Supplementary Materials

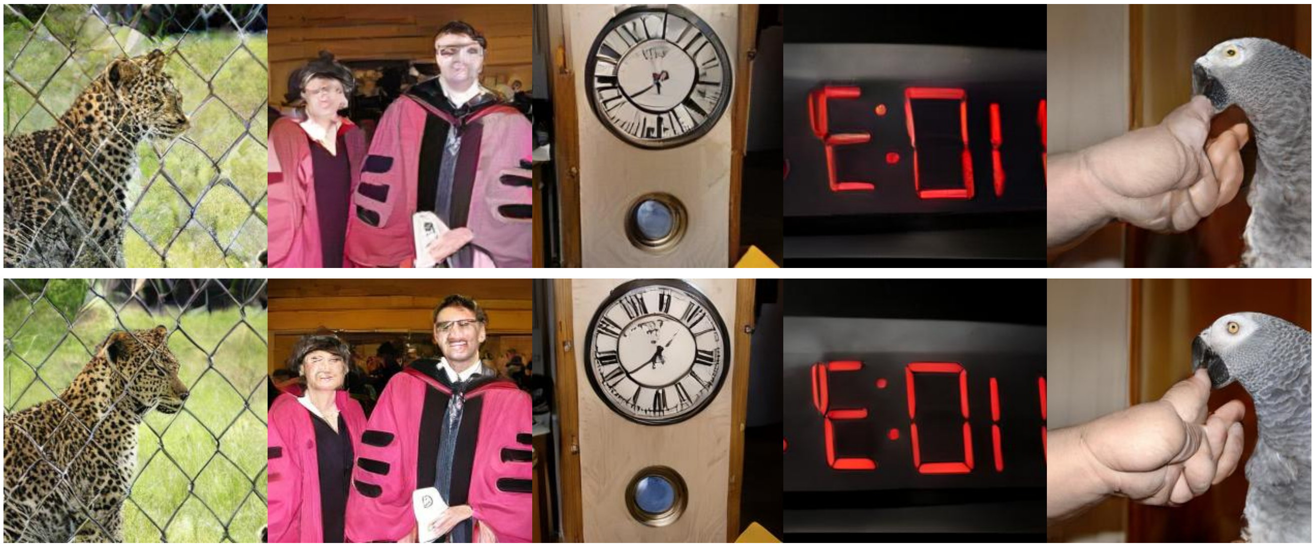

Figure 1. **Demonstration of Failure Cases.** Top: Images generated by RAR-XXL. Bottom: Images generated by our proposed DisCon-L. Although such issues are common in image synthesis, our method exhibits improved performance.

A. Implementation Details

In our experiments, the training is conducted for a default of 800 epochs. For DisCon-L, the batch size per GPU is set to 56, whereas for DisCon-B it is set to 90. Both models share a backbone with 32 transformer blocks and a width of 1024. The primary difference between the two lies in the diffusion head: DisCon-L employs 12 blocks with a width of 1536, while DisCon-B uses 3 blocks with a width of 1024.

B. Failure Cases

Figure 1 illustrates typical failure cases observed in both our method and RAR-XXL, including challenges with human faces, characters, and hands. It is important to emphasize that these issues are inherent to most image generation methods and are not unique to any single approach. Despite the prevalence of these common challenges, our approach consistently outperforms the SOTA RAR-XXL model.

C. Training Process

Figure 2 shows the training loss curve for DisCon-L. The loss stabilizes at around 100 epochs, demonstrating the reduced optimization complexity achieved by our two-stage approach. The efficient training dynamics underscore the benefits of decoupling the modeling of discrete and continuous representations, leading to more reliable and high-quality image synthesis.

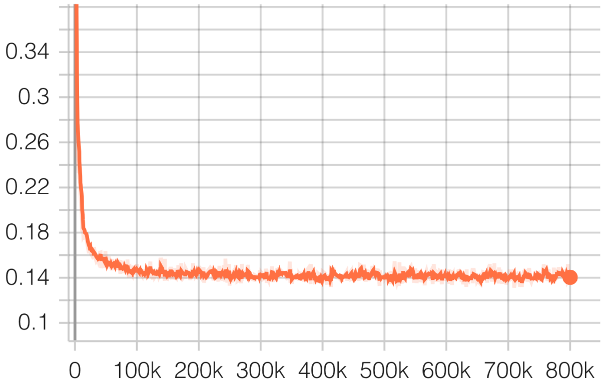

Figure 2. **Training Loss Curve of DisCon-L.** The loss converges at around 100 epochs, demonstrating the reduced optimization complexity of our approach.

