

1 In this document, we provide supplementary material that we could not fit into the main manuscript  
2 due to the page limit. It includes detailed explanations, visualization results, and quantitative  
3 experiments.

## 4 A Robustness to Calibration Data

5 To evaluate the stability of our drafter/refiner identification method across varying calibration set  
6 sizes, we conducted an ablation study using prompts sampled from the LAION-Art dataset. We  
7 systematically evaluated ScaleKV’s performance using calibration sets ranging from a single prompt  
8 to 128 prompts, measuring the resulting FID scores on the MS-COCO validation set. As demonstrated  
9 in Figure 1, the FID score remains stable at 2.53 with zero standard deviation across the entire range of  
10 calibration set sizes. This exceptional consistency indicates that the attention patterns distinguishing  
11 drafters from refiners represent fundamental architectural properties of VAR models rather than  
12 dataset-specific characteristics. The immediate convergence with even a single calibration sample  
13 demonstrates that our Attention Selectivity Index effectively captures the intrinsic scale-dependent  
14 attention behaviors—dispersed attention for drafters and concentrated attention for refiners—without  
15 requiring extensive statistical sampling. These findings validate that ScaleKV’s layer categorization  
16 is both theoretically sound and practically robust.

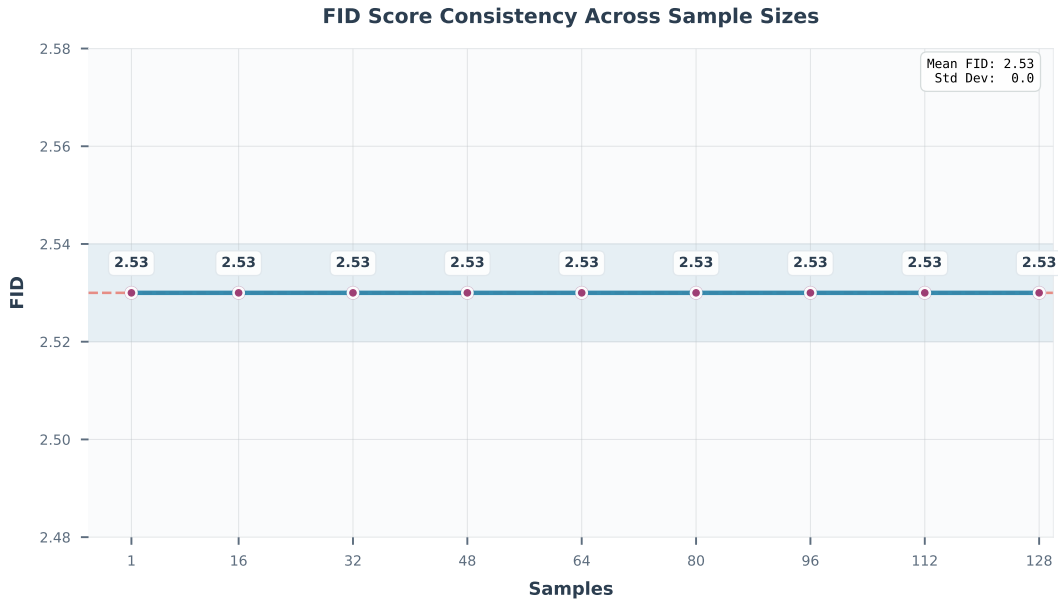


Figure 1: FID score consistency across calibration set sizes. ScaleKV maintains stable performance (FID = 2.53,  $\sigma = 0.0$ ) from 1 to 128 calibration samples, demonstrating the robustness of our drafter/refiner identification method.

## 17 B Attention Map Visualization

18 Figures 2 and 3 present representative attention maps from transformer layers identified as drafters and  
19 refiners at scale 7, providing empirical validation for our layer categorization framework. The drafter  
20 layer (Block 23) exhibits distinctly dispersed attention patterns across multiple attention heads, with  
21 weights distributed broadly across the spatial dimensions to capture global contextual information  
22 from preceding scales. This dispersed mechanism enables comprehensive integration of hierarchical  
23 features, justifying the larger cache allocation for such layers. In contrast, the refiner layer (Block 29)  
24 demonstrates highly concentrated attention patterns, with attention heads focusing predominantly  
25 on localized spatial regions within the current token map. These contrasting attention behaviors  
26 provide strong empirical evidence for our differentiated cache management strategy: drafters require  
27 substantial cache capacity to maintain broad contextual access while refiners can operate effectively  
28 with significantly reduced cache allocation due to their localized processing nature.

