MipKV: A Sparsify-then-Recover Paradigm for Accelerating Large Vision-Language Model Pre-Filling

Supplementary Material

6. Detailed Evaluation Results

Detailed results for regular image-understanding tasks are summarized in Table 2, and high-resolution or video-related benchmarks are reported in Table 1.

7. More Visual Demonstration

Figures 13 to 17 show a complete attention distribution in all decoder layers. The attention sink pattern appears in most layers while intermittent scattering show in layer 14 and 20.

Figures 18 to 22 show the KV Deviation with and without the token recovery in all decoder layers. High KV Deviations are largely suppressed.

8. Limitations

Due to the computational cost and resource constraints associated with evaluating very large multimodal models, our current experiments focus on Qwen2-VL-2B and LLaVA-1.5-7B. Extending MipKV to higher-capacity models–such as 72B-scale variants–and testing its genrality across different LVLM architectures would provide a more comprehensive understanding of its applicability.

Table 1. **Performance of MipKV compared with other acceleartion methods on Qwen2-VL-2B across high-resolution and video benchmarks.** Attention-based methods are excluded for their high memory consumption.

Methods	InfoVQA	DocVQA	HRBench^{4K}	HRBench^{8K}	MVbench	Avg.(%)
upper bound, 100% visual tokens						
Vanilla	66.0	90.1	59.2	58.1	66.0	100
retain or recover 80% visual tokens						
Resize	64.0	**90.0**	**62.1**	56.2	**60.3**	**99.7**
DART	61.0	88.0	58.2	54.9	59.3	96.3
V^{2}Drop	61.0	89.0	58.0	55.8	59.5	96.8
MipKV	**65.0**	**90.0**	59.1	**57.9**	60.0	99.5
retain or recover 60% visual tokens						
Resize	**62.0**	**90.0**	**60.1**	54.8	**59.8**	**97.8**
DART	54.0	80.0	58.4	54.4	58.5	92.0
V^{2}Drop	57.0	85.0	56.9	53.6	57.2	92.8
MipKV	61.0	89.0	58.9	**57.4**	59.1	97.5
retain or recover 40% visual tokens						
Resize	**60.0**	**90.0**	**60.4**	53.1	**59.6**	**96.7**
DART	43.0	66.0	57.1	53.5	55.9	84.0
V^{2}Drop	48.0	78.0	53.6	52.9	53.6	86.0
MipKV	58.0	86.0	59.0	**56.8**	58.5	95.6
retain or recover 20% visual tokens						
Resize	52.0	**89.0**	54.9	51.9	56.1	90.5
DART	30.0	43.0	56.1	50.6	50.7	71.9
V^{2}Drop	35.0	60.0	51.6	49.2	49.6	74.8
MipKV	**55.0**	82.0	**58.0**	**56.0**	**57.5**	**92.8**
retain or recover 10% visual tokens						
Resize	44.0	**84.0**	54.2	46.9	52.1	83.7
DART	24.0	26.0	54.2	48.2	45.8	63.1
V^{2}Drop	27.0	44.0	48.4	46.9	46.4	65.9
MipKV	**52.0**	81.0	**57.4**	**55.6**	**56.9**	**91.2**

Table 2. **Performance of MipKV compared with other acceleration methods on Qwen2-VL-2B across image-understanding benchmarks.** All experiments are conducted on a NVIDIA RTX 4090 GPU. The best results for each task are highlighted in bold, and underlined numbers indicate the second-best performance. "–" denotes out-of-memory errors, and "#" indicates that the input images are too small to be processed.

