Appendix

A. More Details

A.1. RFA-Score

Feature–Residual Alignment (RFA) score is an effective metric for assessing whether cross-scale feature-map feedback is beneficial or detrimental for distinctive representation learning at the current scale in VAR and VAR-like autoregressive models.

Score Definition. Let $\phi(\cdot)$ denote a linear projection that maps both early features F_{early} and the target residual R_{target} into a shared embedding space. Therefore, the RFA score can be defined as:

$$\begin{aligned} \text{RFA}_{(\text{early,target})} &= \cos(\phi(F_{\text{early}}),\, \phi(R_{\text{target}})) \\ &= \frac{\langle \phi(F_{\text{early}}), \phi(R_{\text{target}}) \rangle}{\|\phi(F_{\text{early}})\|\, \|\phi(R_{\text{target}})\|}. \end{aligned} \tag{6}$$

If early features contain beneficial predictive impacts, they should vary in a direction consistent with R_{target}. A positive inner product $\langle \phi(F_{\text{early}}), \phi(R_{\text{target}}) \rangle > 0$ indicates that early features enhance the consistent learning direction, while a negative value illustrates contradictory or detrimental feedback.

Let $\mathcal{L}$ be the scale-wise residual loss. From the back-propagation $\nabla_F \mathcal{L} \propto -\phi(R_{\text{target}})$, we can learn that the gradient direction is aligned with R_{target}. If early features $\phi(F_{\text{early}})$ have high cosine similarity with $\phi(R_{\text{target}})$, they contribute gradients for residual minimization, which is a form of positive consistency. Negative similarity means gradient conflict [78].

A.2. Global History Compensation Mechanism

In Table 4, we evaluate a global history variant to compare with our history compensation mechanism. This variant is implemented as a cross-attention-based memory update block, where the historical state from the previous scale and the current-scale feature map are jointly fed into a cross-attention module to update a new global memory state. This design enables each scale to incorporate aggregated historical information from all preceding scales.

B. More Analysis

Additional Analysis at Higher Resolutions. We provide additional evaluation at 512×512 resolution to further analyze the effectiveness of Markov-VAR. The samples in Figure 11 are generated by our 0.88B Markov-VAR-$d36$ model trained for 150 epochs. As shown in Table 6, compared to VAR-$d36$ (2.3B) trained for 250 epochs, our model achieves comparable performance (FID 3.07 vs. 2.63, IS 298.6 vs. 303.2) with significantly fewer parameters and reduced training cost. Under identical settings (last two rows of Table 6), with the same model scale ($d16$), training epochs (40), and resolution (512×512), Markov-VAR further improves generation quality (FID 9.81 vs. 10.18, IS 197.8 vs. 172.1) while reducing inference time (1.94s vs. 2.93s), demonstrating consistent gains in both efficiency and performance. We also study the impact of sliding window size at higher resolutions. The optimal size remains unchanged (size=3 achieves FID 9.81), while size=2 and size=4 yield similar results (FID 9.85 and 9.83), indicating reduced sensitivity to this hyperparameter at higher resolutions. This suggests that the Markovian formulation is robust and does not require additional tuning when scaling to higher resolutions.

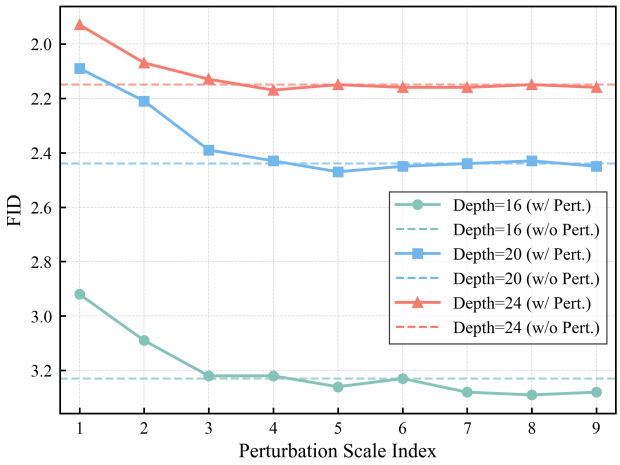


Figure 8. FID Performance comparison of Markov-VAR with different depths under perturbations injected at various scales.

Table 6. Evaluation on ImageNet 512×512 generation

Type	Model	Params.	Epochs	FID↓	IS↑	Time(s)
Diffusion	DiT-XL/2	675M	–	3.04	240.8	–
AR	VQGAN	1.4B	–	26.52	66.8	–
VAR	VAR-$d36$	2.3B	250	2.63	303.2	–
	Our Variant	0.88B	150	3.07	298.6	–
	VAR-$d16$	0.31B	40	10.18	172.1	2.93
	Markov-VAR-$d16$	0.32B	40	9.81	197.8	1.94

An Interesting Observation. While investigating the challenge of continuous error accumulation in the introduction, we observed an interesting phenomenon: as illustrated in Figure 8, when perturbations are injected at early scales, VAR exhibits a performance drop, whereas Markov-VAR achieves a performance improvement. In Figure 9, we carefully select several generated samples with better semantics quality to demonstrate this phenomenon. We speculate that the Markovian modeling process may introduce certain

self-correction opportunities, which could help the model to refine its representations. This interesting phenomenon deserves further investigation in the future.

