A. Discussion

Limitations. We have shown image conditions benefit our distillation learning. However, the distillation learning depends on the adapter architecture that takes conditions, and it is difficult to reduce the inference latency introduced by the adapter network in our current framework. As a future work, we would like to explore lightweight network architectures [14] in our distillation technique to further reduce the inference latency.

Ethics statement. The diffusion distillation technique introduce in this work holds the promise of significantly enhancing the practicality of diffusion models in everyday applications such as consumer photography and artistic creation. While we are excited about the possibilities this model offers, we are also acutely aware of the possible risks and challenges associated with its deployment. Our model's ability to generate realistic scenes could be misused for generating deceptive content. We encourage the research community and practitioners to prioritize privacy-preserving practices when using our method.

B. Proofs

B.1. Notations

We use $\hat{\mathbf{v}}_\theta(\cdot,\cdot)$ to denote a pre-trained diffusion model that learns the unconditional data distribution $\mathbf{x} \sim p_{\text{data}}$ with parameters θ. The signal prediction and the noise prediction transformed by equation 8 are denoted by $\hat{\mathbf{x}}_\theta(\cdot,\cdot)$ and $\hat{\epsilon}_\theta(\cdot,\cdot)$, and they share the same parameters θ with $\hat{\mathbf{v}}_\theta(\cdot,\cdot)$.

B.2. Self-consistency in Noise Prediction

Remark. *If a diffusion model, parameterized by $\hat{\mathbf{v}}_\theta(\mathbf{z}_t, t)$, satisfies the self-consistency property on the noise prediction $\hat{\epsilon}_\theta(\mathbf{z}_t, t) = \alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) + \sigma_t \mathbf{z}_t$, then it also satisfies the self-consistency property on the signal prediction $\hat{\mathbf{x}}_\theta(\mathbf{z}_t, t) = \alpha_t \mathbf{z}_t - \sigma_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t)$.*

Proof. The diffusion model that satisfies the self-consistency in the noise prediction implies:

$$\begin{aligned} \hat{\epsilon}_\theta(\mathbf{z}_{t'}, t') &= \hat{\epsilon}_\theta(\mathbf{z}_t, t), \\ \alpha_{t'} \hat{\mathbf{v}}_\theta(\mathbf{z}_{t'}, t') + \sigma_{t'} \mathbf{z}_{t'} &= \alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) + \sigma_t \mathbf{z}_t, \\ \hat{\mathbf{v}}_\theta(\mathbf{z}_{t'}, t') &= \frac{\alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) + \sigma_t \mathbf{z}_t - \sigma_{t'} \mathbf{z}_{t'}}{\alpha_{t'}}, \end{aligned} \tag{16}$$

Based on the above equivalence, the transformation between the signal prediction $\mathbf{x}_\theta(\mathbf{z}_{t'}, t')$ and $\mathbf{x}_\theta(\mathbf{z}_t, t)$ by using the update ruler in equation 7 and the reparameterization trick is:

$$\begin{aligned} \mathbf{x}_\theta(\mathbf{z}_{t'}, t') &= \alpha_{t'} \mathbf{z}_{t'} - \sigma_{t'} \hat{\mathbf{v}}_\theta(\mathbf{z}_{t'}, t') \\ &= \alpha_{t'} \mathbf{z}_{t'} - \sigma_{t'} \frac{\alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) + \sigma_t \mathbf{z}_t - \sigma_{t'} \mathbf{z}_{t'}}{\alpha_{t'}} && \text{// integrating equation 16} \\ &= \frac{\alpha_{t'}^2 \mathbf{z}_{t'} - \sigma_{t'} \alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) - \sigma_{t'} \sigma_t \mathbf{z}_t + \sigma_{t'}^2 \mathbf{z}_{t'}}{\alpha_{t'}} \\ &= \frac{(1 - \sigma_{t'}^2) \mathbf{z}_{t'} - \sigma_{t'} \alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) - \sigma_{t'} \sigma_t \mathbf{z}_t + \sigma_{t'}^2 \mathbf{z}_{t'}}{\alpha_{t'}} \\ &= \frac{\mathbf{z}_{t'} - \sigma_{t'} (\alpha_t \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t) + \sigma_t \mathbf{z}_t)}{\alpha_{t'}} \\ &= \frac{\mathbf{z}_{t'} - \sigma_{t'} (\hat{\epsilon}_\theta(\mathbf{z}_t, t))}{\alpha_{t'}} && \text{// transformed with equation 8} \\ &= \frac{\alpha_{t'} \mathbf{x}_\theta(\mathbf{z}_t, t) + \sigma_{t'} \hat{\epsilon}_\theta(\mathbf{z}_t, t) - \sigma_{t'} (\hat{\epsilon}_\theta(\mathbf{z}_t, t))}{\alpha_{t'}} && \text{// update ruler equation 9 of DDIM} \\ &= \mathbf{x}_\theta(\mathbf{z}_t, t). \end{aligned}$$