Methods	**MME**	**MMB**	**SQA**	**GQA**	**POPE**	**ChartQA**	**OcrBench**	**VQA^{text}**	**Avg.(%)**
upper bound, 100% visual tokens									
Vanilla	1875.2	71.5	58.8	59.9	86.0	72.3	765	79.5	100
retain or recover 80% visual tokens									
Resize	1865.0	70.1	**59.1**	59.4	85.8	71.9	736.0	78.3	98.9
FastV	-	70.3	58.4	59.6	85.7	68.3	-	-	98.3
VisionZip	1834.8	70.2	58.5	58.7	85.6	64.1	639.0	70.0	94.2
DART	**1916.0**	69.4	58.2	59.4	85.5	69.1	697.0	78.8	97.8
V^{2}Drop	1878.3	70.7	57.9	59.5	85.5	68.4	669.0	78.2	97.1
MipKV	1880.2	**71.1**	59.0	**59.9**	**86.0**	**72.3**	**761.0**	**79.1**	**99.9**
retain or recover 60% visual tokens									
Resize	1867.6	69.1	**60.0**	58.5	84.9	68.3	715.0	77.1	97.4
FastV	-	68.4	58.9	58.6	84.0	55.7	-	-	93.7
VisionZip	1816.4	68.4	58.4	57.6	85.0	60.0	557.0	68.9	91.2
DART	1852.7	66.6	59.0	58.2	83.6	58.9	577.0	76.4	92.5
V^{2}Drop	1880.6	69.3	59.1	58.7	84.5	59.8	615.0	76.0	94.1
MipKV	**1892.1**	**70.8**	59.3	**59.7**	**86.1**	**70.7**	**749.0**	**78.9**	**99.4**
retain or recover 40% visual tokens									
Resize	1832.3	68.2	57.3	55.6	82.0	65.0	680.0	73.0	93.7
FastV	-	66.3	57.4	56.3	81.0	34.0	-	-	85.1
VisionZip	1800.4	65.8	56.9	56.1	84.3	54.5	460.0	65.9	86.9
DART	1779.6	63.1	57.5	55.8	79.7	47.5	455.0	71.1	85.2
V^{2}Drop	1824.6	68.2	57.5	56.5	82.4	43.2	525.0	73.7	87.7
MipKV	**1883.5**	**70.8**	**60.0**	**59.5**	**86.6**	**68.2**	**739.0**	**78.3**	**98.9**
retain or recover 20% visual tokens									
Resize	1776.7	63.9	55.4	52.3	76.2	40.7	#	66.1	84.8
FastV	-	58.2	54.1	50.8	71.1	18.5	-	-	73.3
VisionZip	1763.4	57.6	53.4	51.6	79.5	41.2	276.0	56.4	76.0
DART	1596.2	54.4	57.4	50.4	68.9	32.4	295.0	55.4	72.0
V^{2}Drop	1763.4	62.7	53.8	52.4	77.2	24.4	397.0	66.4	77.5
MipKV	**1846.5**	**70.2**	**59.5**	**59.5**	**86.8**	**64.1**	**710.0**	**77.7**	**97.2**
retain or recover 10% visual tokens									
Resize	1710.0	60.4	53.4	48.6	70.8	24.1	#	56.0	76.2
FastV	-	41.4	50.9	45.0	55.8	15.0	-	-	61.0
VisionZip	1503.9	44.9	53.7	46.2	71.0	29.3	140.0	41.1	63.1
DART	1449.8	47.0	55.6	46.0	55.6	21.3	190.0	38.1	60.2
V^{2}Drop	1605.2	57.2	53.2	48.6	69.1	16.0	286.0	59.0	68.9
MipKV	**1855.1**	**69.7**	**59.2**	**59.3**	**86.8**	**60.0**	**688.0**	**77.0**	**95.8**

