Appendix

A. Performance on Non-Residual VAE Models

To assess whether the token redundancy observed in high-resolution stages originates from the next-scale prediction scheme itself or from the residual VAE architecture (used in models like HART and Infinity), we evaluate SparseVAR on **FlexVAR** [29], a next-scale prediction model that does *not* use residual VAEs. Unlike HART or Infinity, FlexVAR directly predicts the entire token map at each stage without computing residuals. This makes it a suitable testbed for isolating the effect of the next-scale prediction paradigm. As shown in Fig. 9, we visualize the ℓ_1 differences between consecutive stages in FlexVAR. We find that low-frequency regions receive minimal refinement in later stages, while high-frequency regions continue to be updated. This confirms that token redundancy is inherent to the multi-stage refinement design of next-scale prediction, independent of residual VAE components.

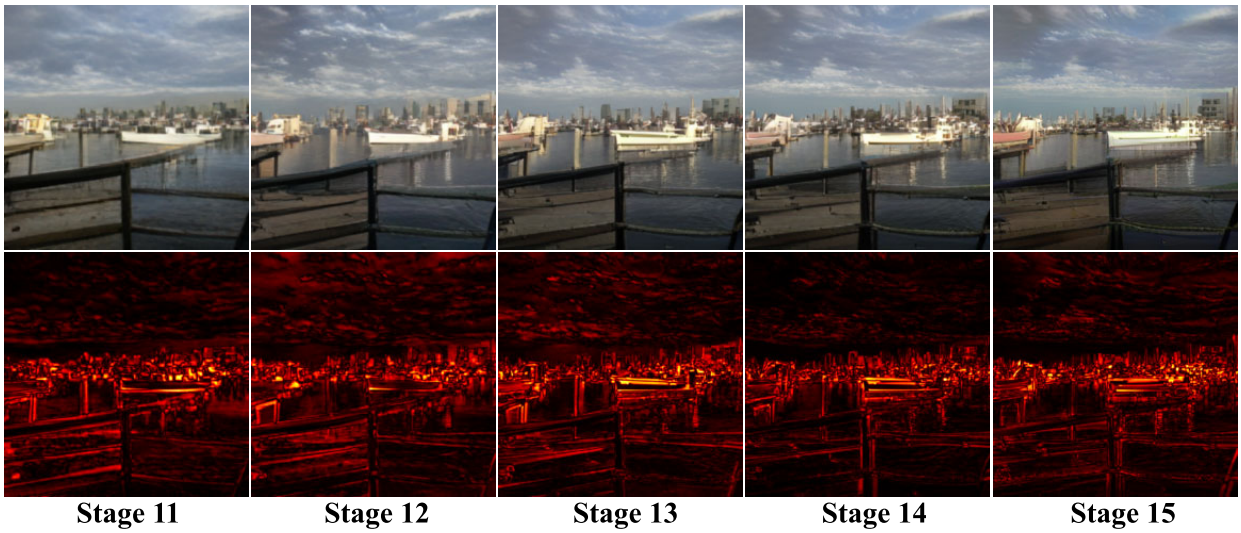

Figure 9. ℓ_1 difference between consecutive stages in FlexVAR. Redder areas indicate greater refinement. Low-frequency regions show minimal change in later stages.

We then apply SparseVAR to FlexVAR by skipping token prediction in low-frequency regions and using interpolated values from the previous stage. As shown in Table 6, this yields inference acceleration with minimal degradation in image quality. Although the performance gain is smaller—due to FlexVAR's lower 512×512 resolution and limited scalability—the results confirm that SparseVAR remains effective even in non-residual VAE settings.

Table 6. Evaluation on ImageNet (512×512) using FlexVAR with and without SparseVAR.