Analysis on Network Reshaping. Since Markov-VAR removes the full-context dependency as in VAR and other VAR-like models, its intra-scale computation and attention complexity are significantly reduced. To further improve parameter efficiency, we reset the network structure as: the width $w = 32d$, the number of attention head $h = 2/3d$ and the dropout rate $dr = 0.1 \cdot d/24$. Based on this configuration strategy, we construct two variants: a small-scale model with depth $= 24$ (Makrov-VAR-S) and a medium-scale model with depth $= 30$ (Makrov-VAR-M). The evaluation results shown in Table 7, for our non-full-context dependency modeling paradigm, increasing network depth is more beneficial than increasing width. Interestingly, deeper models achieve better generation quality even with fewer parameters. This trend indicating that depth-oriented scaling is inherently more effective for Markov-VAR's non–full-context design.

Table 7. Ablation study on the of Markov-VAR performance under different network structural configurations.

Model	Param	FID↓	IS↑	Precision↑	Recall↑
Markov-VAR-S	267.4M	3.17	263.5	0.84	0.51
Markov-VAR-M	516.1M	2.58	275.8	0.83	0.56
Markov-VAR-*d16*	329.0M	3.23	256.2	0.84	0.52
Markov-VAR-*d20*	623.2M	2.44	286.1	0.83	0.56

C. Inference Algorithm

Algorithm 2 shows the inference process. It illustrates how Markov-VAR aggregates historical information, forms the Markovian state, and predicts residuals for each scale.

Algorithm 2: Inference Process of Markov-VAR

Input: Condition $\langle \text{sos} \rangle$, window size N, scales $(S_t \times S_t)_{t=1}^{T}$, temperature τ
Output: Generated image I

$R_{t0} \leftarrow \langle \text{sos} \rangle$; $M_{t0} \leftarrow \langle \text{sos} \rangle$;
$\mathcal{W} \leftarrow$ *Queue*() ;
$\hat{R} \leftarrow [\,]$;
for $t = 1$ **to** T **do**
  if $\mathcal{W}.size() == N$ **then**
    $Queue\_Pop(\mathcal{W})$;
  end
  $T_t = Concat(\mathcal{W})$;
  if T_t *is not empty* **then**
    $h_{tt} = Attn(q_\theta, T_t, T_t)$;
  else
    $h_{tt} = \mathbf{0}$;
  end
  $E_{t-1} = Embed(Up(R_{tt-1}, S_t))$;
  $H_{t-1} = Broadcast(h_{tt})$;
  $M_{tt-1} = Concat(E_{t-1}, H_{t-1})$;
  $logits_t = Model(M_{tt-1})$;
  $R_{tt} \sim Softmax(logits_t/\tau)$;
  $\hat{R}.Append(R_{tt})$;
  $X_t = Tokenize(E_{t-1})$;
  $Queue\_Push(\mathcal{W}, X_t)$;
end
$I = Decode(\hat{R})$;
return I

D. More Visualization

More Main Generation Visualization. Figure 10 and Figure 11 show more generated images with 256×256 or 512×512 via Markov-VAR-d24. Given the scale and quality of the training dataset ImageNet, the performance of Markov-VAR can be regarded as competitive enough. With larger-scale and higher-quality training datasets, Markov-VAR is expected to become more promising.

Visualization of Various Classifier-Free Guidance. To further investigate the impact of Classifier-Free Guidance (cfg) in Markov-VAR during inference, we visualize the impact of Classifier-Free Guidance on generated images. As shown in Figure 12, increasing cfg generally improves sharpness, which is aligned to other VAR-like models [38, 56, 58]. However, what is different is that the semantic quality of the generated images shows strong robustness to the cfg setting: its FID varies more smoothly across a wide range of guidance scales, and the overall performance remains consistently favorable under different cfg values. These stable results and phenomenon indicate that Markov-VAR is more robust to sampling noise and reduces the need for extensive hyper-parameter tuning when switching between different guidance strengths, thereby lowering the computational and resource cost in practice.

Visualization across Various Model Sizes. To further illustrate the performance of Markov-VAR with different model sizes, we provide visualizations generated by models of depth 16, 20, and 24, to assess how increasing the model depth influences semantic fidelity, texture richness, and structural coherence. As shown in Figure 13, across model sizes, Markov-VAR consistently generates images with stable semantic structures and visually pleasing textures. As the depth increases, we observe a clear improvement in fine-grained details, suggesting that larger models benefit more from the coarse-to-fine residual modeling pro-

