A. Appendix

A.1. Further comparison

L	1	2	3	4	5	1^*
GFlops	26.82	27.13	28.17	29.61	33.98	35.11
Param(M)	104	208	313	417	523	135
FID ↓	31.13	16.52	15.50	13.87	**9.87**	19.74

Table 6. **Comparison of multiple model configurations over model depth L.** Unlike the baseline diffusion model ($L = 1^*$) whose computational complexity (GFlops) grows linearly with model parameters, our efficient hierarchical design only yields minimal GFlops growth with deeper models but achieves much better image quality than the baseline model with the same GFlops.

Table 6 presents a comparison of model parameters and computational complexity for models with varying depths L. In contrast to the baseline model 1^*, whose computational complexity scales linearly with the number of parameters, our efficient hierarchical design incurs only a modest computational overhead for deeper models. Under a comparable computational budget, our model with $L = 5$ demonstrates significantly better performance than 1^*.

A.2. Derivation of formulas

Derivation for $\mathcal{L}_{\text{ELBO}}$ (Eqn. 2). Let $\mathbf{x} = \mathbf{z}_1$ be the observed data and $\mathbf{z}_2, \mathbf{z}_3, \ldots, \mathbf{z}_L$ be the latent variables with $\mathbf{z}_{>l} := \{\mathbf{z}_m\}_{m=l+1}^{L}$. We assume the joint distribution of data and latent variables can be modeled as follows:

$$p_\theta(\mathbf{x}, \mathbf{z}_{>1}) = p_\theta(\mathbf{x}|\mathbf{z}_{>1}) \prod_{l=2}^{L-1} p_\theta(\mathbf{z}_l|\mathbf{z}_{>l}) p_\theta(\mathbf{z}_L), \quad (5)$$

with the corresponding posterior written as:

$$q(\mathbf{z}_{>1}|\mathbf{x}) = q(\mathbf{z}_L|\mathbf{x}) \prod_{l=2}^{L-1} q(\mathbf{z}_l|\mathbf{z}_{>l}, \mathbf{x}). \quad (6)$$

For the derivation of the ELBO, we proceed in a similar way as Pervez and Gavves [50], Takida et al. [65], Vahdat and Kautz [68] by relying on Jensen's equality:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}, \mathbf{z}_{>1}) \mathrm{d}\mathbf{z}_{>1} && (7)\\
&= \log \int q(\mathbf{z}_{>1}|\mathbf{x}) \frac{p_\theta(\mathbf{x}, \mathbf{z}_{>1})}{q(\mathbf{z}_{>1}|\mathbf{x})} \mathrm{d}\mathbf{z}_{>1} && (8)\\
&\geq \mathbb{E}_{q(\mathbf{z}_{>1}|\mathbf{x})} \log \frac{p_\theta(\mathbf{x}, \mathbf{z}_{>1})}{q(\mathbf{z}_{>1}|\mathbf{x})} && (9)\\
&\equiv \text{ELBO}. && (10)
\end{aligned}$$

By plugging in Eqn. 5 and Eqn. 6:

$$\begin{aligned}
\text{ELBO} &= \mathbb{E}_{q(\mathbf{z}_{>1}|\mathbf{x})} \Bigg[\log p_\theta(\mathbf{x}|\mathbf{z}_{>1}) && (11)\\
&\quad + \sum_{l=2}^{L-1} \log \frac{p_\theta(\mathbf{z}_l|\mathbf{z}_{>l})}{q(\mathbf{z}_l|\mathbf{z}_{>l}, \mathbf{x})} + \log \frac{p_\theta(\mathbf{z}_L)}{q(\mathbf{z}_L|\mathbf{x})} \Bigg] \\
&= \mathbb{E}_{q(\mathbf{z}_{>1}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}_{>1}) && (12)\\
&\quad - \sum_{l=2}^{L-1} \mathbb{E}_{q(\mathbf{z}_{>l}|\mathbf{x})} D_{\text{KL}} \left(q(\mathbf{z}_l|\mathbf{z}_{>l}, \mathbf{x}) | p_\theta(\mathbf{z}_l|\mathbf{z}_{>l}) \right) \\
&\quad - D_{\text{KL}} \left(q(\mathbf{z}_L|\mathbf{x}) | p_\theta(\mathbf{x}_L) \right).
\end{aligned}$$

To incorporate diffusion models to parameterize $p_\theta(\mathbf{z}_l|\mathbf{z}_{>l})$, we further decompose the KL divergence for each level l. Since we utilize a pre-trained encoder that computes each latent variable $\mathbf{z}_l$ directly from the observed data $\mathbf{x}$ (see Section 3.3), we can simplify the conditional posterior distribution by removing the dependence on $\mathbf{z}_{>l}$:

$$\begin{aligned}
&- D_{\text{KL}} \left(q(\mathbf{z}_l|\mathbf{z}_{>l}, \mathbf{x}) | p_\theta(\mathbf{z}_l|\mathbf{z}_{>l}) \right) && (13)\\
&= \int q(\mathbf{z}_l|\mathbf{x}) \Big[\log p_\theta(\mathbf{z}_l|\mathbf{z}_{>l}) - \log q(\mathbf{z}_l|\mathbf{x}) \Big] \mathrm{d}\mathbf{z}_l. && (14)
\end{aligned}$$

