FREE: Uncertainty-Aware Autoregression for Parallel Diffusion Transformers

Supplementary Material

A. Training Details

We train all drafter models on ImageNet-1K using a randomly sampled 30% subset of the training split. Each image is center-cropped to the target resolution (*e.g.* 256^2 or 512^2), normalized to [-1, 1], and encoded using a frozen SD-VAE following the standard DiT preprocessing pipeline. At each iteration, a diffusion timestep $t \in \{1, \ldots, 1000\}$ is sampled. The frozen DiT-XL/2 first processes a noisy latent x_t, and a second forward pass at $t-1$ yields the teacher supervision $(\epsilon_{ref}, f_{ref})$. The drafter receives the intermediate latent x_{t-1} without normalization, matching the teacher model's inference behavior and allowing it to exploit fine-grained information along the denoising trajectory. Classifier-free guidance also follows the DiT formulation, where 10% of labels are randomly dropped before the teacher's first forward pass. The resulting label embedding computed by DiT is then provided directly to the drafter, which does not maintain its own label-embedding module.

Although one might consider precomputing latents and intermediate DiT features to avoid online VAE encoding, this is highly impractical. Even at 256^2 resolution, the top-block hidden representation of DiT-XL/2 used during training produces a flattened feature tensor of shape [1024, 1152], amounting to $\sim$ 4.7MB per sample in FP32—over an order of magnitude larger than a raw ImageNet image (typically 100–200 KB). Moreover, because these intermediate features depend on the sampled timestep and the (possibly dropped) class label, each stored item has essentially no reuse value. For these reasons, we generate both latents and teacher features on the fly from raw images.

The drafter itself is lightweight, containing $\sim$ 32M trainable parameters compared with the 676M parameters of the DiT-XL/2 teacher. For the 256^2 configuration, we train for 120 epochs and reach convergence in approximately 56.8 hours on a $4 \times$ RTX 4090 (24GB) setup. For 512^2, training runs for 60 epochs and requires about 138 hours under the same hardware. These runtimes indicate that FREE can be trained efficiently even under modest computational budgets. Throughout training we follow the same optimization settings used for DiT. AdamW is employed with a fixed learning rate of 1×10^{-4}, and an EMA decay of 0.9999 is applied. These standard choices provide stable optimization and help maintain close alignment between the drafter and the frozen teacher model.

As mentioned in 3.2, the training loss is defined as: $\mathcal{L}_{\text{FREE}} = \mathcal{L}_{\text{noise}} + \lambda_f \mathcal{L}_{\text{feat}} + \lambda_s \mathcal{L}_{\text{smooth}}$. For the 256 resolution, we set $\lambda_f = 0.5$ and $\lambda_s = 0.005$, whereas for the 512 resolution, we kept λ_f unchanged and increased λ_s to 0.01. We plotted the partial loss curves for both settings, as shown below 7.