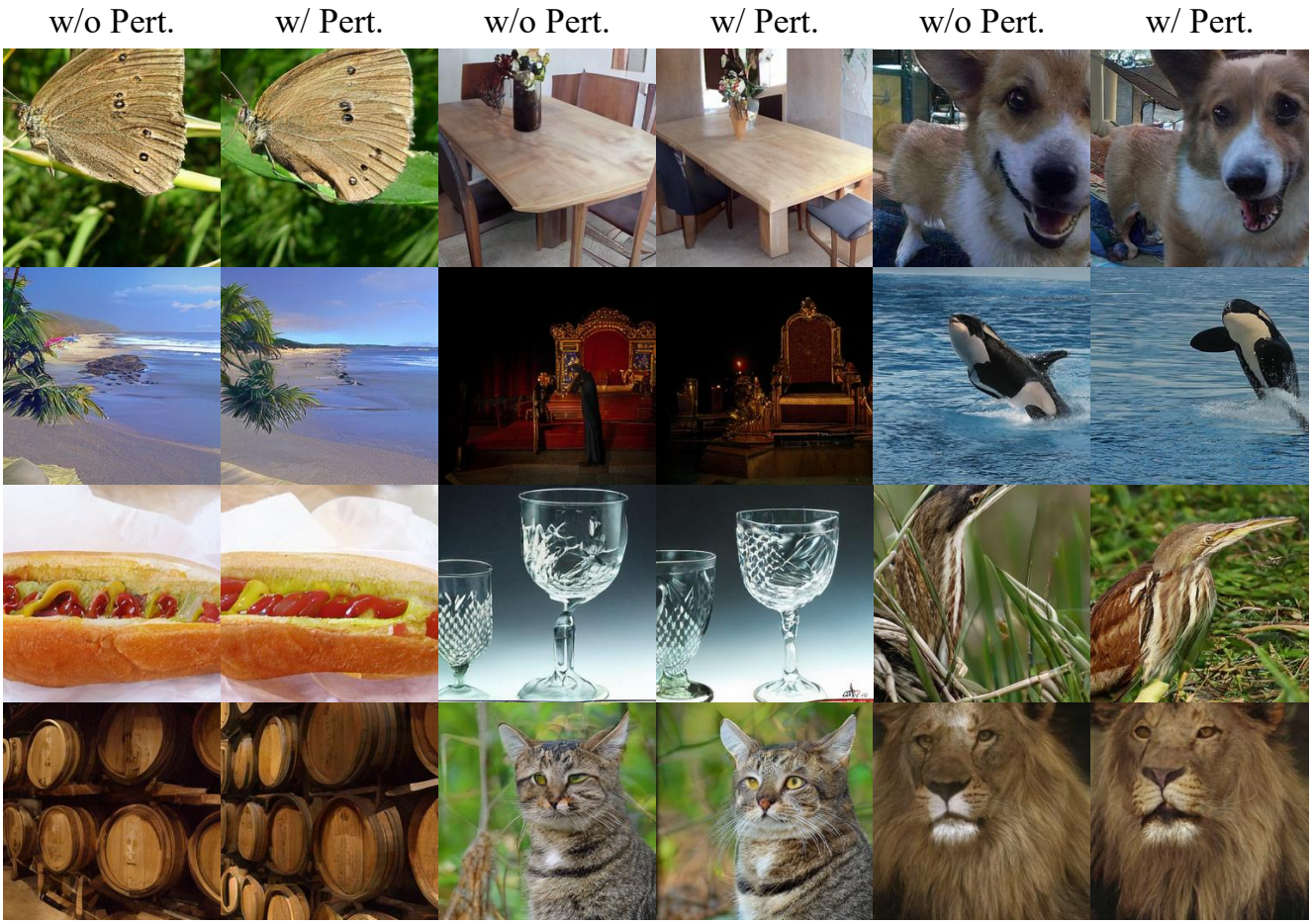


Figure 9. Visual comparison of selected perturbed generated samples in Markov-VAR with better semantic quality.

cess. This also provides intuitive evidence that Markov-VAR preliminarily follows the scaling law on model size.

More Visual Comparison between VAR and Markov-VAR. Figure 14 shows more comparisons on the generated images between VAR and Markov-VAR. We provide more image samples of the same class.

Zero-shot Task Generalization. Following [27, 58], we also test Markov-VAR on zero-shot task generalization, including image class-conditional editing, unconditional editing and in-painting. Markov-VAR is forced to generate tokens only in the bounding box conditional on related requirements. Figure 15 shows that Markov-VAR can produce plausible content that fuses well into surrounding contexts, achieving favorable results in these downstream tasks and verifying the generality of Markov-VAR.

Visualization of Generation Process To better illustrate the generation process in Markov-VAR, we visualize the Markovian scale prediction process at 256×256, as shown in Figure 16, the model predicts residuals scale by scale and incrementally refines the image. This visualization provides an intuitive understanding of how Markov-VAR constructs semantic structure and fills in fine-grained details across multiple scales. From the visualization, we observe that the earliest scales capture only very coarse global structures, such as rough color distributions and object silhouettes. As the scale increases, the model gradually refines the semantics. Later scales enrich high-frequency textures and refine local details such as fur patterns, grass textures, and cloud shapes. Notably, the refinement is both smooth and consistent, demonstrating that each Markovian scale effectively preserves essential historical information without relying on full-context dependency. This provides compelling evidence that Markov-VAR achieves stable and progressive construction throughout the generation hierarchy.