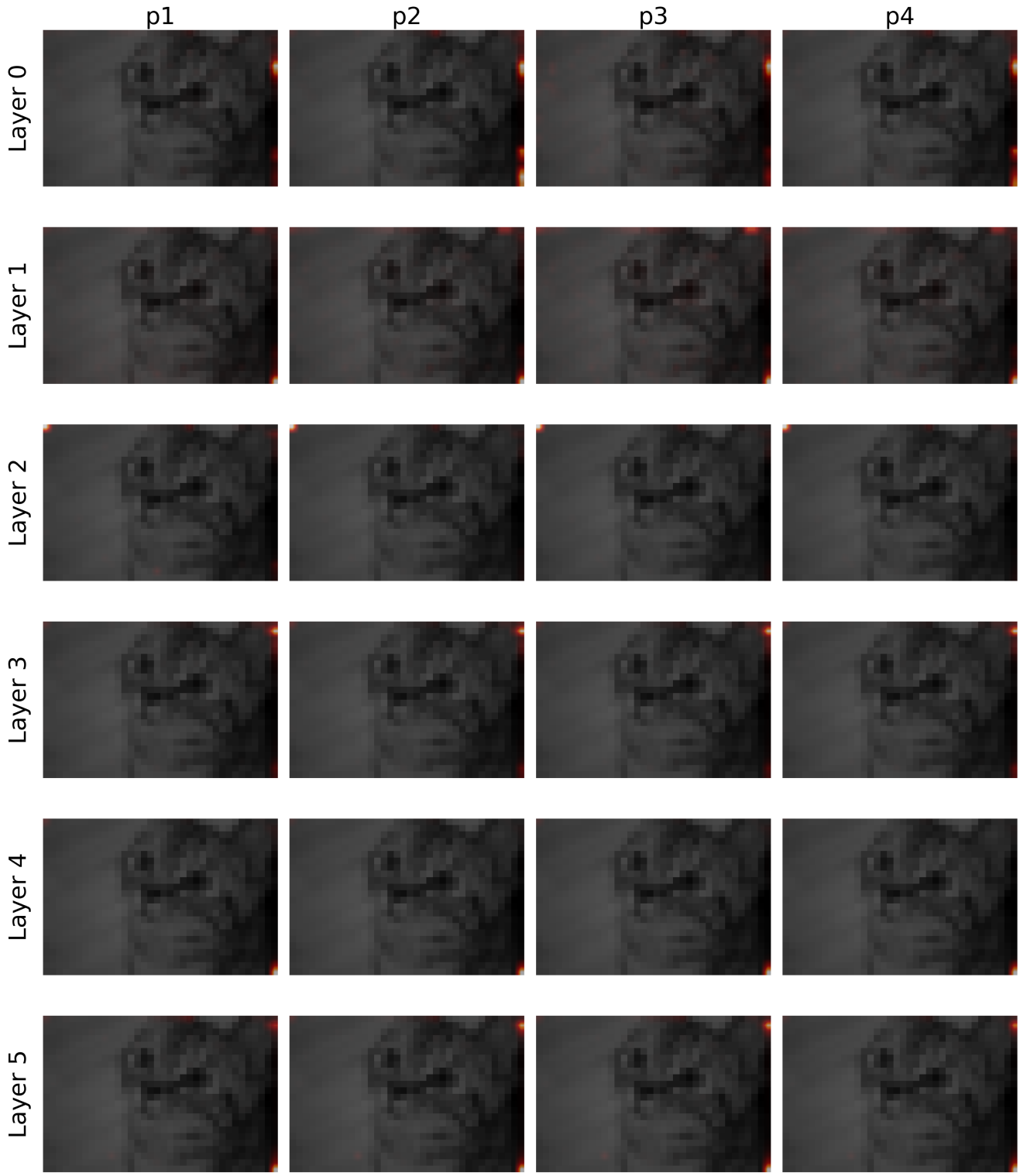

Figure 13. Attention distribution (Layer 0 – Layer 5)