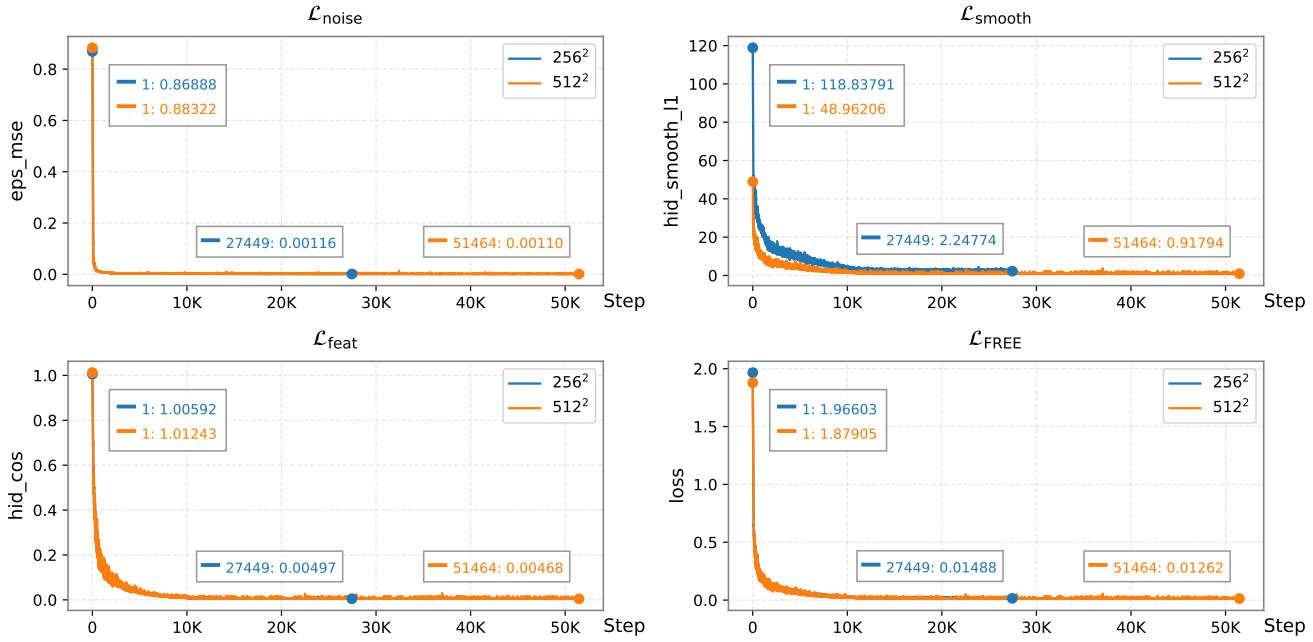

Figure 7. Training loss curves for the first 30 epochs at 256^2 resolution and the first 15 epochs at 512^2 resolution.

Meanwhile, we evaluated the model checkpoints at different training stages, and the resulting parallel efficiency is shown below 5. The results indicate that a larger $\mathcal{L}_{\text{smooth}}$ leads to slower convergence for the 256^2 drafter.

Table 5. Scaling behavior of parallel efficiency during drafter training (FREE, speculation length = 8, cfg = 1.5).

(a) 256^2 resolution

Epoch	30	60	90	119
Parallel efficiency	1.36	2.41	2.59	2.66

(b) 512^2 resolution

Epoch	15	30	45	60
Parallel efficiency	1.88	2.16	2.29	2.27

B. Theoretical Proofs for Reflection Coupling

Theorem 1 (Maximal coupling). For any two probability densities p and q on $\mathbb{R}^d$, the total variation distance satisfies

$$\|p - q\|_{TV} = 1 - \int \min\{p(x), q(x)\}\, dx = \int \max\{p(x) - q(x), 0\}\, dx. \tag{10}$$

Moreover, for any coupling (X, Y) with marginals p and q,

$$\mathbb{P}(X \neq Y) \;\geq\; \|p - q\|_{TV}, \tag{11}$$

and the equality is achievable; any coupling attaining equality is called a *maximal coupling*.

Recall Algorithm 2, where the draft distribution and the target distribution are given by the two Gaussian densities:

$$p(x) = \mathcal{N}(x; \hat{m}, \sigma^2 I_d), \qquad q(x) = \mathcal{N}(x; m, \sigma^2 I_d), \tag{12}$$

which share the same covariance matrix. In each RMC step, a draft sample $\hat{x} \sim p(x)$ is accepted ($x = \hat{x}$) with probability

$$a(\hat{x}) = \min\left(1, \frac{q(\hat{x})}{p(\hat{x})}\right). \tag{13}$$

Otherwise, upon rejection, we apply the reflection update

$$x = m + (I_d - 2ee^\top)(\hat{x} - \hat{m}), \qquad e = \frac{\hat{m} - m}{\|\hat{m} - m\|}, \tag{14}$$

where e is referred to as the mean-discrepancy direction. Let μ denote the distribution of the output x. We will show (1) unbiasedness: $\mu = q$, and (2) maximal coupling: $\mathbb{P}(x \neq \hat{x}) = \|p - q\|_{TV}$.

Lemma 1 (Geometry of the reflection). $S(z)$ is the mirror image of z across the decision hyperplane $\{x : p(x) = q(x)\}$.

Proof. The condition $p(x) = q(x)$ is equivalent to

$$\|x - \hat{m}\|^2 = \|x - m\|^2 \iff (x - \frac{\hat{m} + m}{2})^\top \cdot \frac{\hat{m} - m}{\|\hat{m} - m\|} = 0. \tag{15}$$

Hence the set $\{x : p(x) = q(x)\}$ is exactly the hyperplane

$$H = \{x : \langle x - mid, e\rangle = 0\}, \qquad mid = \tfrac{m + \hat{m}}{2}, \tag{16}$$

which passes through mid and has unit normal e. Define the reflection matrix

$$R = I_d - 2ee^\top. \tag{17}$$

It satisfies

$$R^\top R = I_d, \qquad Re = -e, \qquad Rv = v \;\text{ for all } v \perp e. \tag{18}$$

Thus R is the Householder reflection across H. The reflection map

$$S(z) = m + R(z - \hat{m}) \tag{19}$$

shifts z to the coordinate frame centered at $\hat{m}$, applies the reflection R, and then shifts the result to the frame centered at m. Therefore, $S(z)$ is exactly the mirror image of z across H.