The derived equivalence shows that enforcing the self-consistency in the noise prediction, which is implemented by learning to minimize our distillation loss in equation 15, enforces the self-consistency in the signal prediction and can distill the pre-trained diffusion model. □

C. Difference between Consisntecy Models

Algorithm 1 Conditional Diffusion Distillation (CDD)

Input: conditional data $(\mathbf{x}, c) \sim p_{\text{data}}$, adapted diffusion model $\hat{\mathbf{w}}_\theta(\mathbf{z}_t, c, t)$, learning rate η, distance functions $d_\epsilon(\cdot,\cdot)$ and $d_{\mathbf{x}}(\cdot,\cdot)$, and EMA γ

$\boldsymbol{\theta}^- \leftarrow \boldsymbol{\theta}$ // target network initlization

repeat

Sample $(\mathbf{x}, c) \sim p_{\text{data}}$ and $t \sim [\Delta t, T]$ // empirically $\Delta t = 1$

Sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$

$s \leftarrow t - \Delta t$

Sample $\mathbf{z}_t \leftarrow \alpha_t \mathbf{x} + \sigma_t \epsilon$

\- $\hat{\mathbf{x}}_t \leftarrow \alpha_t \mathbf{z}_t - \sigma_t \Phi(\mathbf{z}_t, c, t)$

\- $\hat{\epsilon}_t \leftarrow \alpha_t \Phi(\mathbf{z}_t, c, t) + \sigma_t \mathbf{z}_t$

\+ $\hat{\mathbf{x}}_t \leftarrow \alpha_t \mathbf{z}_t - \sigma_t \hat{\mathbf{w}}_\theta(\mathbf{z}_t, c, t)$ // signal prediction in equation 8

\+ $\hat{\epsilon}_t \leftarrow \alpha_t \hat{\mathbf{w}}_\theta(\mathbf{z}_t, c, t) + \sigma_t \mathbf{z}_t$ // noise prediction in equation 8

$\hat{\mathbf{z}}_s \leftarrow \alpha_s \hat{\mathbf{x}}_t + \sigma_s \hat{\epsilon}_t$ // update rule in equation 9

\- $\hat{x}'_t \leftarrow \alpha_t \mathbf{w}_\theta(\mathbf{z}_t, c, t) + \sigma_t \mathbf{z}_t$

\- $\hat{x}'_s \leftarrow \alpha_t \mathbf{w}_{\theta^-}(\hat{\mathbf{z}}_s, c, s) + \sigma_s \hat{\mathbf{z}}_s$

\+ $\hat{\epsilon}_s \leftarrow \alpha_s \mathbf{w}_{\theta^-}(\hat{\mathbf{z}}_s, c, t) + \sigma_s \hat{\mathbf{z}}_s$ // noise prediction in equation 8

\- $\mathcal{L}(\theta, \theta^-) \leftarrow d_{\mathbf{x}}(\hat{x}'_t, \hat{x}'_s)$

\+ $\mathcal{L}(\theta, \theta^-) \leftarrow d_\epsilon(\hat{\epsilon}_t, \hat{\epsilon}_s) + d_{\mathbf{x}}(\mathbf{x}, \hat{\mathbf{x}}_t)$