Model	FID ↓	Latency (s) ↓
FlexVAR	4.33	0.52
+ SparseVAR	4.38	0.43

B. Comparison with Token Reduction Methods

We compare SparseVAR with representative token reduction strategies, including both token merging and token selection methods, applied during high-resolution stages. For token merging, we evaluate ToMe [3] and PiToMe (NeurIPS 2024), sweeping sparsity levels from 10% to 50%. For token selection, we test SparseViT (CVPR 2023) and ZipVL [12], both of which are compatible with FlashAttention. We explore sparsity levels ranging from 10% to 80%. As shown in Fig. 10, token merging tends to degrade image quality significantly under higher sparsity settings, while selection methods based on attention scores or token norms fail to identify redundancy effectively in next-scale prediction models like VAR. This is largely due to the strong local dependency patterns in these models, which differ from typical vision transformers.

In contrast, SparseVAR introduces a new insight: that redundancy predominantly stems from low-frequency tokens in later high-resolution stages. It leverages a frequency-aware metric (MSE change in high-frequency blocks) to perform dynamic token exclusion, achieving a better balance between quality and computational efficiency.

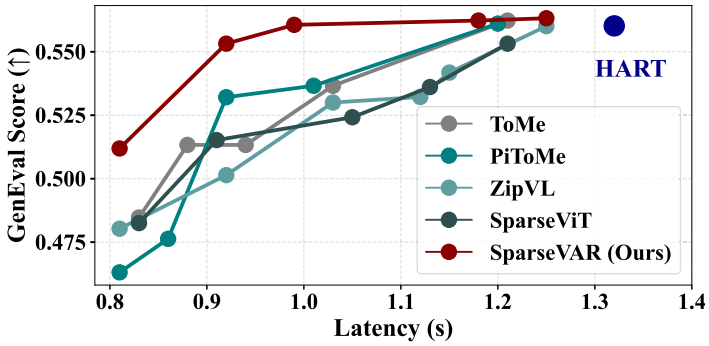

Figure 10. Comparison of token reduction methods on GenEval using HART-0.7B.

C. Performance on Complex Scenes

To further evaluate the robustness of SparseVAR in real-world settings, we conduct qualitative analysis on complex scenes involving multiple subjects and high-frequency content. These include images with crowded human figures, multiple animal faces, textual elements, and fine-grained visual structures. As shown in Fig. 11, SparseVAR is able to maintain high-fidelity generation quality even in these more challenging scenarios. Despite early exiting in low-frequency regions, critical high-frequency areas such as facial features and text remain well-preserved. These results suggest that SparseVAR generalizes beyond simple single-object scenes, and remains effective under diverse and complex image distributions.

Figure 11. Visualization results on complex scenes with multiple subjects and high-frequency details (e.g., text, faces, groups). SparseVAR maintains visual quality while providing acceleration.

D. More Experimental Details

D.1. Implementation Details of Figure 6

Assume that tokens in the k-th stage ($k \geq P$) are candidates for early exclusion. For logits, the logits from the $(k-1)$-th stage are interpolated to match the resolution of the k-th stage. The average cosine similarity between each token and its neighboring 3×3 region is computed. Tokens are classified as low-frequency and marked for early exclusion if their neighboring similarity exceeds τ times the maximum observed similarity across all tokens, where $\tau \in [0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$. For the ℓ_1 metric, the cumulative token maps from the $(k-1)$-th and $(k-2)$-th stages are passed through the VAE decoder to generate

images. The ℓ_1 difference between these two images is then computed. Tokens are classified as low-frequency and marked for early exclusion if their ℓ_1 difference is below τ times the maximum observed ℓ_1 difference across all tokens, where $\tau \in [0.04, 0.05, 0.06, 0.07, 0.1, 0.2, 0.3]$. For the MSE metric, $\tau \in [0.2, 0.3, 0.4, 0.5, 0.6, 0.7]$.

E. More Visualizations of Observations

This subsection provides additional visualization examples corresponding to the observations discussed in the main text.

E.1. Minimal Impact of High-Resolution Residuals on Low-Frequency Regions

In this section, we present additional visualizations of the ℓ_1 changes in the generated images of the accumulated token maps between each of the last four stages and the previous stage. As shown in Figures 12 and 13, it is evident that the feature map r_k generated in high-resolution stages have a limited impact on the low-frequency regions in both HART and Infinity.

Stage 9 Stage 10 Stage 11 Stage 12 Generated Image

Figure 12. More visualizations of HART-0.7B.