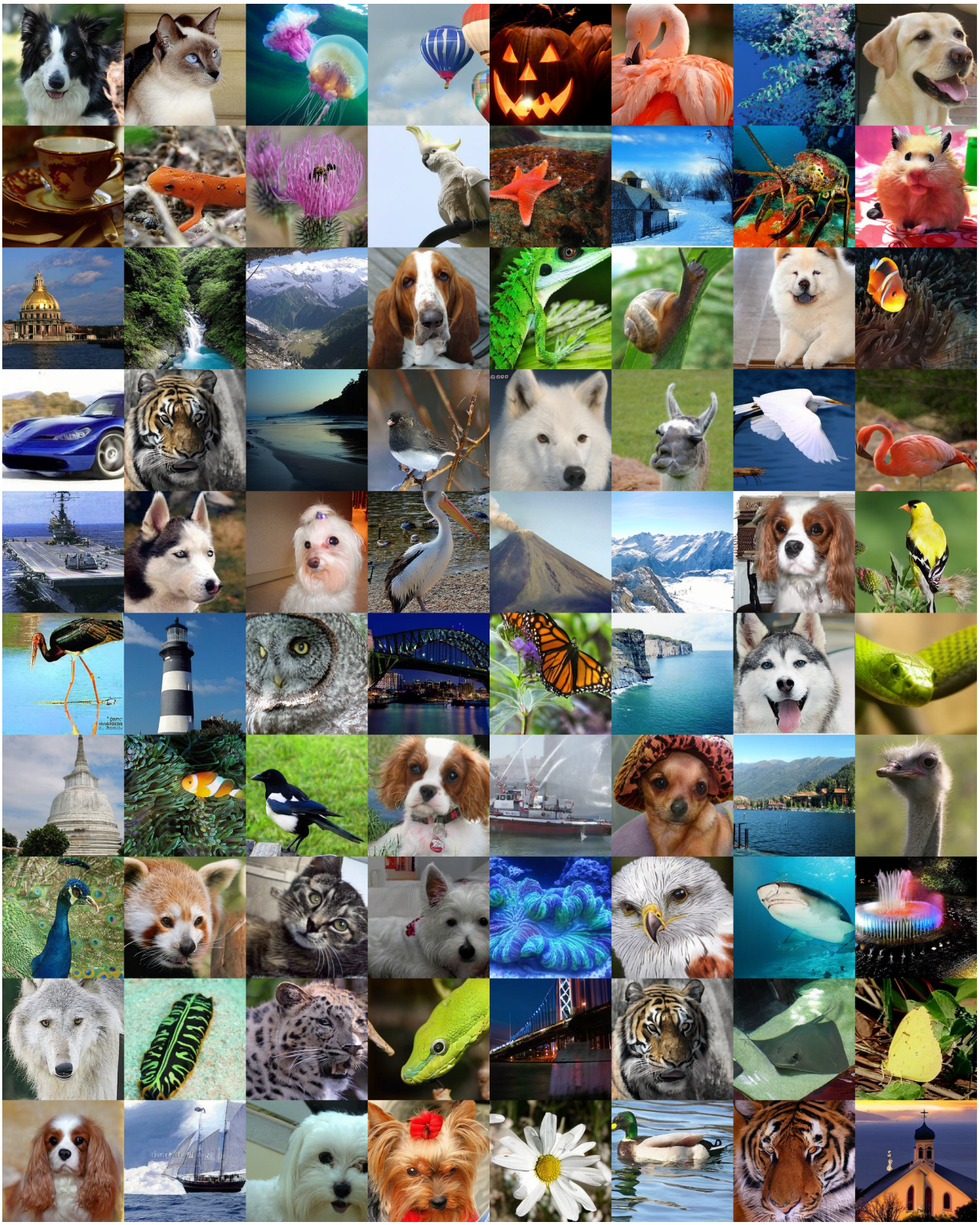


Figure 10. Visualization of 256×256 images generated by Markov-VAR.