$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$

$\boldsymbol{\theta}^- \leftarrow \text{stopgrad}(\gamma \boldsymbol{\theta}^- + (1-\gamma)\boldsymbol{\theta})$ // exponential moving average

until convergence

D. Implementation Details

Skip Connections. We implement the skip connections as follows, which is same as the consistency models [44] and EDMs [10] for satisfying the boundary condition but f_ϕ could be either the signal prediction or noise prediction:

$$f'_\phi(\mathbf{z}_t, t) = c_{\text{skip}}(t)\mathbf{x} + c_{\text{out}}(t) f_\phi(\mathbf{z}_t, t), \tag{17}$$

where

$$c_{\text{skip}}(t) = \frac{\sigma_{\text{data}}}{t^2 + \sigma_{\text{data}}^2}, c_{\text{out}}(t) = \frac{\sigma_{\text{data}} t}{\sqrt{t^2 + \sigma_{\text{data}}^2}}. \tag{18}$$

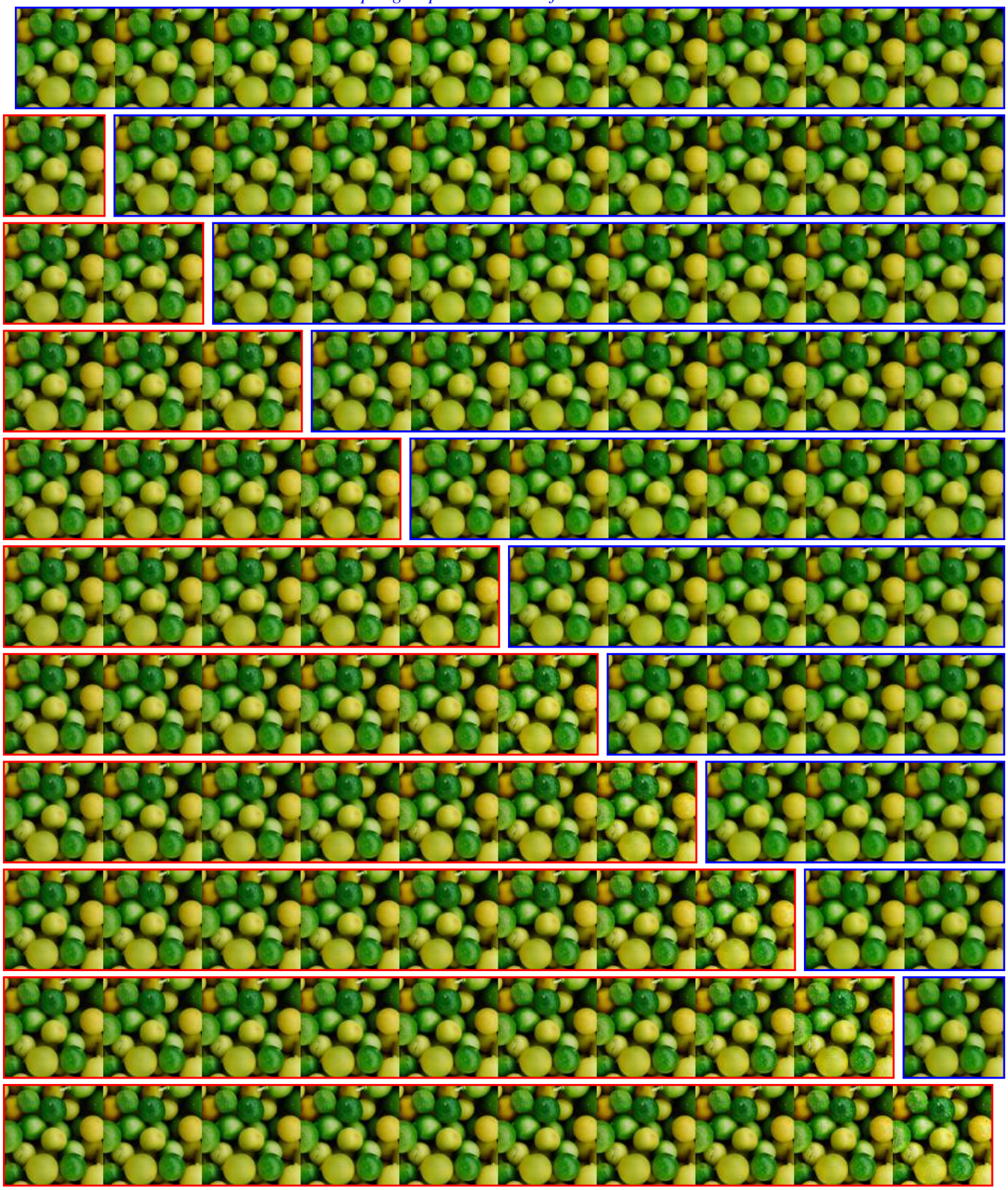
We use $\sigma_{\text{data}} = 0.5$.