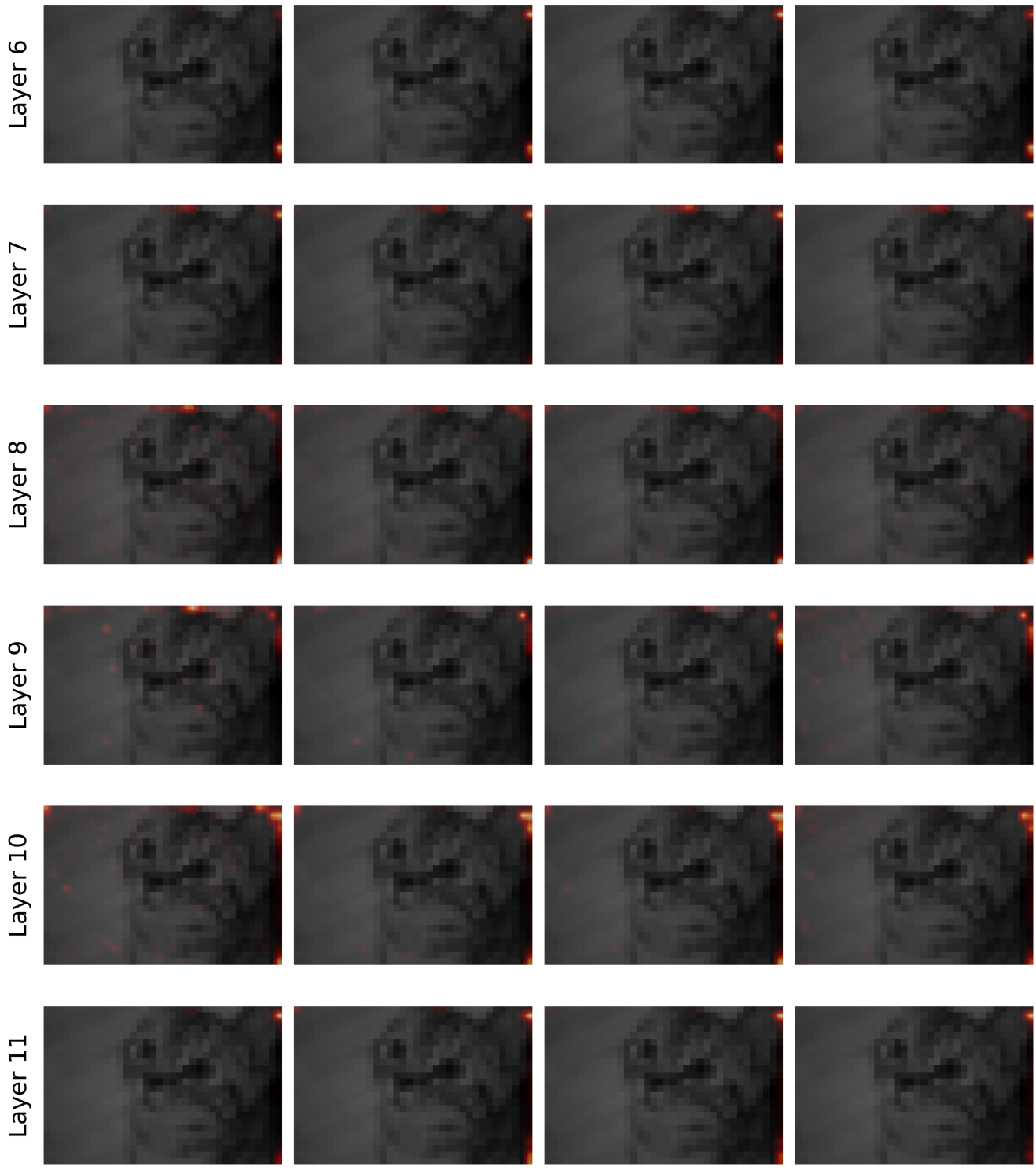

Figure 14. Attention distribution (Layer 6 – Layer 11)