## 29 C Additional Qualitative Results

30 Figures 4 and 5 present comprehensive galleries of images generated by ScaleKV-compressed Infinity-  
31 8B and Infinity-2B models, respectively, demonstrating the practical effectiveness of our compression  
32 framework across diverse visual content. These compressed models operate with merely 10% of  
33 the original KV cache memory requirement, yet maintain exceptional generation quality across  
34 various image categories including natural scenes, objects, portraits, and artistic compositions. These  
35 results demonstrate that ScaleKV achieves substantial memory reduction without compromising the  
36 generative capabilities of VAR models, making high-quality image synthesis feasible in memory-  
37 constrained deployment scenarios.

## 38 D Limitations

39 In this analysis, we critically examine the constraints of our methodology.

40 First, while ScaleKV demonstrates robust compression performance across models of varying  
41 capacities, evaluation on larger VAR models would provide additional insights into our method’s  
42 scalability. Due to the limited availability of large-scale models, our evaluation was restricted to  
43 Infinity-8B, currently the largest available VAR model. Testing on models with greater capacity,  
44 such as those with 20B parameters, would enable more comprehensive assessment of ScaleKV’s  
45 scalability. Second, ScaleKV functions as a post-training KV cache compression solution that relies  
46 on pre-trained VAR models and mirrors the original model’s outputs. Therefore, if the baseline  
47 quality of the original VAR models is unsatisfactory, achieving high-quality results with our method  
48 could be challenging.

## 49 E Societal impacts

50 This work introduces a new KV cache compression framework for VAR models that addresses critical  
51 memory bottlenecks in high-resolution image generation. By reducing memory requirements to 10%  
52 of the original capacity while maintaining generation quality, our method enhances the accessibility  
53 of advanced image synthesis technologies and carries several important societal implications. First, it  
54 democratizes access to high-quality image generation by enabling deployment on consumer-grade  
55 hardware and edge devices, thereby benefiting creative industries and educational institutions that  
56 previously lacked the computational resources for such applications. Second, the reduced memory  
57 footprint results in lower energy consumption during inference, contributing to more sustainable  
58 AI deployment practices. Third, by enabling ultra-high resolution generation at scales up to 4K,  
59 our framework creates new opportunities for professional content creation, medical imaging, and  
60 scientific visualization applications.