Stage 9 Stage 10 Stage 11 Stage 12 Generated Image

Figure 13. More visualizations of Infinity-2B.

E.2. Distinct Regional Focus of Blocks in Next-Scale Prediction Models

In this section, we present more detailed visualizations of the MSE changes for some blocks in the final five stages. As shown in Figure 14 and Figure 15, it is evident that different blocks focus on distinct regions, with some emphasizing high-frequency regions and others targeting low-frequency regions. Based on the most effective clarification, we select the 3rd block for Infinity-2B and the 16th block for HART-0.7B.

E.3. High Logits Similarity in Certain Low-Frequency Regions

Low-frequency regions refer to areas in the image with slow variations, where an intuitive hypothesis is that the predicted residuals for these regions at high-resolution stages should be relatively similar. To explore the similarity of predicted residuals

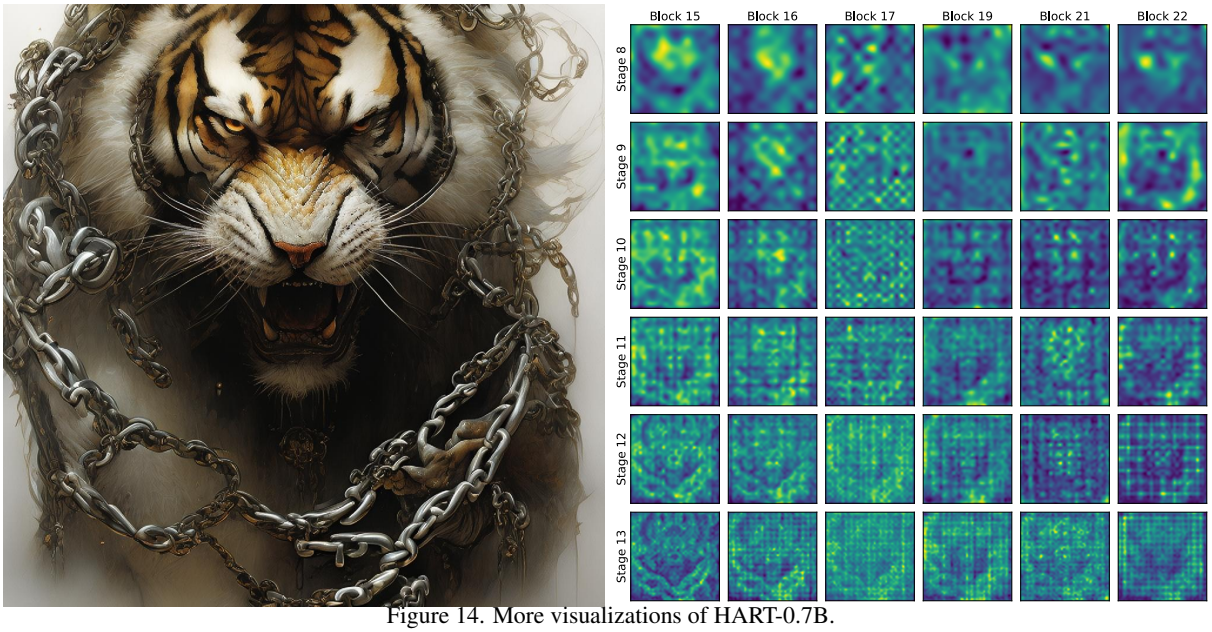

Figure 14. More visualizations of HART-0.7B.