E. Sampling Process Visualization

In order to provide a comprehensive understanding about the sampling process of our distilled model, as well as the difference between ours and the finetuned conditional diffusion model, here we visualize their predicted clean image $\hat{\mathbf{x}}_0$ at each sampling steps in equation 8.

As the results shown in Figure 8, we can find that our method constantly adds more details into the predicted $\hat{\mathbf{x}}_0$ when samples more steps. In contrast, such a constanly refinement is less visible in the results of the finetuned undistilled model. The different demonstrate that our method indeed can reduce the sampling time by learning to replicate the iterative refinement effects.

sampling steps $0 \leftarrow T$ *with finetuned model*



sampling steps $T \rightarrow 0$ *with our conditional distilled model*

Figure 8. Sampling process visualization of the distilled model by using our conditional diffusion distillation and the finetuned conditional diffusion model. The results belong to the same row come from the predicted $\hat{\mathbf{x}}_0$ at different time of the same sampling process, while different row denotes different sampling process that uses different the total number of the sampling time, which are increased from $T = 0$ into $T = 10$ and decreased from $T = 10$ into $T = 0$, respectively.

F. Additional results

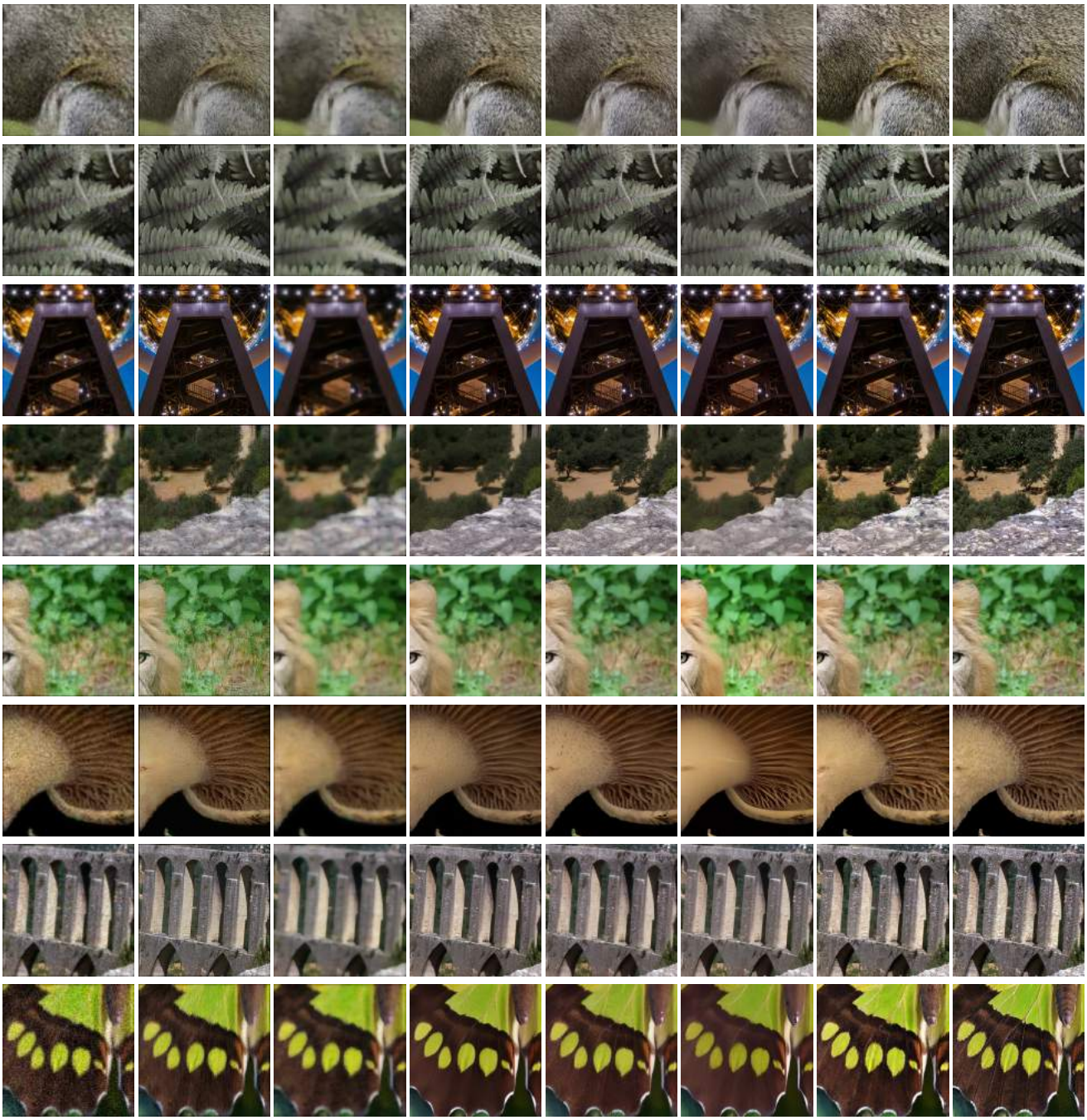


Figure 9. Visual comparisons of various diffusion-based methods on the simulated real-world super-resolution benchmark. The input of all methods is a 'Bicubic'-upsampled image.