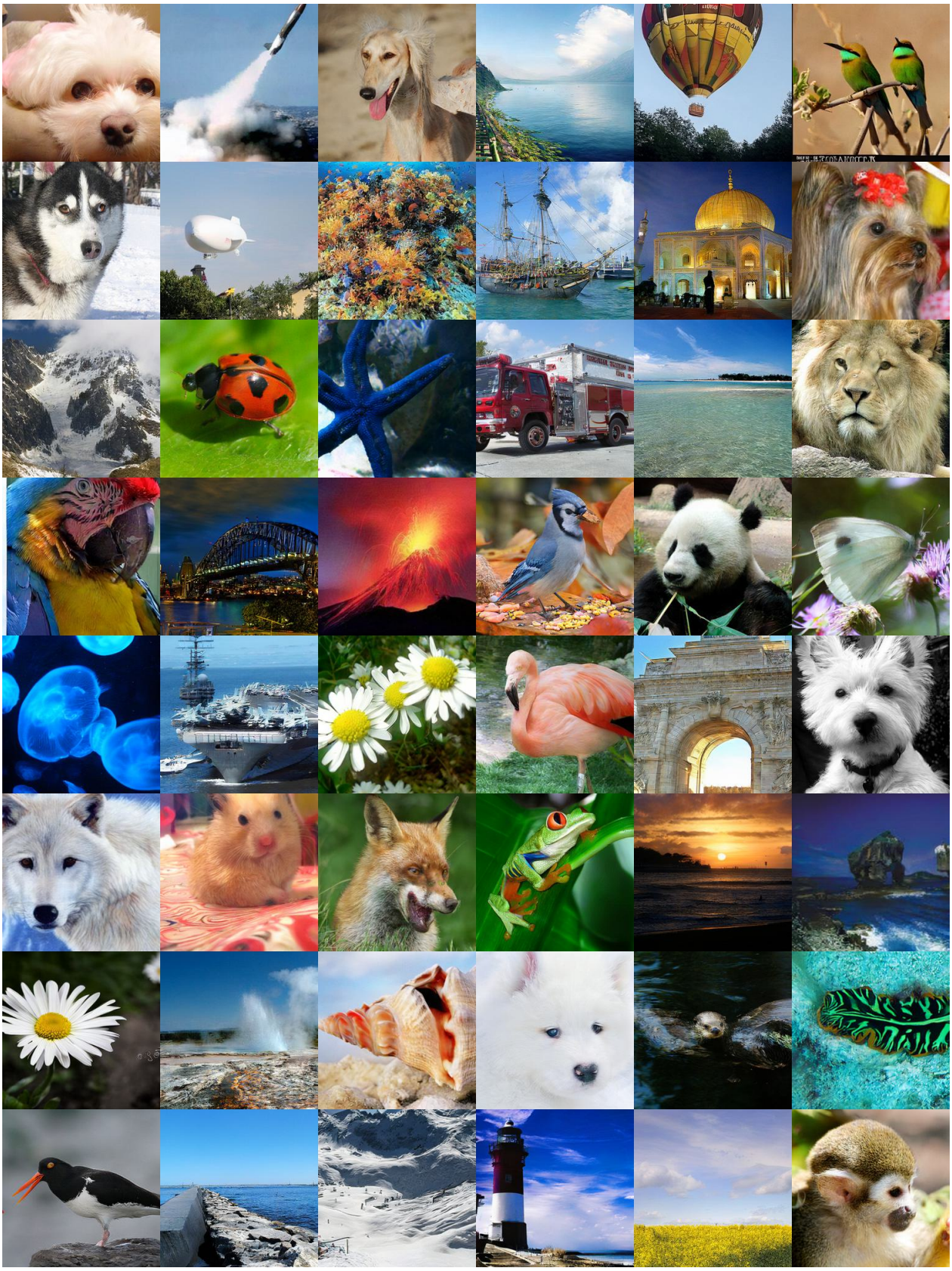


Figure 11. Visualization of 512×512 images generated by Markov-VAR.

