}_j\mathbf{u}_j^\top\right)\left(\hat{\mathbf{C}}_1+\hat{\mathbf{C}}_2\right)\hat{\mathbf{D}}^{-1} \\
&= \sum_{i=1}^{r}\frac{t-(1-t)\lambda_i}{(1-t)^2\lambda_i+t^2}\mathbf{v}_i\mathbf{u}_i^\top + \sum_{i=r+1}^{d}\frac{\lambda_i}{t}\mathbf{v}_i\mathbf{u}_i^\top \\
&\quad \underbrace{-\sum_{i=r+1}^{d}\sum_{j=1}^{r}\frac{(1-t)\lambda_i}{(1-t)^2\lambda_j+t^2}\left[\mathbf{u}_i^\top\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\mathbf{u}_j\right]\mathbf{v}_i\mathbf{u}_j^\top}_{\mathcal{I}_1} \\
&\quad \underbrace{-\sum_{i=r+1}^{d}\sum_{j=r+1}^{d}\frac{(1-t)\lambda_i}{t^2}\left[\mathbf{u}_i^\top\left(\mathbf{P}^{(s)}\right)^\top\mathbf{P}^{(s)}\mathbf{u}_j\right]\mathbf{v}_i\mathbf{u}_j^\top}_{\mathcal{I}_2}.
\end{aligned} \tag{30}$$

We finish the proof. □

C. More Implementation Details

DiT Architecture	
Input dim	256×256×3
Num. layers	26
Hidden dim.	1152
Num. heads	16
Patch Detailer Head Architecture	
DownSampling path	16→8→4→2→1
UpSampling path	1→2→4→8→16
DownSampling channel	3→64→128→256→512
Bottleneck	(512+1152)→512
UpSampling channel	512→256→128→64→64
Output Layer	64→3
Optimization	
Optimizer	AdamW
Learning rate	0.0001
Weight decay	0
Batch size	256
Interpolants	
Diffusion sampler	Euler
Diffusion steps	100
Evaluation suite	ADM

Table 1. Hyperparameter settings.

Hypermarameters. Table 1 reports the detailed hyperparameters of DiP, including the DiT Architecture, Patch Detailer Head Architecture, Optimization, and Interpolants.

Objective. DiP follows the training objectives of DDT [**?**]. It is trained using flow matching as the objective function and regularized using representation alignment techniques. Further improvements could be made by introducing adversarial loss [**?**], perceptual loss [**?**].

Sampler. We use Euler-Maruyama ODE sampler with 100 sampling steps by default. For DiT-only and DiP, we used the same inference hyperparameters.

Classifier-Free Guidance. In our experiments, we employ Interval-based Classifier-Free Guidance [**?**] (Interval-CFG). Specifically, we set the guidance scale to cfg=2.9. The guidance is activated exclusively within the normalized timestep interval of [0.11, 0.97].

D. How to Preserving High-Frequency Signal: Patch or Image

While the theoretical analysis in Appendix A establishes the need for all-frequency raw signals to refine missing high-frequency details, we also focus on how can this information be injected in the most effective way. Specifically, we are interested in whether patch-level input is better than image-level input, or vise versa, as shown in Figure 1. Intuitively, the transformer structure in DiT has captured the long-distance dependencies, therefore we only need the specific high-frequency signals or details of the image. A toy experiment in Figure 2 verifies this intuition.

In Figure 2(a), through the patch-level input, the learned manifold (black) tightly adheres to the ground truth structure (orange), effectively capturing intricate branching patterns and sharp boundaries. In contrast, Figure 2(b) reveals that global processing leads to over-smoothing. The learned distribution is more dispersed and struggles to lock onto fine structural details. This suggests that we only need the refinement structure to dedicate its capacity to high-frequency sensing without being distracted by long-distance dependencies. On the contrary, with image-level input the network tends to average out features across a broader spatial regime, resulting in a loss of sharp details in high-frequency regions.