Lemma 2 (Density swap). For all $z \in \mathbb{R}^d$, $p(z) = q(S(z))$, $q(z) = p(S(z))$.

Proof. By Lemma 1, R is an orthogonal reflection, hence

$$S(z) - m = R(z - \hat{m}) \quad \Rightarrow \quad \|S(z) - m\| = \|z - \hat{m}\|, \tag{20}$$

and similarly

$$S(z) - \hat{m} = (S(z) - m) - (\hat{m} - m) = R(z - \hat{m}) + R(\hat{m} - m) = R(z - m) \;\Rightarrow\; \|S(z) - \hat{m}\| = \|z - m\|. \tag{21}$$

Thus S is an isometry. Using the norm equalities established above,

$$\begin{aligned} q(S(z)) &\propto \exp\left(-\frac{|S(z) - m|^2}{2\sigma^2}\right) = \exp\left(-\frac{|z - \hat{m}|^2}{2\sigma^2}\right) \propto p(z), \\ p(S(z)) &\propto \exp\left(-\frac{|S(z) - \hat{m}|^2}{2\sigma^2}\right) = \exp\left(-\frac{|z - m|^2}{2\sigma^2}\right) \propto q(z). \end{aligned} \tag{22}$$

Since the normalization constants of the two Gaussians match, we obtain the exact equalities

$$p(z) = q(S(z)), \qquad q(z) = p(S(z)). \tag{23}$$

Unbiasedness. For any bounded measurable test function f,

$$\int f(x) d\mu(x) = \mathbb{E}[f(x)] = \mathbb{E}[f(\hat{x})\mathbf{1}_{\text{acc}}] + \mathbb{E}[f(S(\hat{x}))\mathbf{1}_{\text{rej}}]. \tag{24}$$

Conditioned on $\hat{x}$, the acceptance probability is $a(\hat{x})$ and the rejection probability is $1 - a(\hat{x})$, so

$$\int f(x) d\mu(x) = \int f(\hat{x}) a(\hat{x}) p(\hat{x}) d\hat{x} + \int f(S(\hat{x}))(1 - a(\hat{x})) p(\hat{x}) d\hat{x}. \tag{25}$$

From Eq. (13), the acceptance part equals $\int f(x) \min(p(x), q(x)) dx$, and the rejection part is

$$\int f(S(\hat{x}))(1 - a(\hat{x})) p(\hat{x}) d\hat{x} = \int f(S(\hat{x})) \max\big(p(\hat{x}) - q(\hat{x}), 0\big)\, d\hat{x}. \tag{26}$$

Let $x = S(\hat{x})$. Since S is an isometry, it is volume-preserving, hence $d\hat{x} = dx$. By Lemma 2, we also have $p(\hat{x}) = q(x)$ and $q(\hat{x}) = p(x)$. Thus the rejection term becomes $\int f(x) \max(q(x) - p(x), 0) dx$. Combining the acceptance and rejection contributions,

$$\int f(x) d\mu(x) = \int f(x) \left[\min(p(x), q(x)) + \max(q(x) - p(x), 0)\right] dx = \int f(x) q(x) dx. \tag{27}$$

Since this holds for all test functions f, we obtain $\mu = q$. Unbiasedness is proved.

Maximal coupling. By construction we have $\hat{x} \sim p$, and by the unbiasedness result above the output satisfies $x \sim q$. Using the acceptance rule Eq. (13) together with Theorem 1, we obtain

$$\mathbb{P}(x \neq \hat{x}) = \int (1 - a(\hat{x}))\, p(\hat{x})\, d\hat{x} = \int \max\{p(x) - q(x), 0\}\, dx = \|p - q\|_{TV}. \tag{28}$$

Hence, the RMC step realizes a maximal coupling between p and q.

C. Variance Setting

Perhaps you have noticed that the image quality metrics (*e.g.* FID) for DiT in Tab. 2 are lower than those reported in the original DiT paper. This is because we fix the reverse-process variance to the posterior value $\tilde{\beta}_t I_d$ (as in DDPM), rather than using the learned variance $\sigma_\theta(x_t, t)$ from the original DiT implementation, which provides marginal improvements in image fidelity. It is important to emphasize that this does not weaken the DiT baseline: the underlying DiT model we evaluate is identical to the original one, and the difference comes solely from a variance-setting choice during sampling.