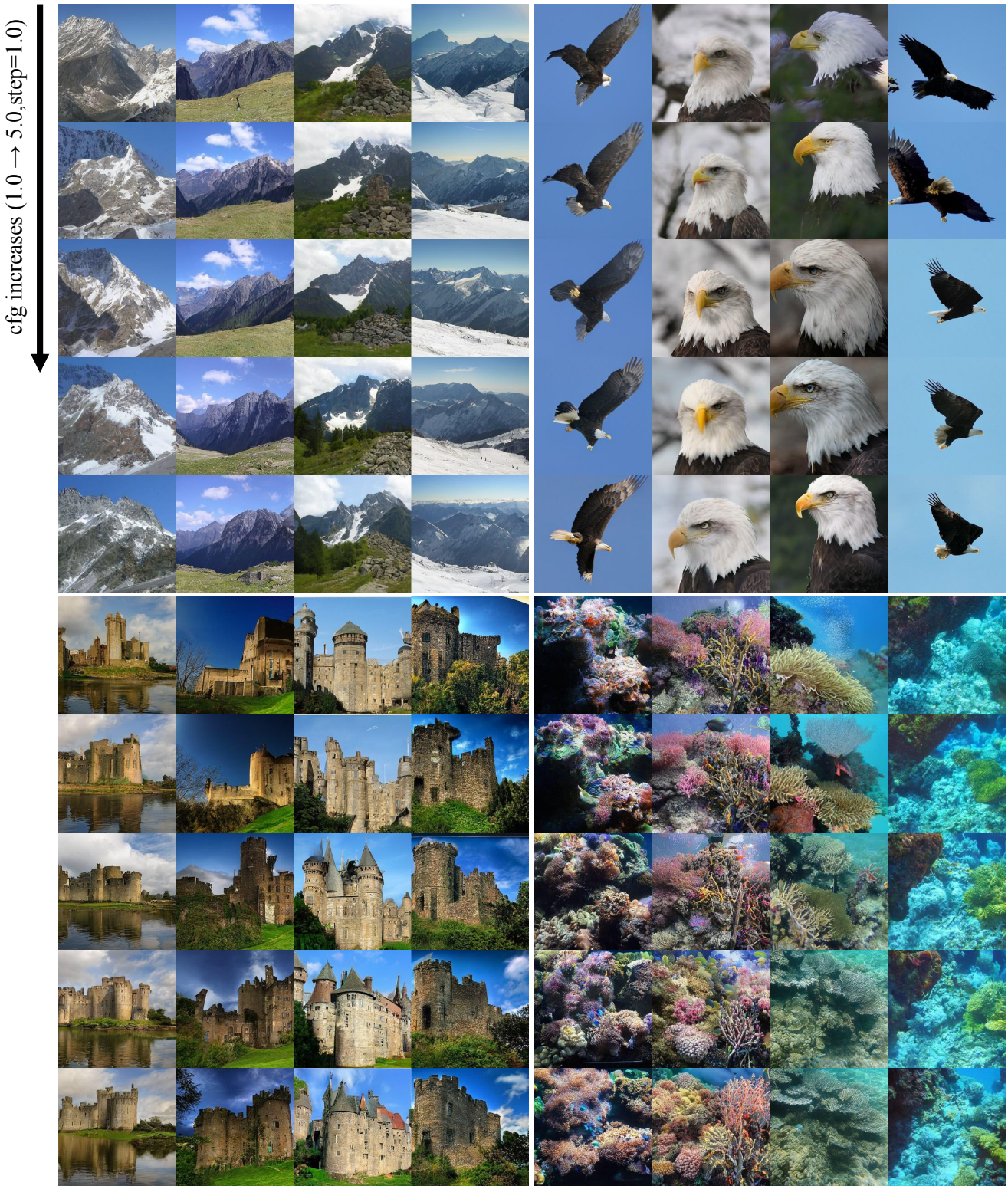
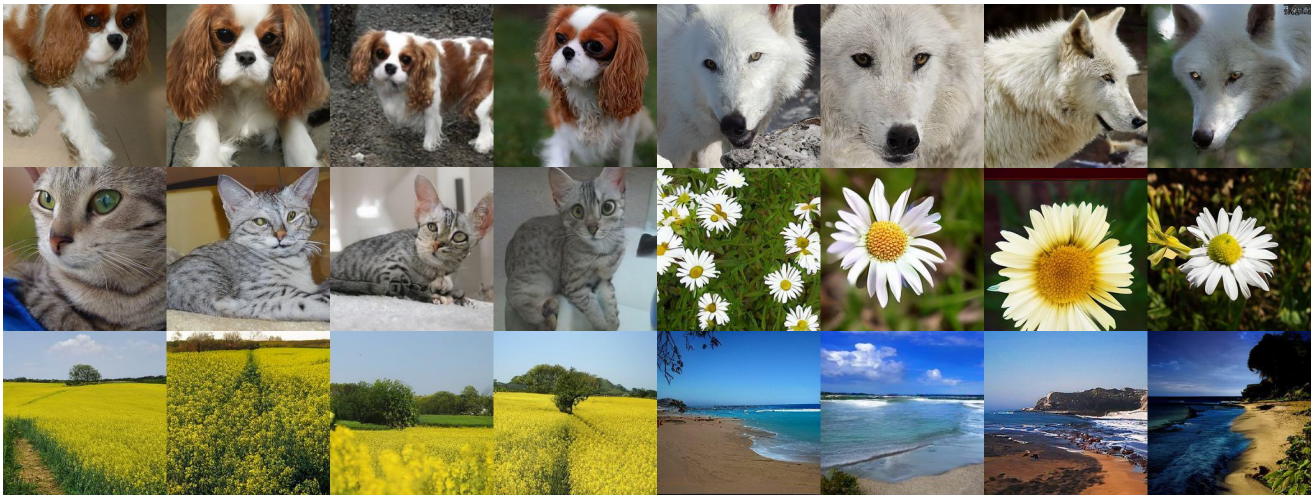
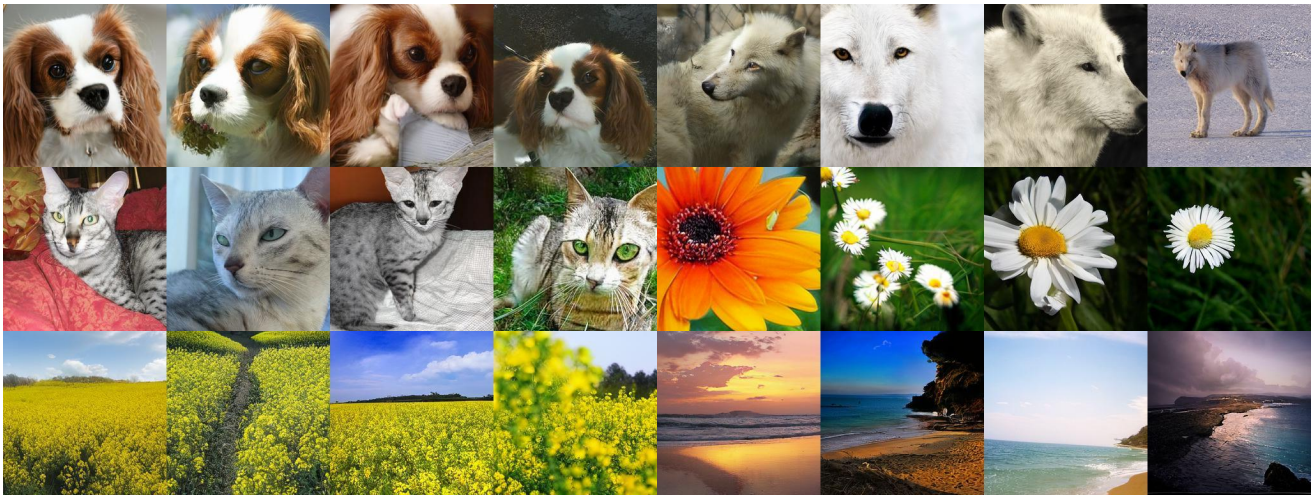


Figure 12. Images generated by Markov-VAR with increasing classifier-free guidance settings.