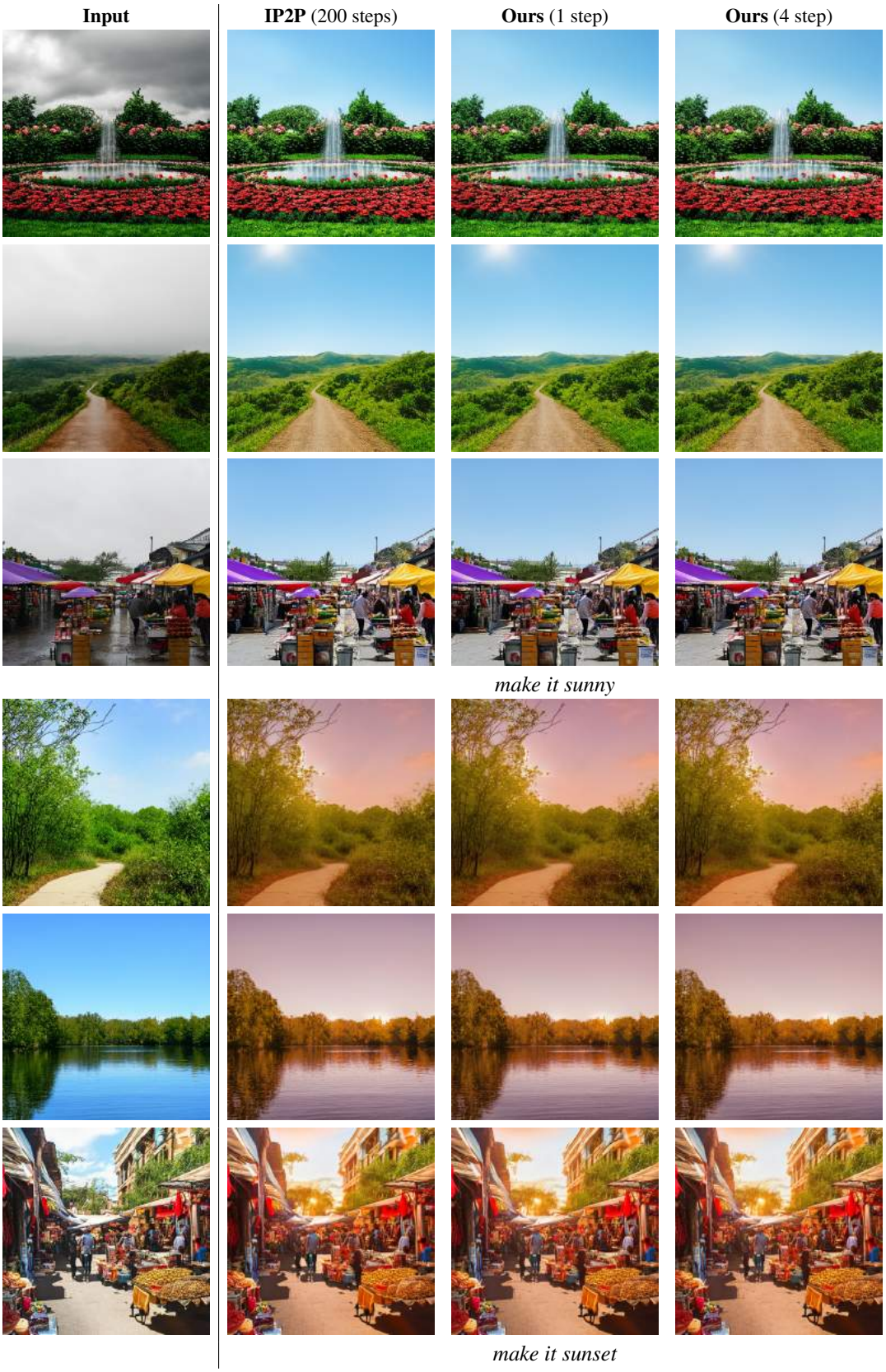


Figure 10. Visual comparisons with the IP2P model and our conditional distilled model.