As the posterior q has no learnable parameters θ, then maximizing the negative KL divergence equals maximizing

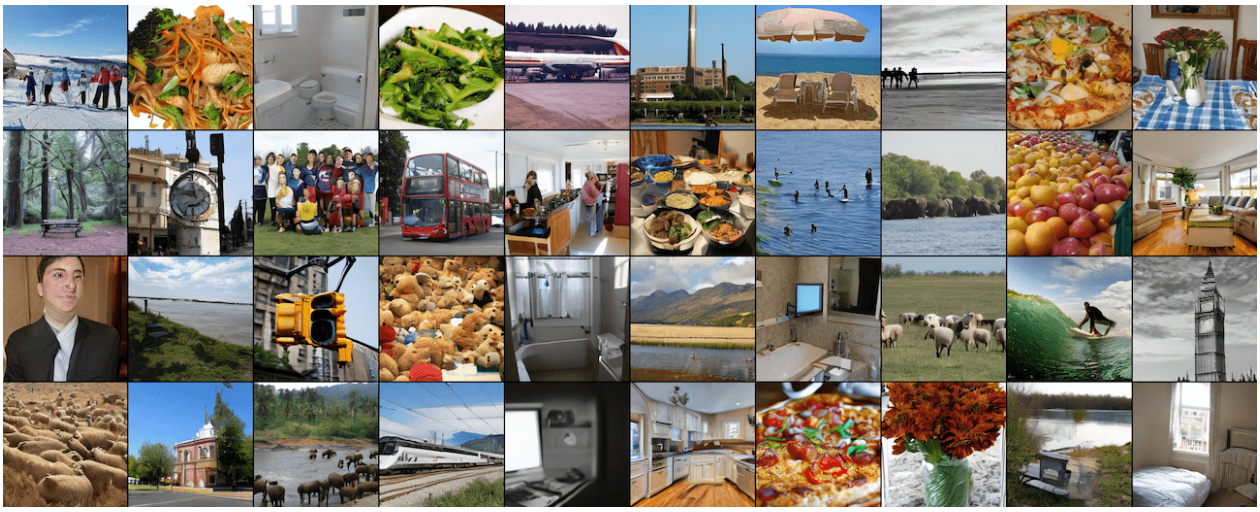
$$\int q(\mathbf{z}_l|\mathbf{x}) \log p_\theta(\mathbf{z}_l|\mathbf{z}_{>l}) \mathrm{d}\mathbf{z}_l. \quad (15)$$

Now assume that the latent variable $\mathbf{z}_l$ is modeled through a diffusion process:

$$p_\theta(\mathbf{z}_l|\mathbf{z}_{>l}) = \int p_\theta(\mathbf{z}_l^{(0:T)}|\mathbf{z}_{>l}) \mathrm{d}\mathbf{z}_1^{(1:\mathrm{T})}, \quad (16)$$

where $\mathbf{z}_l^{(0)} = \mathbf{z}_l$, and $\mathbf{z}_l^{(t)}$ denotes the noise latent variable at time step $t, \forall t \in \{0, 1, \ldots, T\}$. Then maximizing the likelihood in Eqn. 15 amounts to maximizing

$$\begin{aligned}
&\int q(\mathbf{z}_l^{(0)}|\mathbf{x}) \log p_\theta(\mathbf{z}_l^{(0)}|\mathbf{z}_{>l}) \mathrm{d}\mathbf{z}_l^{(0)} \\
&= \int \mathrm{d}\mathbf{z}_l^{(0)} q(\mathbf{z}_l^{(0)}|\mathbf{x}) && (17)\\
&\quad \log \left[\mathrm{d}\mathbf{z}_l^{(1:T)} \frac{p_\theta(\mathbf{z}_l^{(0:T)}|\mathbf{z}_{>l})}{q(\mathbf{z}_l^{(1:T)}|\mathbf{z}^{(0)}, \mathbf{x})} q(\mathbf{z}_l^{(1:T)}|\mathbf{z}^{(0)}, \mathbf{x}) \right] \\
&\geq \int \mathrm{d}\mathbf{z}_l^{(0:T)} q(\mathbf{z}_l^{(0:T)}|\mathbf{x}) && (18)\\
&\quad \log \left[p_\theta(\mathbf{z}_l^{(T)}|\mathbf{z}_{>l}) \prod_{t=1}^{T} \frac{p_\theta(\mathbf{z}_l^{(t-1)}|\mathbf{z}_t^{(t)}, \mathbf{z}_{>l})}{q(\mathbf{z}_t^{(t)}|\mathbf{z}_t^{(t-1)}, \mathbf{x})} \right] \\
&\equiv - \mathcal{L}_l. && (19)
\end{aligned}$$



(a) $L = 2$

(b) $L = 3$

Figure 7. **Visualization of text-to-image generation on COCO-2014.** We present example images generated by hierarchical diffusion models of 2 and 3 levels.