As discussed in Sec. 3.3, fixing the variance is necessary to facilitate the formulation of reflection coupling (RMC), which requires the draft and target distributions to share the same variance at each timestep. To better clarify what "fixing the variance" means in the sampling procedure, we recall the standard one-step DDPM update expressed in terms of the noise-prediction model ϵ_θ:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t)\Big) + \sigma_t z, \quad z \sim \mathcal{N}(0, I). \tag{29}$$

where $\sigma_t = \sqrt{\tilde{\beta}_t I_d}$ is precisely the posterior variance term, ensuring that the reverse dynamics remain consistent with the assumed diffusion posterior. This formulation makes explicit that the stochasticity injected at each step is entirely governed by σ_t. Thus, "fixing the variance" simply means choosing a particular, predefined schedule for these σ_t values.

In fact, the posterior choice $\sigma_t^2 = \tilde{\beta}_t I_d$ is merely one particular fixed-variance schedule, not a fundamental requirement. From the quality standpoint, more expressive fixed-variance schedules beyond this posterior choice could be adopted. One possibility is to analyze the temporal patterns in DiT's learned variance and design a fixed variance schedule that mimics these patterns, ensuring the fixed variance adapts consistently with the learned behavior of the DiT. Alternatively, a small auxiliary network can be employed to predict the variance for each denoising step based on the current state and timestep, without interacting with the generative model. These approaches would offer flexibility and could potentially improve image fidelity while preserving the tractability of reflection coupling. Moreover, from the perspective of acceleration, the minimum posterior variance is actually the least favorable for speculative inference. Intuitively, a very small variance makes the reverse transitions too concentrated, which reduces the overlap between the draft and target kernels, leading to poorer acceptance.

Overall, the use of a fixed variance in our work is solely motivated by the analytical requirements of RMC, and should not be interpreted as a limitation of our method. More expressive and fidelity-oriented variance schedules can be incorporated seamlessly, without affecting the structure or generality of the proposed framework.

D. Stochasticity Hyperparameter

A complementary mechanism for enhancing speculative inference is to introduce a stochasticity hyperparameter ϵ into the reverse dynamics. To place this idea in context, we briefly review how diffusion models are typically defined through a forward noising process that gradually transforms data into an approximately Gaussian distribution. This forward process is described by the stochastic differential equation (SDE):

$$dX_t = f_t X_t dt + g_t dW_t, \qquad t \in [0, 1], \tag{30}$$

where f_t and g_t control signal decay and noise injection, and W_t is a standard Wiener process. Let q_t denote the marginal distribution of X_t, and satisfy

$$X_0 \sim q_0 = p_{\text{data}}, \qquad X_1 \sim q_1 \approx \mathcal{N}(0, I). \tag{31}$$

By the theory of time reversal for diffusion processes, the corresponding denoising process $Y_t := X_{1-t}$ evolves according to the reverse-time SDE:

$$dY_t = \big(-f_{1-t} Y_t + g_{1-t}^2 s_{1-t}(Y_t)\big)\, dt + g_{1-t} dB_t, \tag{32}$$

where $s_t(x) = \nabla_x \log q_t(x)$ is the Stein score (*i.e.* the gradient of the log-density) of the intermediate marginal distribution, and B_t is another independent Wiener process. If the true score were available, integrating this reverse SDE would recover samples from the data distribution. Since the reverse-time SDE cannot be solved analytically, sampling relies on a time-discretized numerical approximation of the continuous reverse dynamics. Let the interval $[0, 1]$ be partitioned into K steps:

$$0 = t_0 < t_1 < \cdots < t_K = 1, \qquad \gamma_k = t_k - t_{k-1}. \tag{33}$$

Defining the reverse-time drift

$$b_t(x) := -f_{1-t} x + g_{1-t}^2 s_{1-t}(x), \tag{34}$$

a single reverse step can be approximated using the Euler–Maruyama scheme, which induces a Gaussian transition kernel

$$q(x_{t_{k-1}}|x_{t_k}) = \mathcal{N}(m_k(x_{t_k}), \sigma_k^2 I), \tag{35}$$

with drift mean and diffusion variance given by

$$m_k(x_{t_k}) = x_{t_k} + \gamma_k b_{t_k}(x_{t_k}), \qquad \sigma_k^2 = \gamma_k g_{1-t_k}^2. \tag{36}$$

In practice, however, the true score $s_t(x)$ is not accessible. Diffusion models therefore replace it with a learned approximation $s_\theta(x,t)$ when evaluating the drift and simulating the reverse Markov chain during sampling.