Markov-VAR-d16



Markov-VAR-d20



Markov-VAR-d24

Figure 13. Visualization of 256×256 images generated by Markov-VAR with different depths.

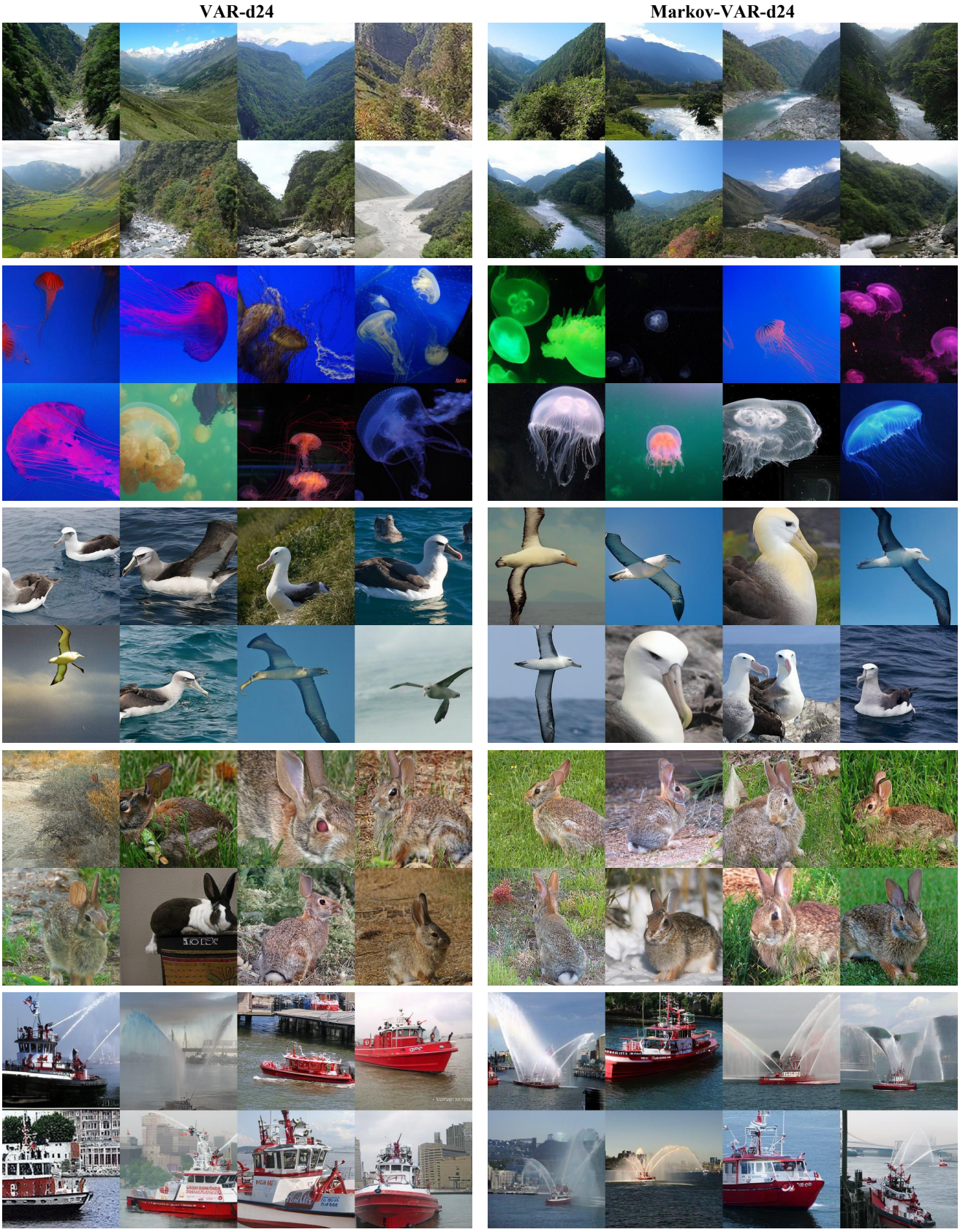


Figure 14. Comparison of image samples generated by Markov-VAR and VAR at 256×256 resolution.