Following the derivation in Sohl-Dickstein et al.[1], the loss at each level l can be further reduced as

$$\mathcal{L}_l \leq \sum_t \int \mathrm{d}\mathbf{z}_l^{(0)} \mathbf{z}_l^{(t)} q(\mathbf{z}_l^{(0)}, \mathbf{z}_l^{(t)}) \\ D_{\mathrm{KL}}\left(q(\mathbf{z}_l^{(t-1)}|\mathbf{z}_l^{(t)}, \mathbf{z}_l, \mathbf{x}) || p_\theta(\mathbf{z}_l^{(t-1)}|\mathbf{z}_l^{(t)}, \mathbf{z}_{>l})\right). \tag{20}$$

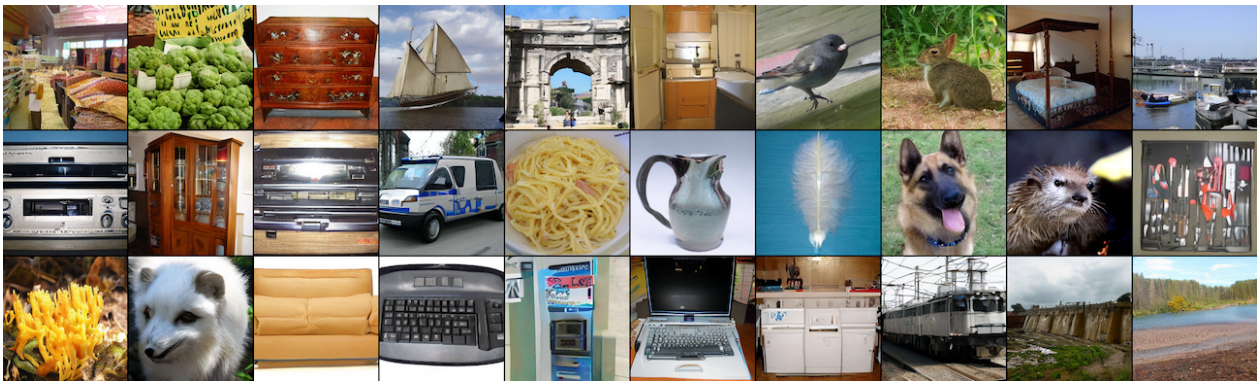
By plugging this reduced form of the loss at each level into Eqn. 12, we arrive at Eqn. 2.

A.3. Qualitative evaluation

Additional visualizations of images generated by our model at different depths are provided: Fig. 7 illustrates text-to-image generation on the COCO-2014 dataset, Fig. 8 displays conditional generation on ImageNet-1k, and Fig. 9 displays unconditional generation on ImageNet-1k.

[1] [82] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. ICML, 2015.

(a) $L = 2$



(b) $L = 3$

(c) $L = 4$

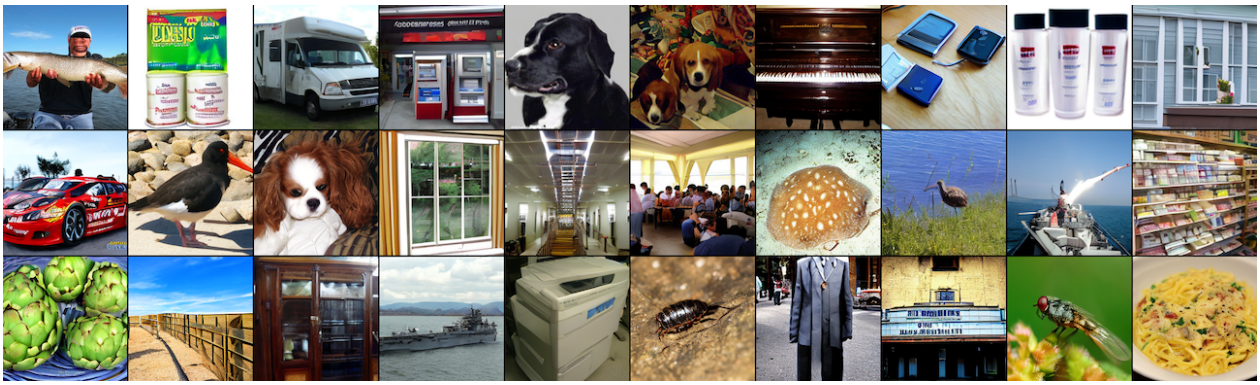
(d) $L = 5$

Figure 8. **Visualization of conditional image generation on ImageNet-1K.** We present example images generated by hierarchical diffusion models containing from 2 to 5 levels.

(a) $L = 2$

(b) $L = 3$

(c) $L = 4$



(d) $L = 5$

Figure 9. **Visualization of unconditional image generation on ImageNet-1K.** More example images generated by hierarchical diffusion models containing from 2 to 5 levels.