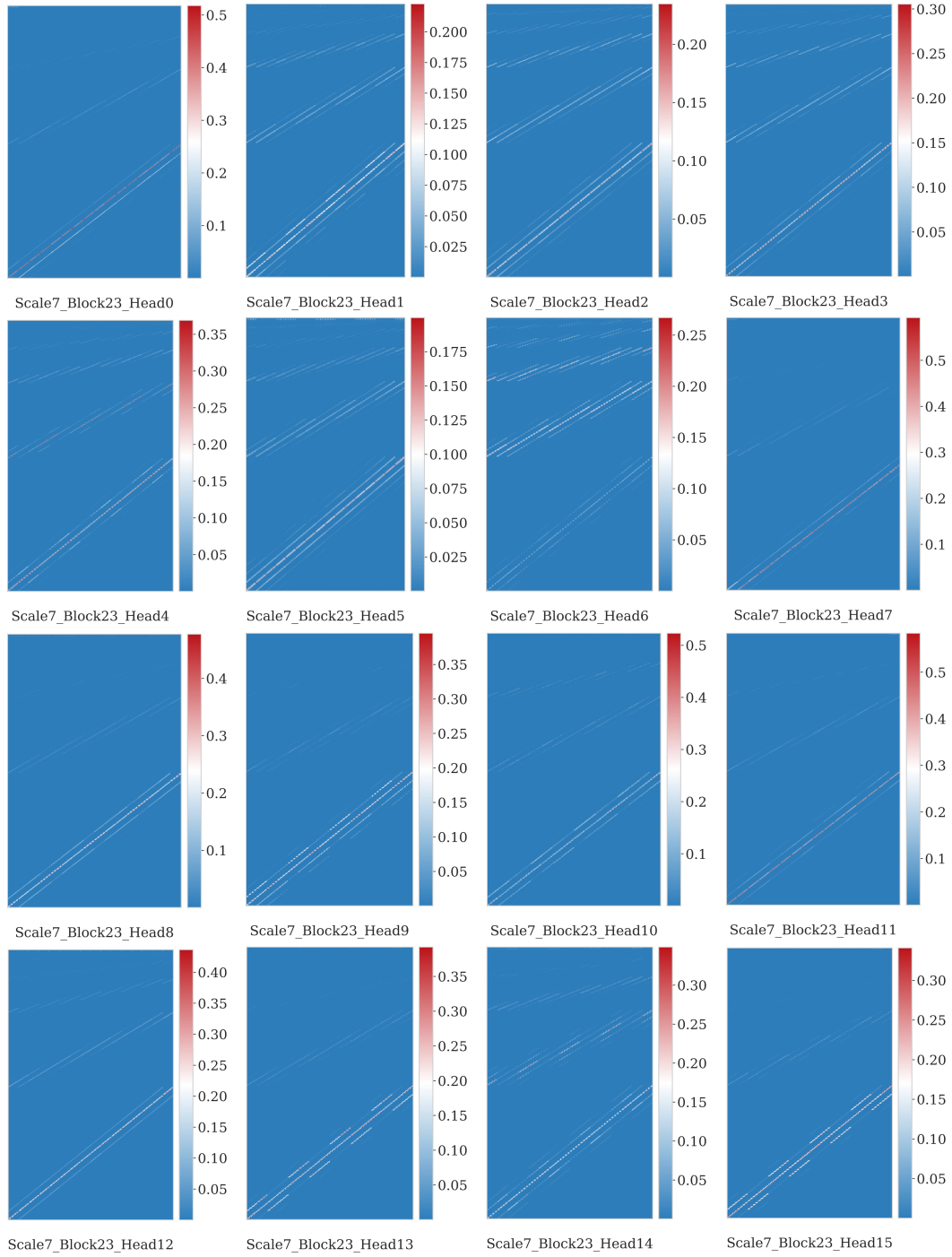


Figure 2: Visualization of Drafter Layer Attention Maps.

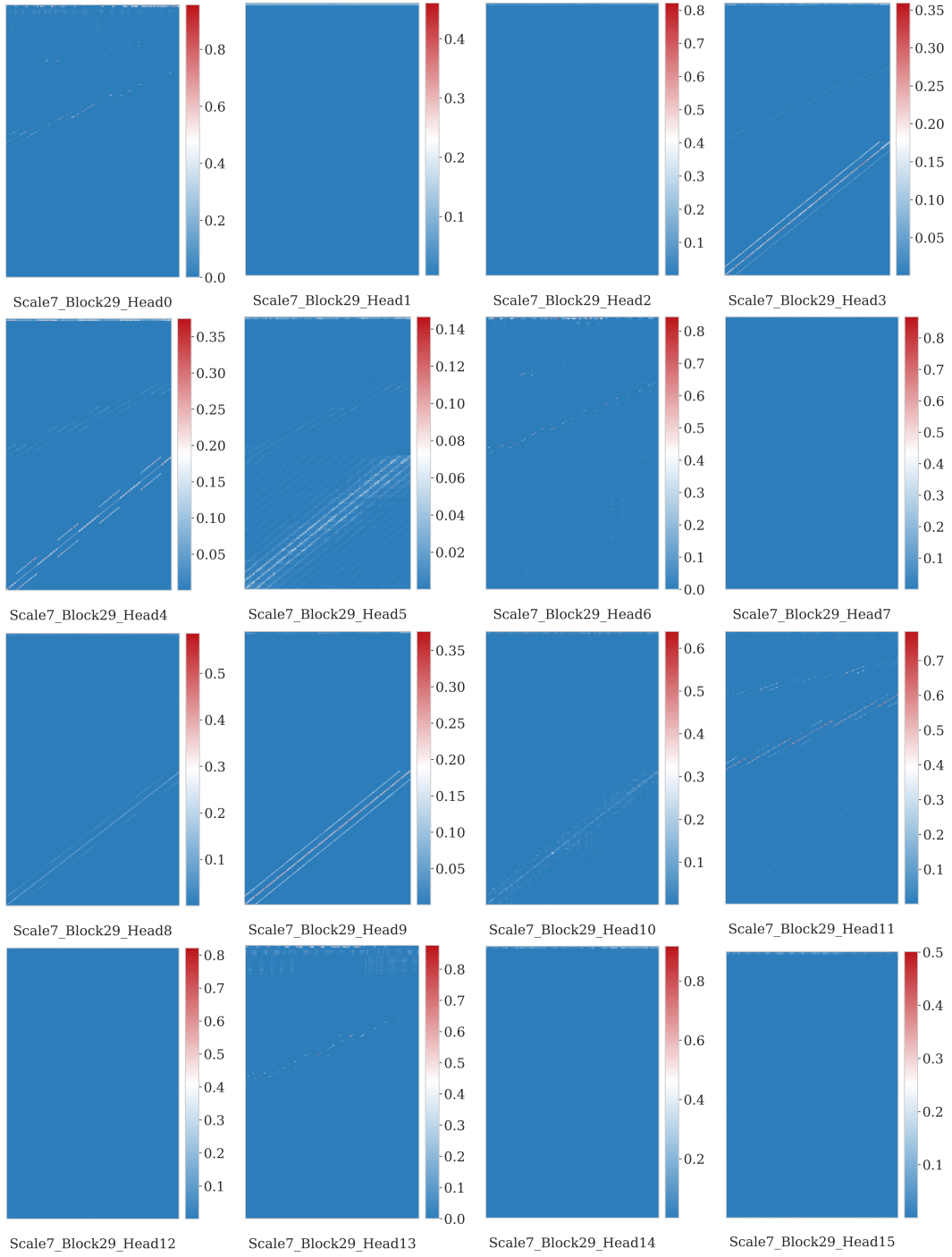


Figure 3: Visualization of Refiner Layer Attention Maps.