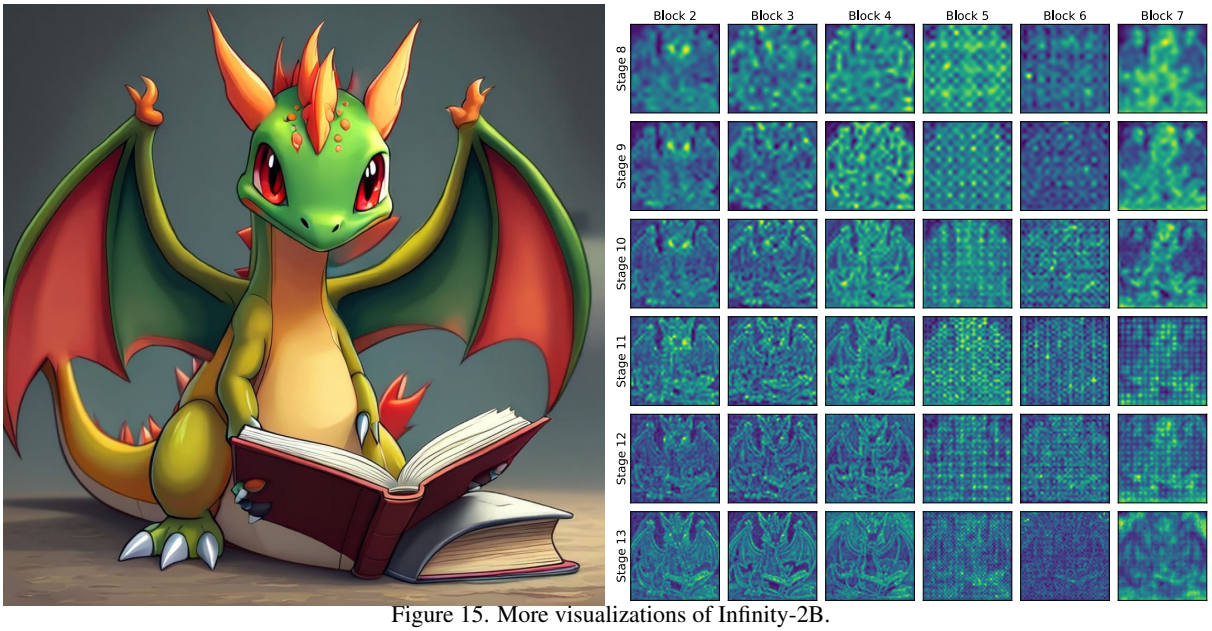

Figure 15. More visualizations of Infinity-2B.

in low-frequency regions, we visualize the average cosine similarity of logits between each token and its neighboring 3×3 region at stages 8-11 of HART and Infinity. As shown in Figure 16 and Figure 17, for certain low-frequency regions, the logits predicted at each token position exhibit high similarity with those of neighboring tokens. Based on this observation, *we propose retaining a subset of anchor tokens to represent the predictions of their neighboring regions, thereby improving the generation quality of low-frequency areas.*

F. Visualizations of Attention Maps in Next-Scale Prediction

We visualize the attention maps of three blocks from HART-0.7B across the last four stages. As shown in the figure, the next-scale prediction model exhibits high attention scores for nearly all tokens with respect to their neighboring tokens, while the attention scores for more distant tokens are close to zero.