D. Discrete AR Models

Our method leverages discrete tokens generated by discrete AR models. We explore performance under different RAR models in Figure 3. Additionally, we present inference results obtained by decoding these discrete tokens, which consistently demonstrate improved performance of our method.

E. Generated Results

Figure 4 presents sample images generated under different class labels, showcasing the high-fidelity synthesis and diversity achieved by our method.

F. Limitations and Future Directions.

Despite these advantages, our method still relies on a diffusion head for generating continuous tokens. Although we use a lightweight diffusion head to mitigate computational overhead, its inclusion inevitably impacts overall efficiency. Moreover, while the two-step approach simplifies the continuous modeling process, it may not be the optimal solution for all scenarios. Future research could explore alternative strategies to further balance efficiency and quality, as well as investigate novel conditioning mechanisms for continuous token generation to enhance image synthesis performance.

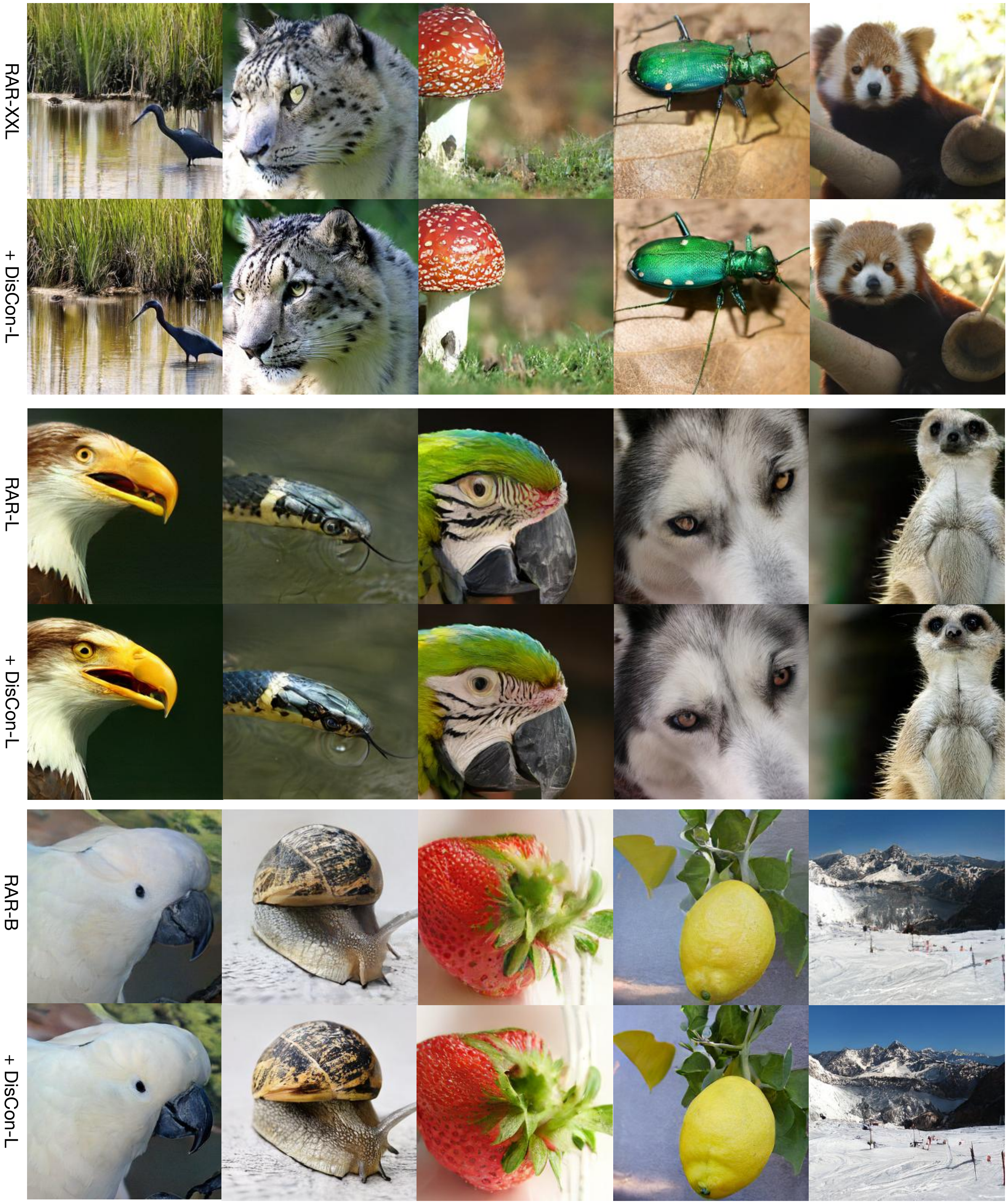

Figure 3. **Results conditioned on discrete tokens generated by different AR models.** From top to bottom: RAR-XXL, RAR-L, and RAR-B. For each model, the top row shows results generated by the respective RAR model, while the bottom row displays outputs from our DisCon method. Zoom in for better visualization to observe the significant improvements in generation quality.