Figure 4: Generated Images from ScaleKV-Compressed Infinity-8B.



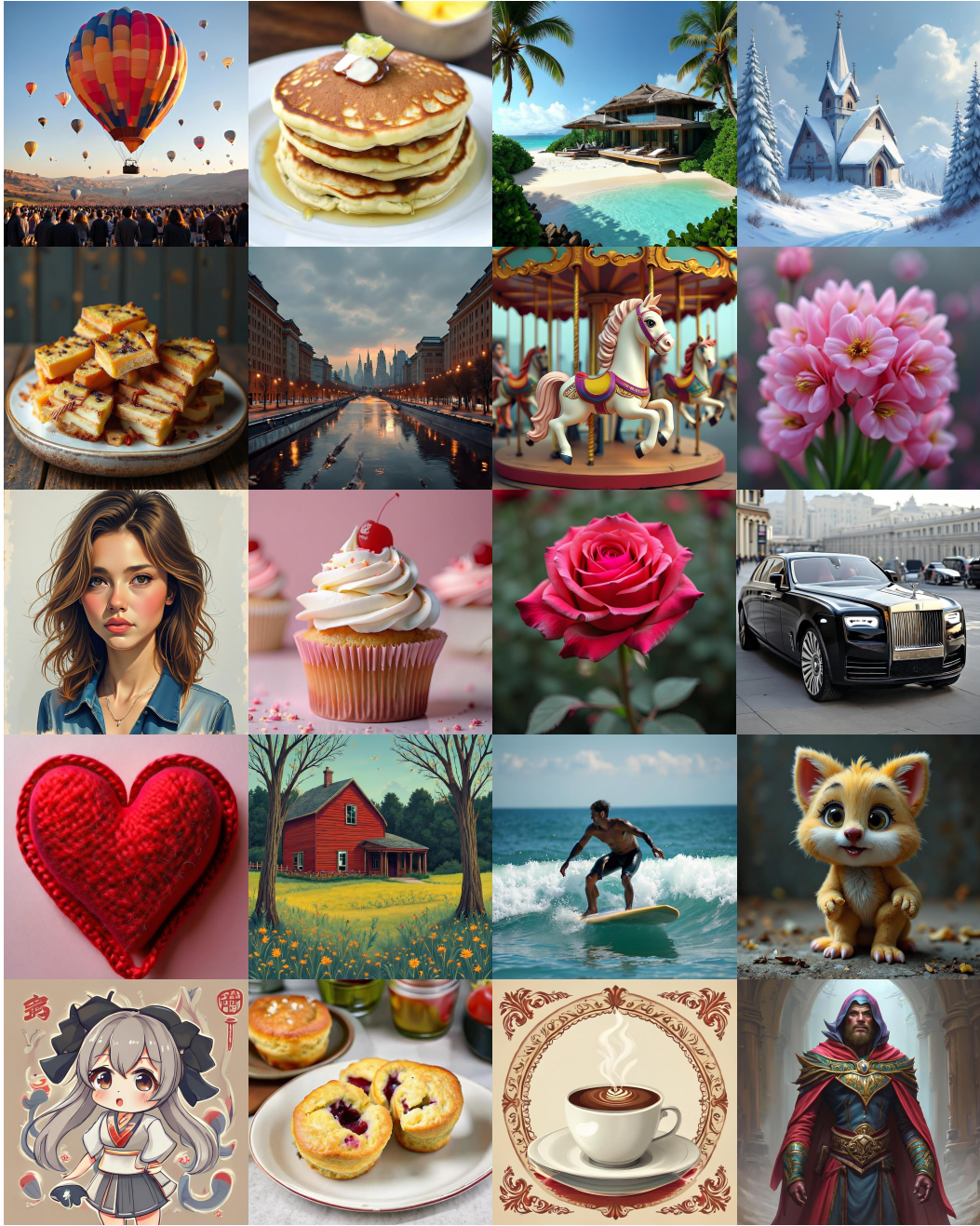


Figure 5: Generated Images from ScaleKV-Compressed Infinity-2B.