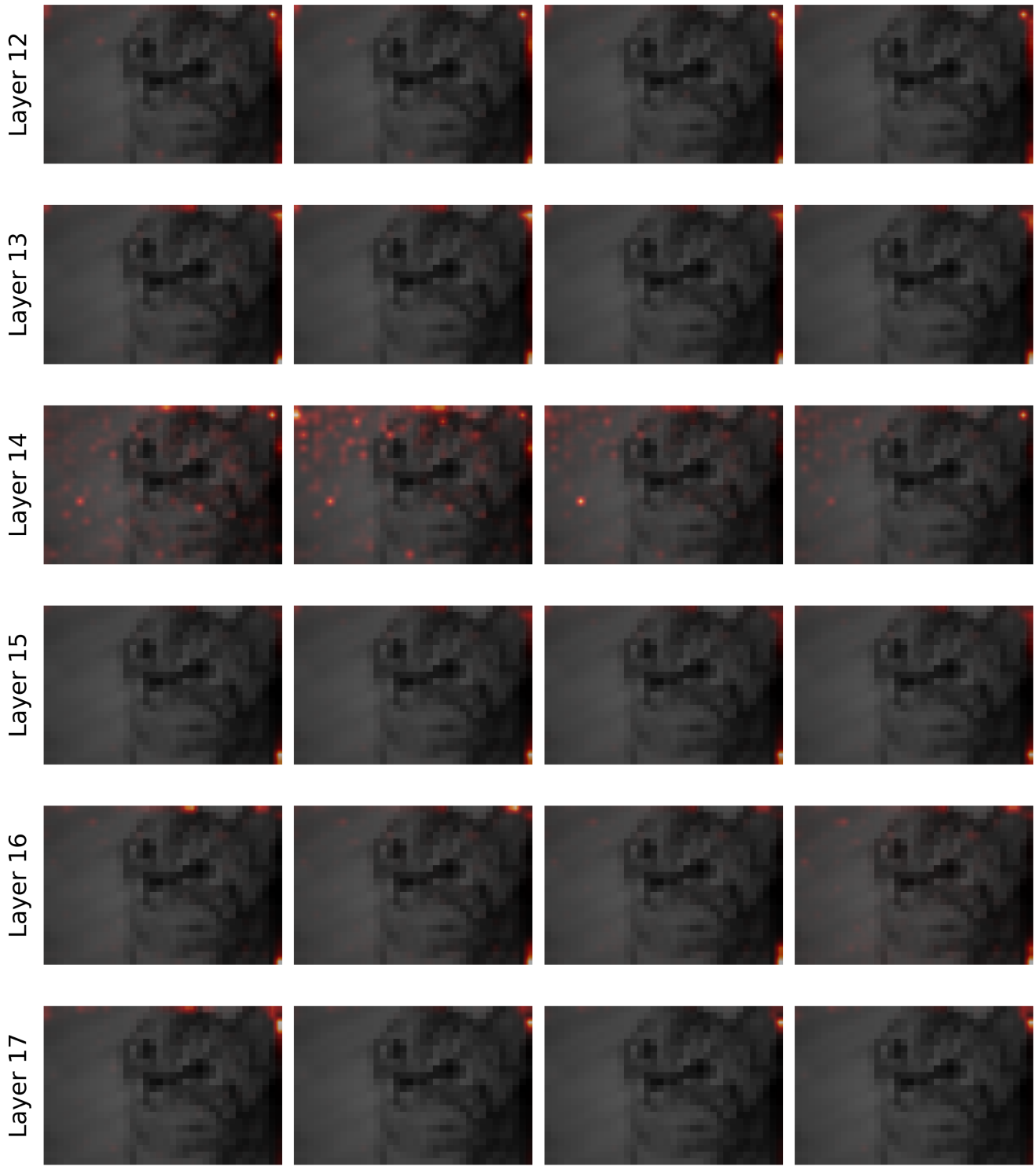

Figure 15. Attention distribution (Layer 12 – Layer 17)