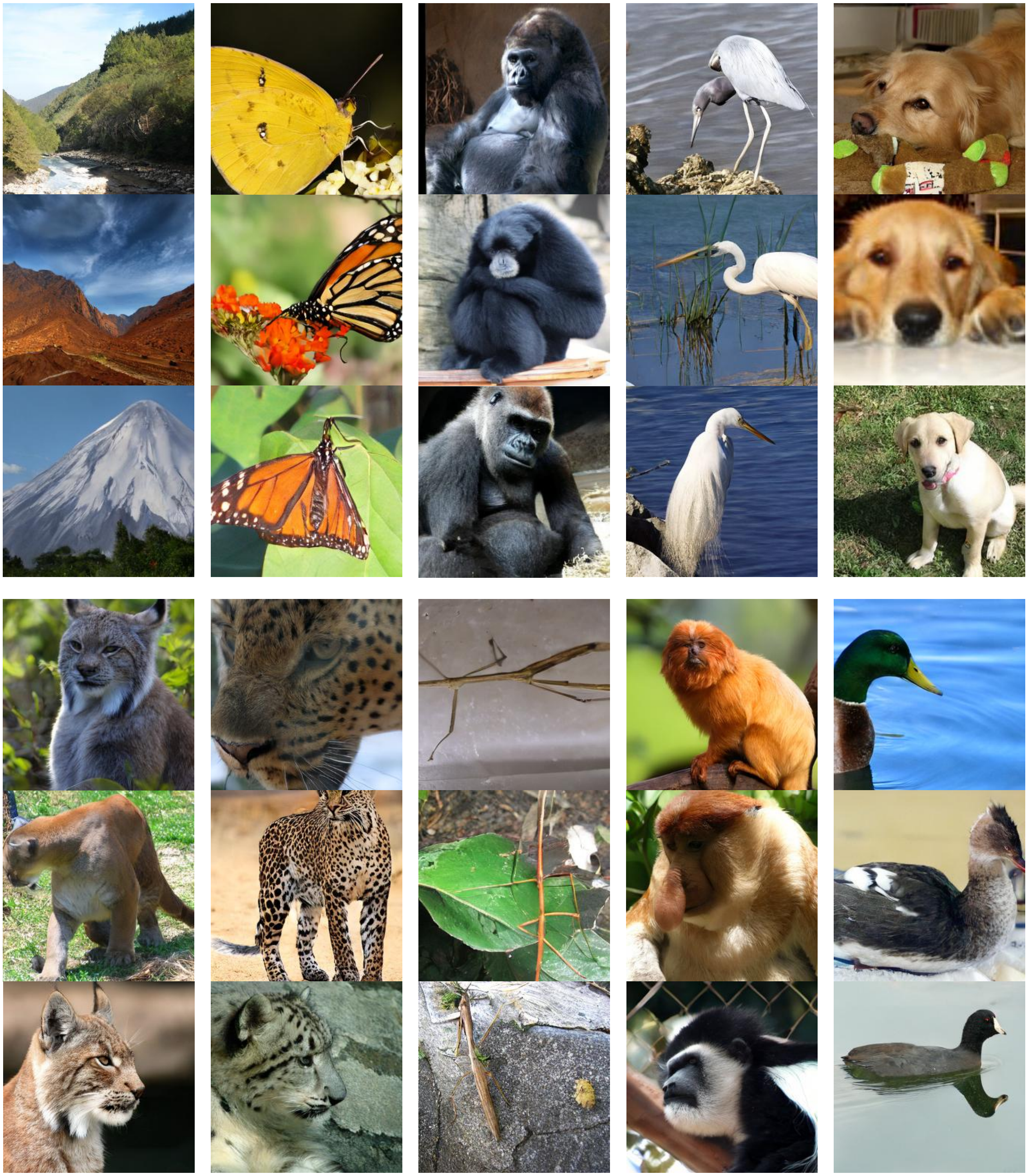

Figure 4. **Class-Conditioned Generation.** This figure showcases images generated by DisCon-L across various classes, demonstrating the high fidelity and diversity achieved by our approach.