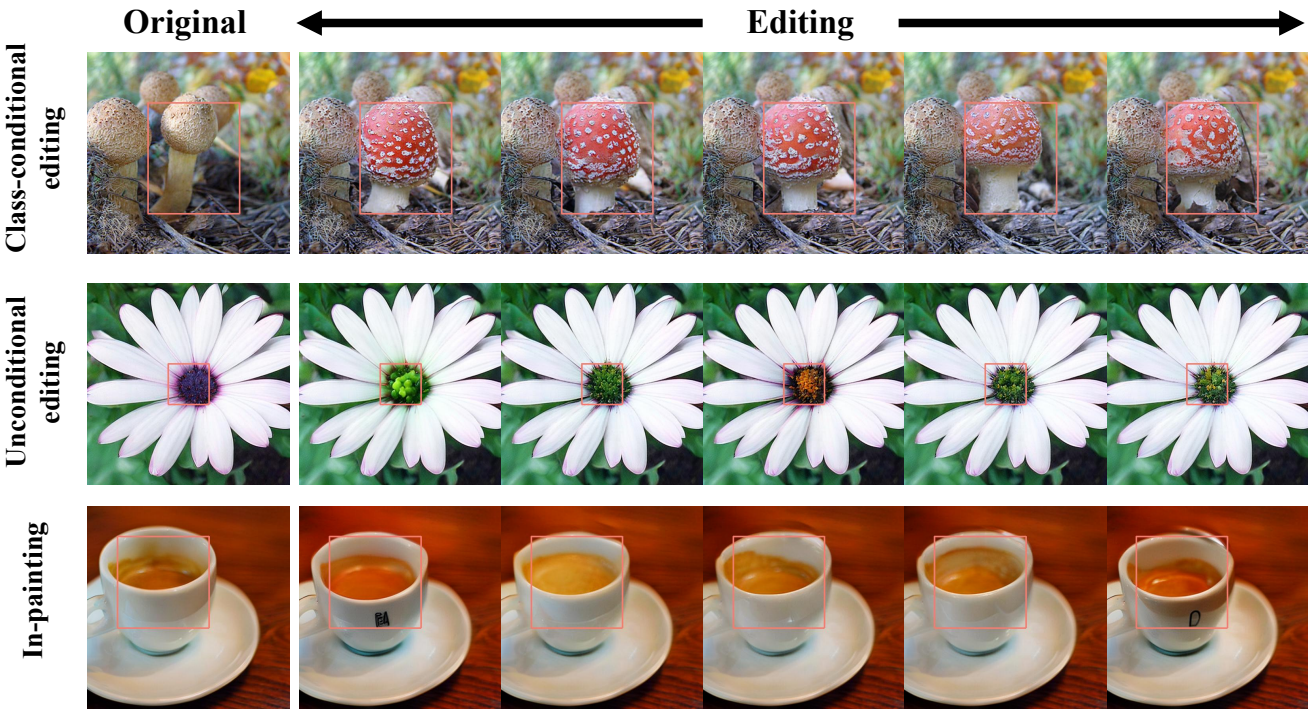


Figure 15. Applying Markov-VAR on zero-shot image editing tasks at 512×512 resolution.

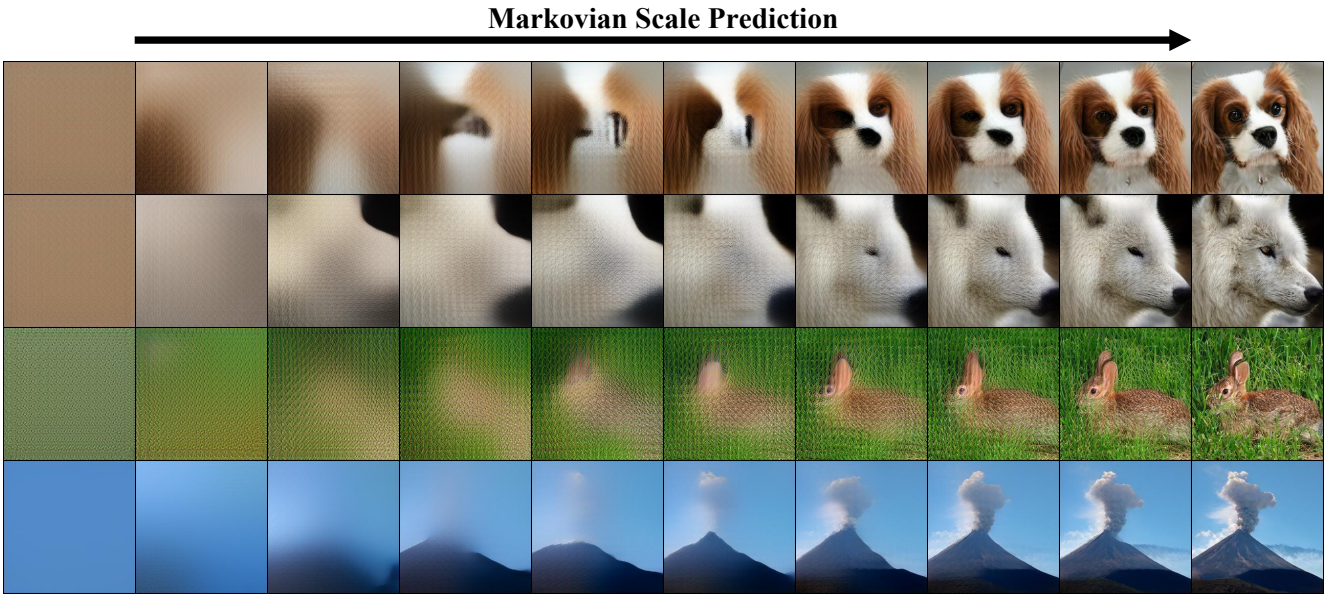


Figure 16. Visualization of generation process in Markov-VAR at 256 × 256 resolution.