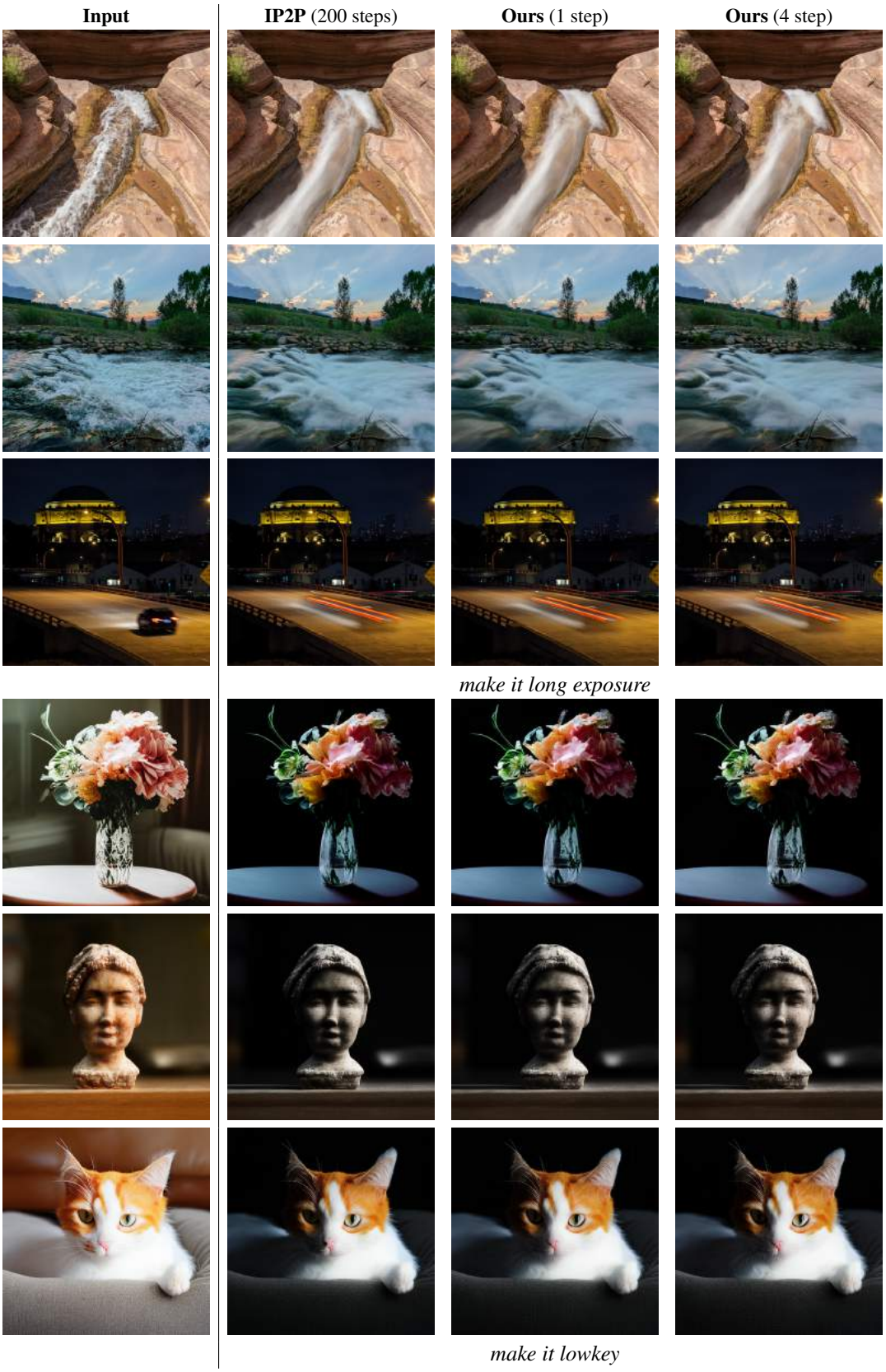


Figure 11. Visual comparisons with the IP2P model and our conditional distilled model.