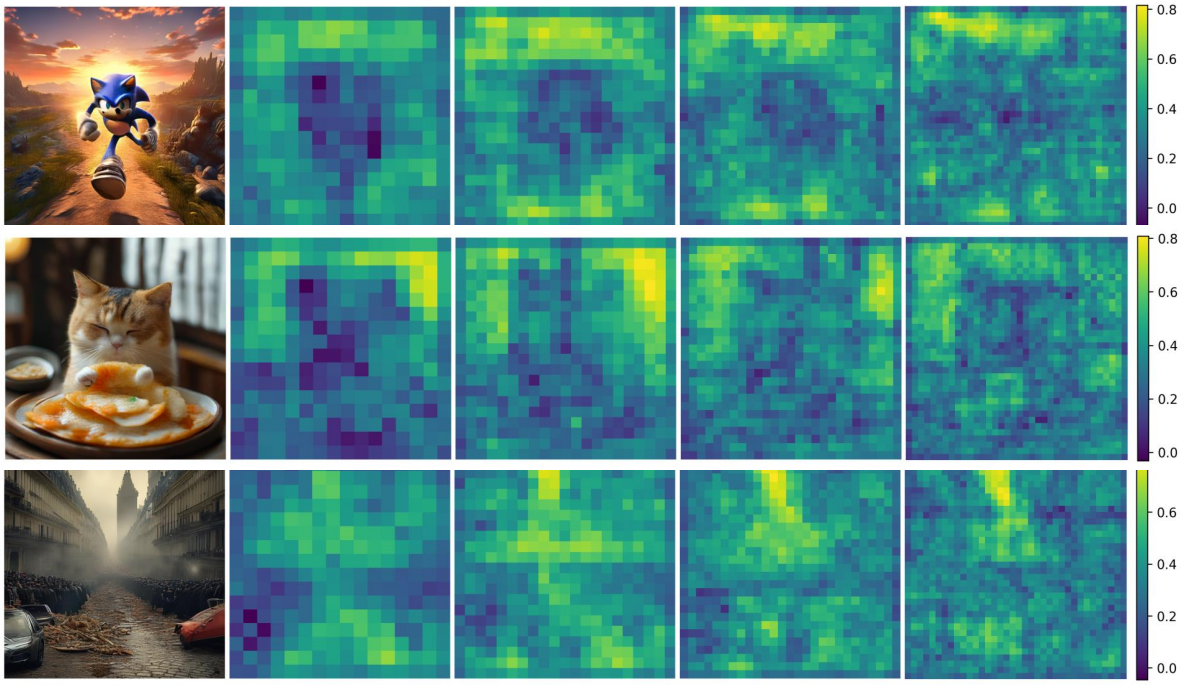

Figure 16. More visualizations of HART-0.7B.

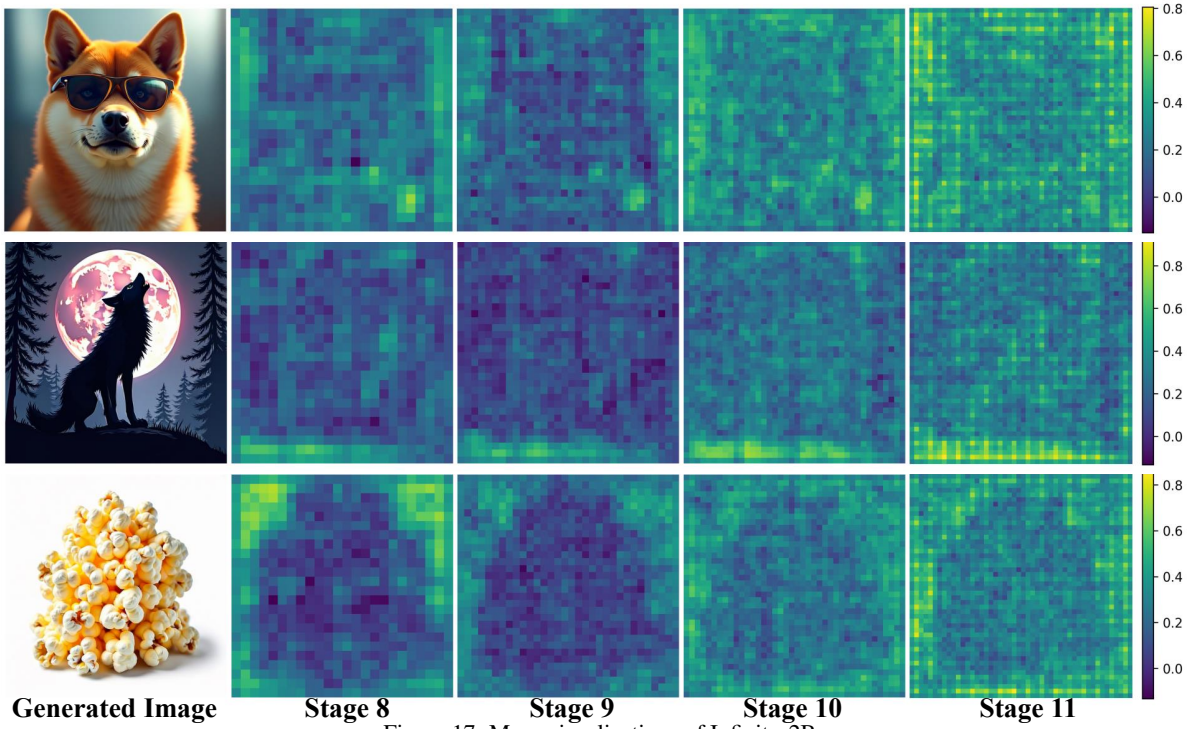

Figure 17. More visualizations of Infinity-2B.