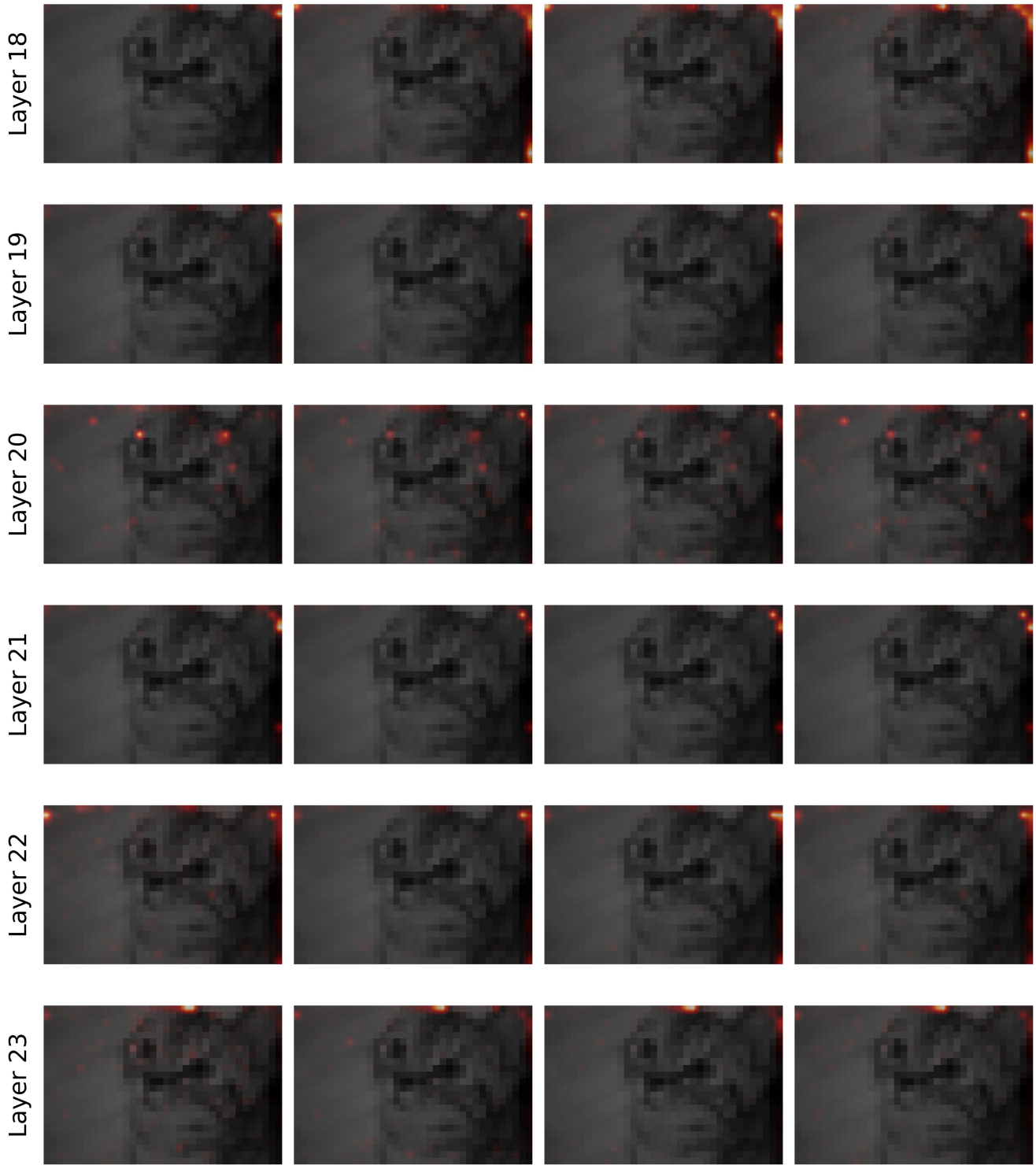

Figure 16. Attention distribution (Layer 18 – Layer 23)