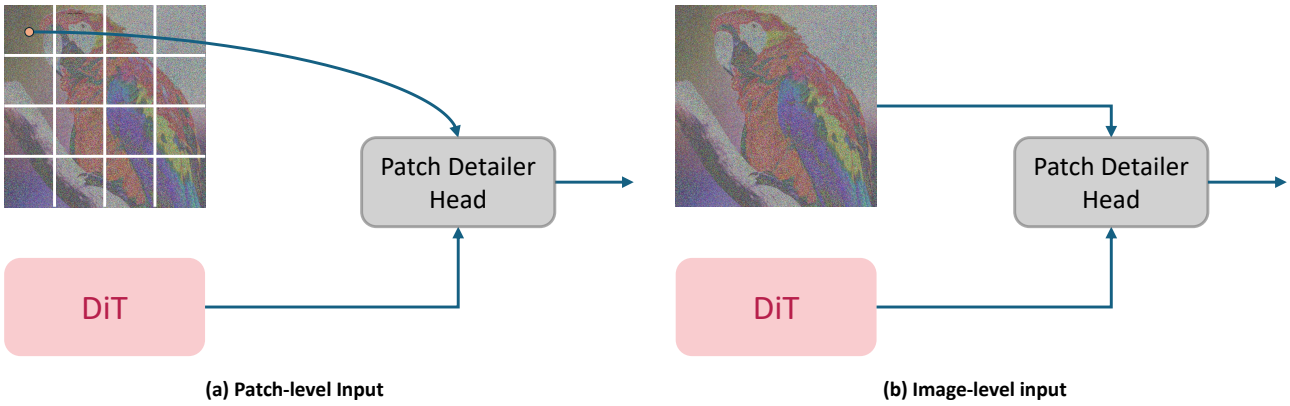

Figure 1. Different input formats of Patch Detailer Head.

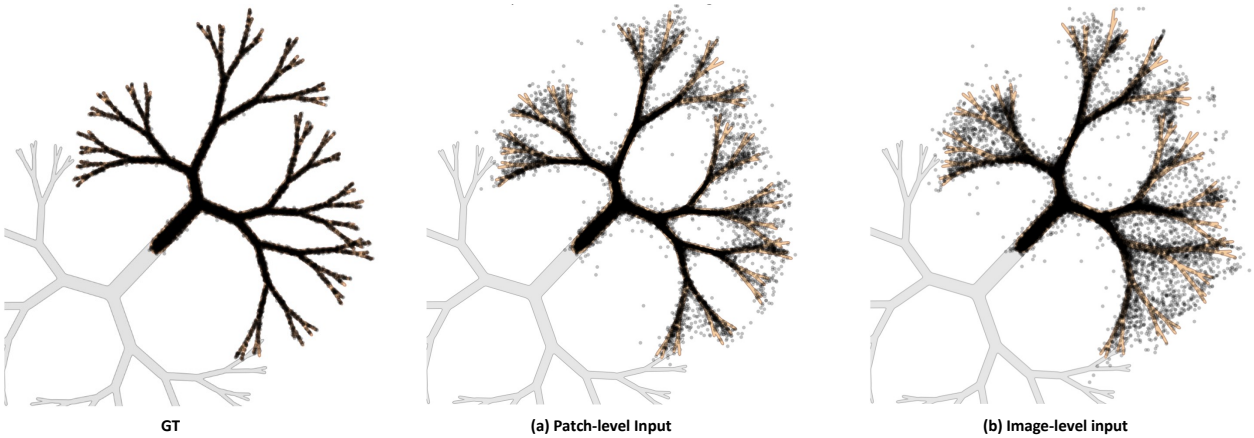

Figure 2. Toy experiment. (a) Visualization of manifold fitting with Patch-level input. The model precisely captures high-frequency branches. (b) Visualization of manifold fitting with Image-level input. The model exhibits over-smoothing and fails to resolve fine details.