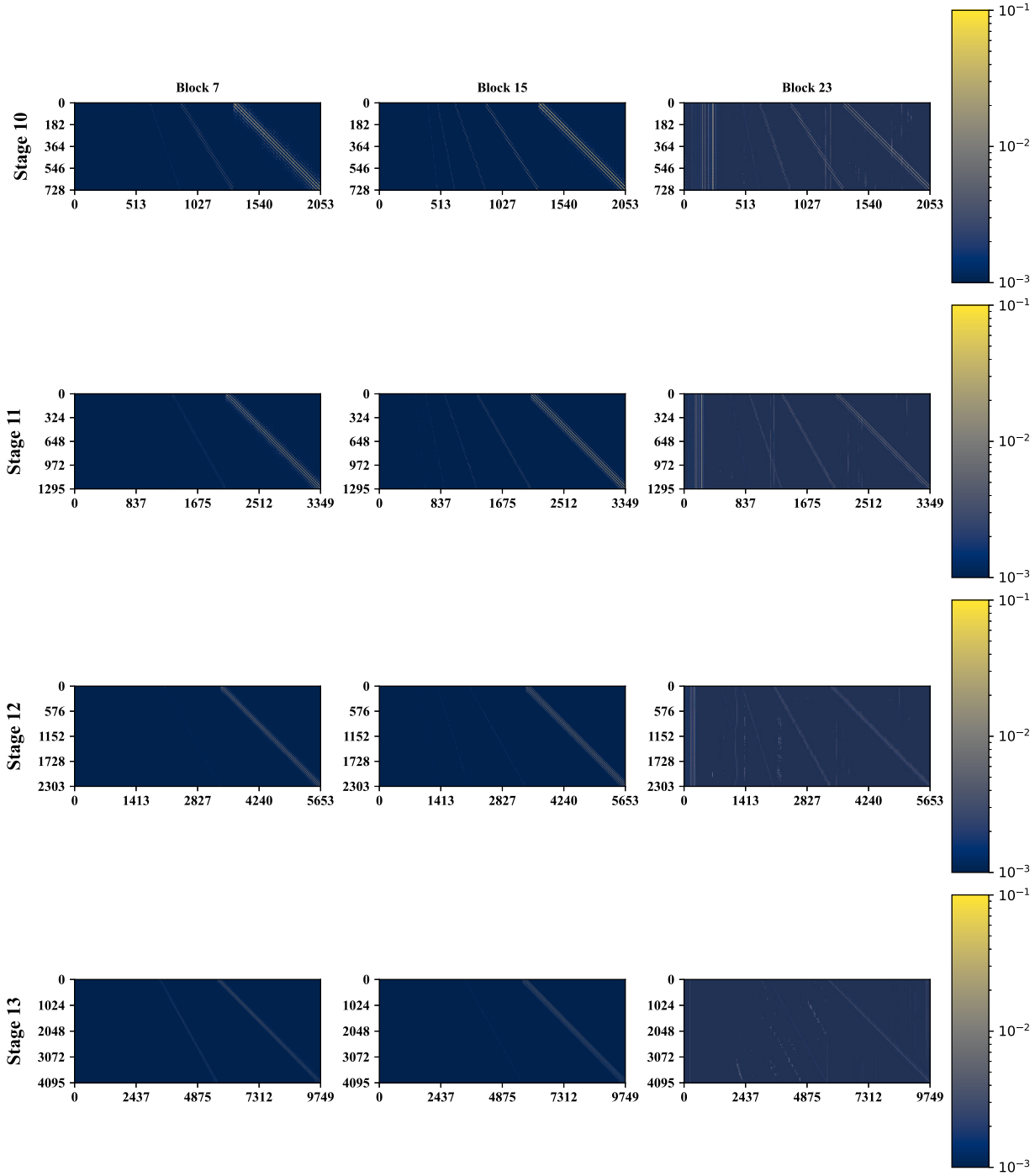

Figure 18. Visualization of attention maps of SparseVAR. The x-axis represents the indices of all tokens in the current stage and the KV cache, while the y-axis represents the token indices in the current stage.