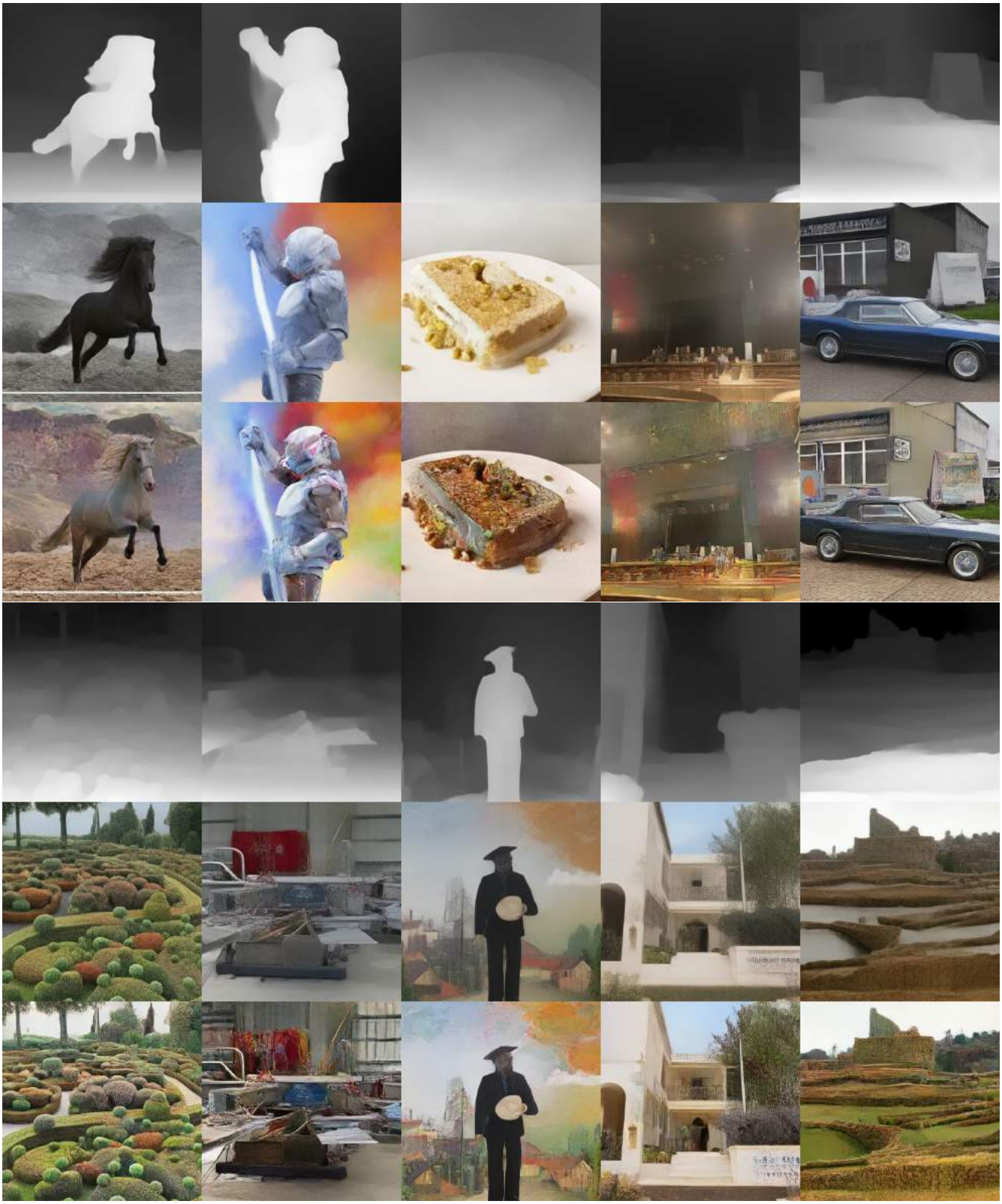


Figure 12. Visual comparisons of depth to image generation with the native ControlNet (central row of each item) and our conditional distilled model (bottom row of each item) in 4 sampling steps.