E. Alternative Patch Detailer Head (PDH).

Details of the alternative PDH are in Fig. 3. The performance gap arises from the Convolutional U-Net's built-in inductive biases and hierarchical architecture, which better capture spatial detail and preserve local continuity. By contrast, alternative variants lack spatial information or are less effective at modeling local patterns, as explained in Sec. 3.4 (paper).

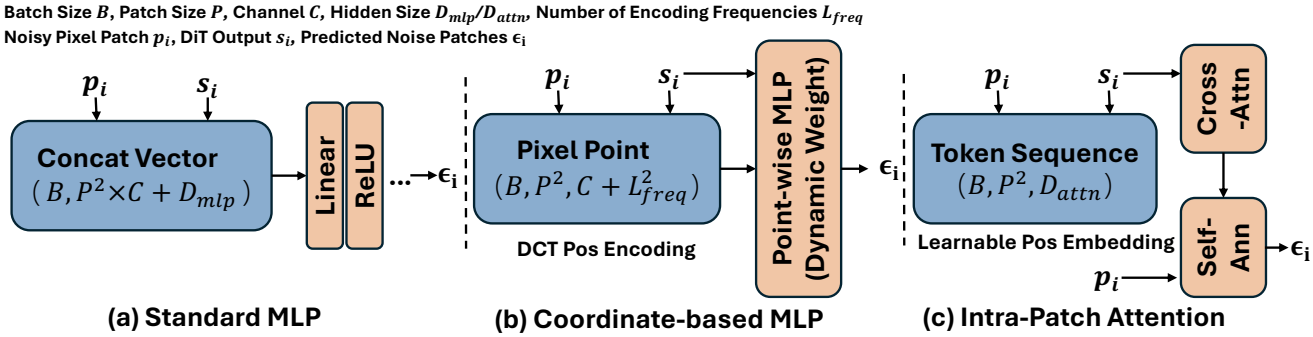

Figure 3. Details of other PDH.

F. More Visualization Results

Figure 4. 256×256 samples. Class lable = “goldfish, Carassius auratus” (1). CFG = 4.0.

Figure 5. 256×256 samples. Class lable = “junco, snowbird” (13). CFG = 4.0.

Figure 6. 256×256 samples. Class lable = “chickadee” (19). CFG = 4.0.

Figure 7. 256×256 samples. Class lable = “tree frog, tree-frog” (30). CFG = 4.0.

Figure 8. 256×256 samples. Class lable = “mud turtle” (35). CFG = 4.0.

Figure 9. 256×256 samples. Class lable = “teddy, teddy bear” (859). CFG = 4.0.

Figure 10. 256×256 samples. Class lable = “cauliflower” (938). CFG = 4.0.

Figure 11. 256×256 samples. Class lable = “potpie” (964). CFG = 4.0.

Figure 12. 256×256 samples. Class lable = “bolete” (997). CFG = 4.0.

Figure 13. 512×512 samples. Class lable = “ptarmigan” (81). CFG = 4.0.

Figure 14. 512×512 samples. Class lable=”jellyfish” (107). CFG=4.0.

Figure 15. 512×512 samples. Class lable=”Maltese dog, Maltese terrier, Maltese” (153). CFG=4.0.

Figure 16. 512×512 samples. Class lable = “lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens” (387). CFG = 4.0.

Figure 17. 512×512 samples. Class lable = “barn” (425). CFG = 4.0.

Figure 18. 512×512 samples. Class lable = “beacon, lighthouse, beacon light, pharos” (437). CFG = 4.0.

Figure 19. 512×512 samples. Class lable = “beer glass” (441). CFG = 4.0.

Figure 20. 512×512 samples. Class lable = “wool, woolen, woollen” (911). CFG = 4.0.

Figure 21. 512×512 samples. Class lable = “trifle” (927). CFG = 4.0.