G. More Qualitative Visualizations

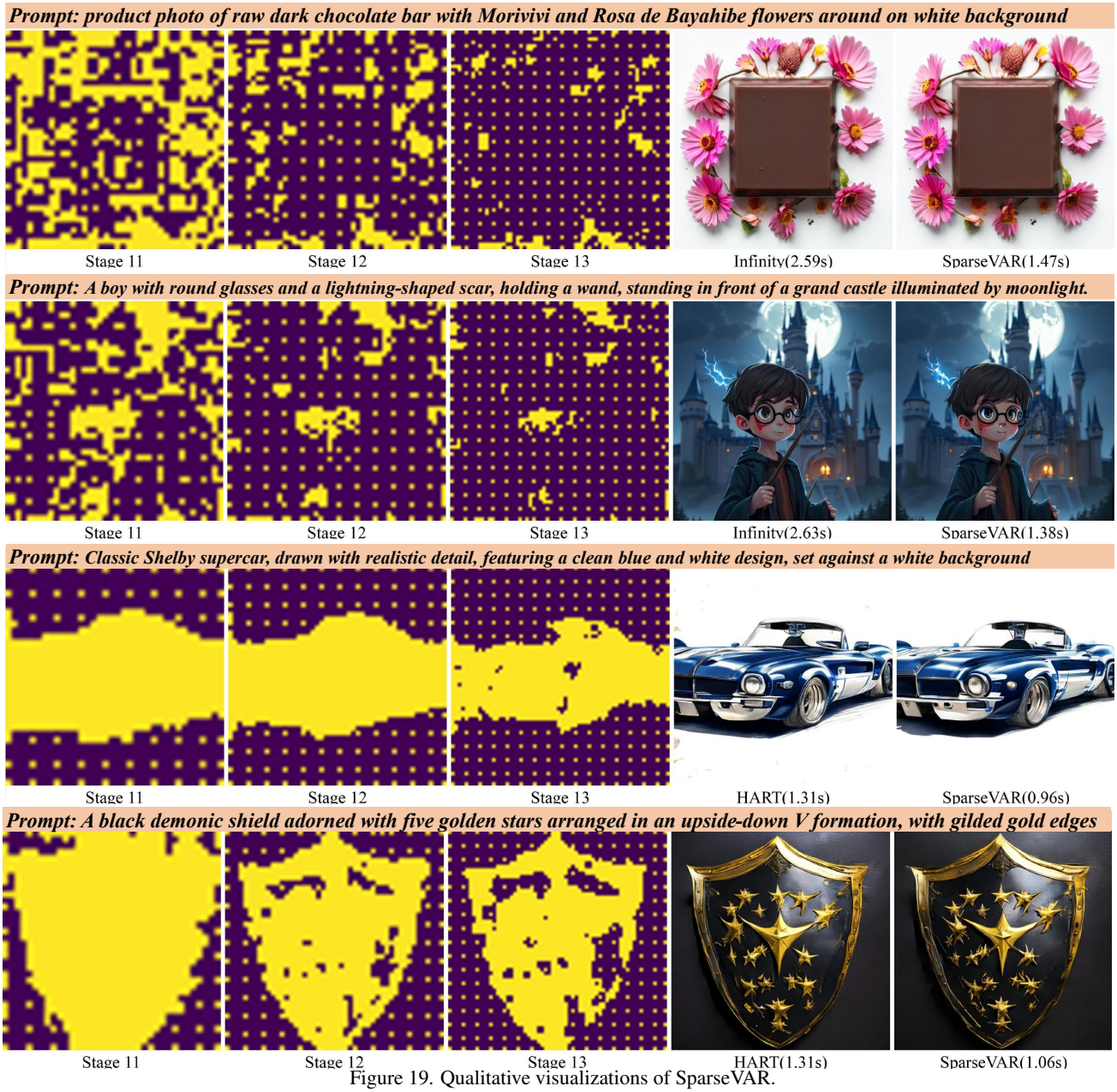

Figure 19. Qualitative visualizations of SparseVAR.