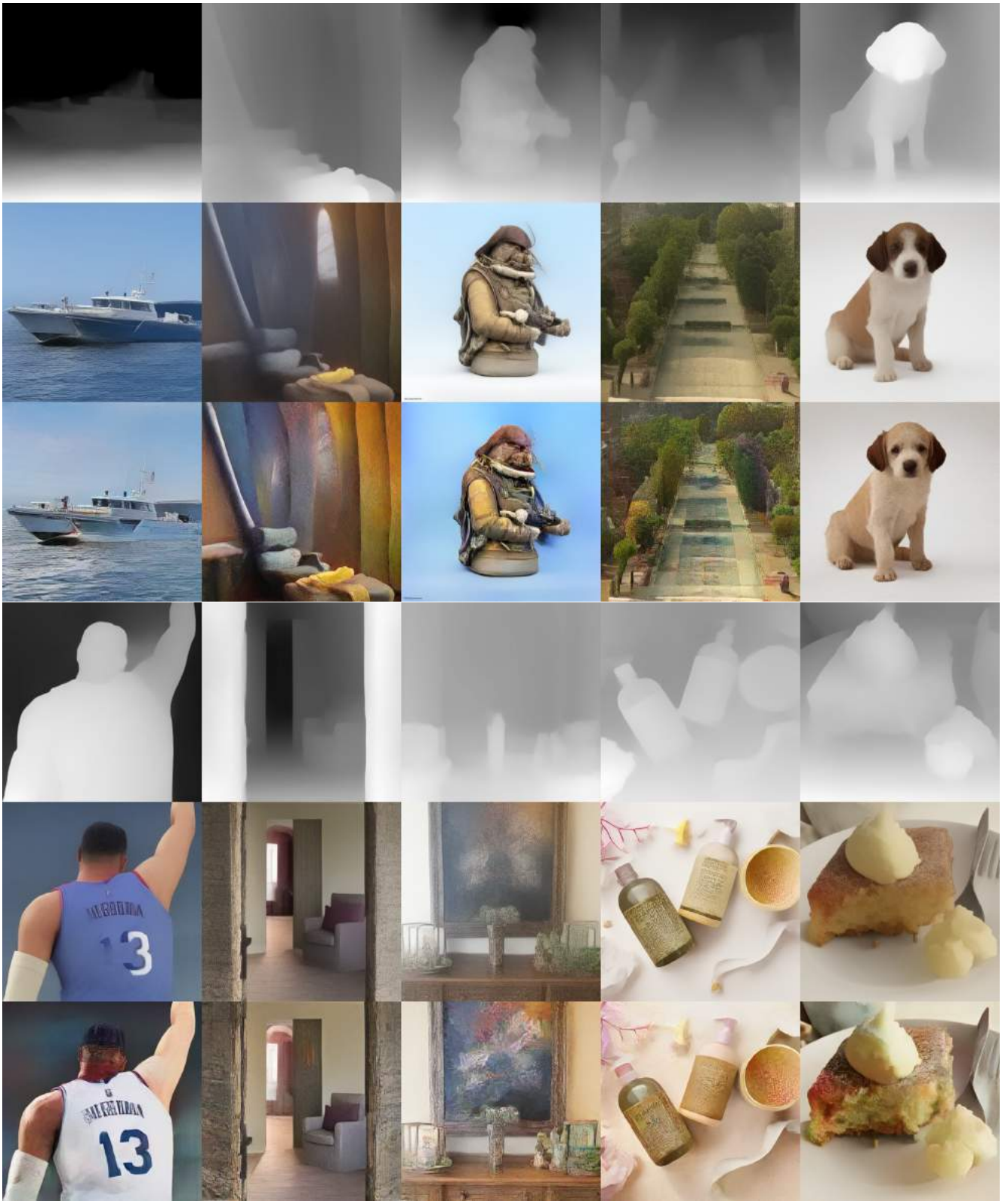


Figure 13. Visual comparisons of depth to image generation with the native ControlNet (central row of each item) and our conditional distilled model (bottom row of each item) in 4 sampling steps.