A tunable stochasticity hyperparameter $\epsilon \geq 0$ can be introduced by generalizing the reverse SDE. Following prior work, we obtain a family of reverse-time processes

$$dY_t^{(\epsilon)} = b_t^{(\epsilon)}(Y_t^{(\epsilon)})dt + \epsilon g_{1-t} dW_t, \qquad b_t^{(\epsilon)}(x) := -f_{1-t}x + \frac{1+\epsilon^2}{2} g_{1-t}^2 s_{1-t}(x). \tag{37}$$

Under ideal continuous-time conditions, and assuming access to the exact score $s_t(x)$, this construction guarantees that for any $\epsilon \geq 0$,

$$Y_{1-t}^{(\epsilon)} \sim q_t, \qquad \forall t \in [0,1], \tag{38}$$

so that all members of the family share the same set of marginal distributions. The case $\epsilon = 1$ recovers the standard reverse-time SDE; $\epsilon = 0$ removes the stochastic term and yields a deterministic ordinary differential equation (ODE); $0 < \epsilon < 1$ interpolates between the ODE and SDE regimes; and $\epsilon > 1$ corresponds to a more strongly stochastic reverse process.

After introducing the stochasticity parameter ϵ, the drift mean and diffusion variance in the corresponding discrete reverse (Eq. (36)) update becomes

$$m_k(x_{t_k}) = x_{t_k} + \gamma_k b_{t_k}^{\epsilon}(x_{t_k}), \qquad \sigma_k^2 = \epsilon^2 \gamma_k g_{1-t_k}^2, \tag{39}$$

so different values of ϵ effectively modulate the level of stochasticity in the discrete reverse transitions.

In speculative sampling, the acceptance probability is determined by how well the drafter's proposals align with the transition distribution of the target model. The stochasticity hyperparameter ϵ provides a convenient means for adjusting this alignment. When ϵ is too small, the reverse kernels become overly sharp and the overlap between the drafter and target transitions decreases. Conversely, excessively large ϵ injects too much noise, pushing proposals into low-probability regions of the target model, where even small mean mismatch leads to sharply reduced acceptance. In practice, intermediate values of ϵ often strike a favorable balance, leading to more stable behavior and higher acceptance rates in speculative inference.

We can further quantify how the expected acceptance probability depends on the stochasticity parameter ϵ. Let the proposal (drafter) distribution and the target distribution be denoted by $p(x) = \mathcal{N}(x; m^p, \sigma^2 I)$ and $q(x) = \mathcal{N}(x; m^q, \sigma^2 I)$, respectively. Under RMC, the mean acceptance probability is

$$\mathbb{E}[a] = \int p(x) \min\left(1, \frac{q(x)}{p(x)}\right) dx = 2\Phi\left(\frac{-|m^p - m^q|}{2\sigma}\right), \tag{40}$$

which is entirely determined by the quantity

$$\Delta := \frac{m^p - m^q}{\sigma}. \tag{41}$$

Substituting the expressions from Eq. (39), we obtain

$$\|\Delta\|^2 = \left\| \frac{\gamma \frac{1+\epsilon^2}{2} g^2 (s^p(x) - s^q(x))}{\epsilon \sqrt{\gamma} g} \right\|^2 = \frac{1}{4}\gamma \left(\epsilon + \frac{1}{\epsilon}\right)^2 g^2 \|s^p(x) - s^q(x)\|^2, \tag{42}$$

where $s^p(x)$ and $s^q(x)$ denote the scores computed by the proposal model and the target model, respectively. Although the analytic expression in Eq. (42) suggests that $\epsilon = 1$ minimizes $\left(\epsilon + \frac{1}{\epsilon}\right)^2$ and is therefore optimal in the idealized continuous-time setting, this conclusion does not hold in practical diffusion models. In discrete time, the dominant source of deviation between the proposal and target transitions arises from drift errors, which include both score-approximation error and the numerical discretization error of Euler–Maruyama. These errors enter the drift term through the factor

$$\frac{1+\epsilon^2}{2}. \tag{43}$$

Consequently, when $\epsilon > 1$ they are amplified quadratically, causing the proposal mean to deviate substantially from that of the target. When $\epsilon < 1$, this amplification is suppressed and the drift becomes more stable, leading to a significantly smaller mean mismatch. As a result, the empirical optimum in practical speculative sampling tends to shift toward values $\epsilon < 1$.