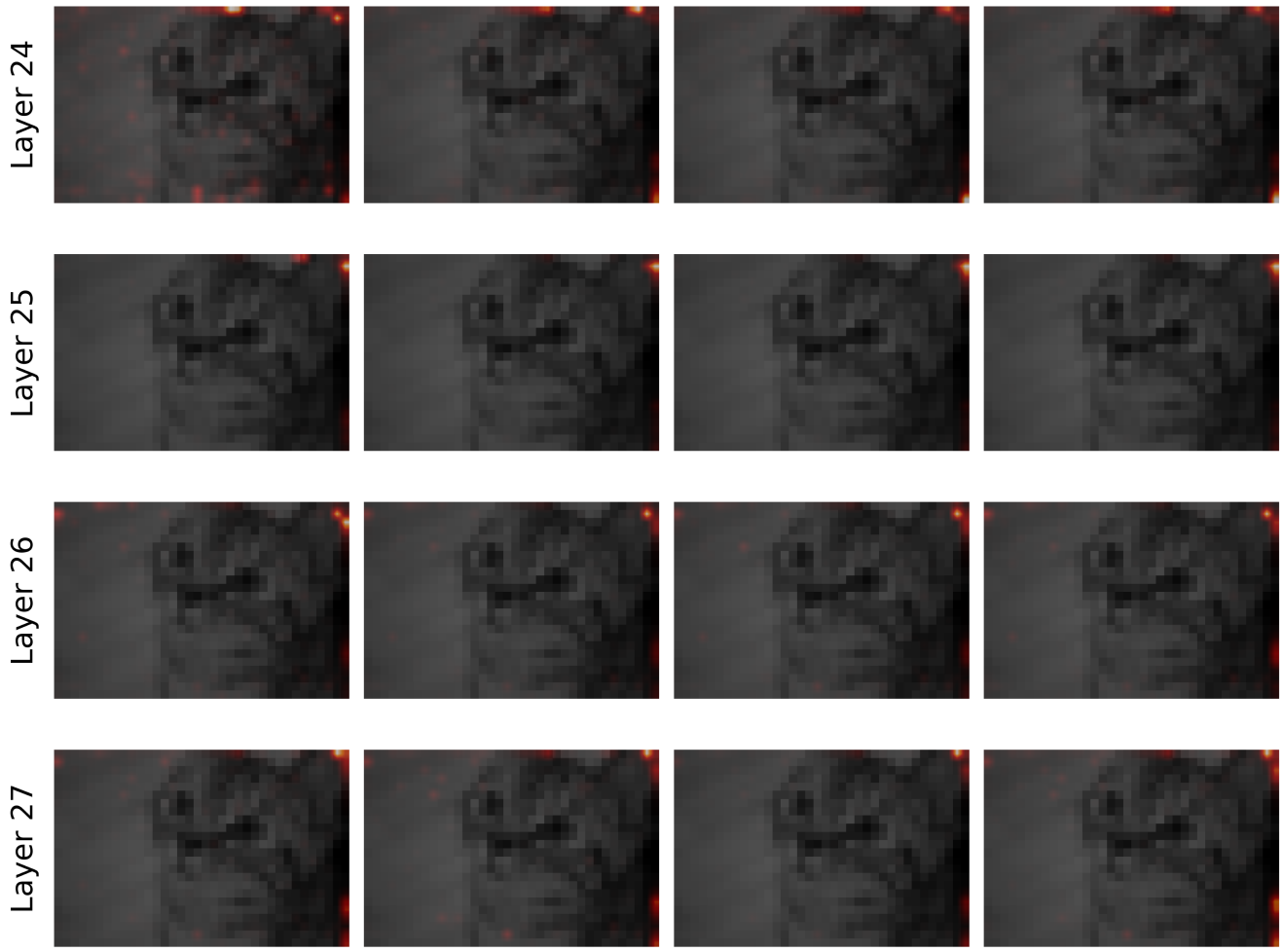

Figure 17. Attention distribution (Layer 24 – Layer 27)

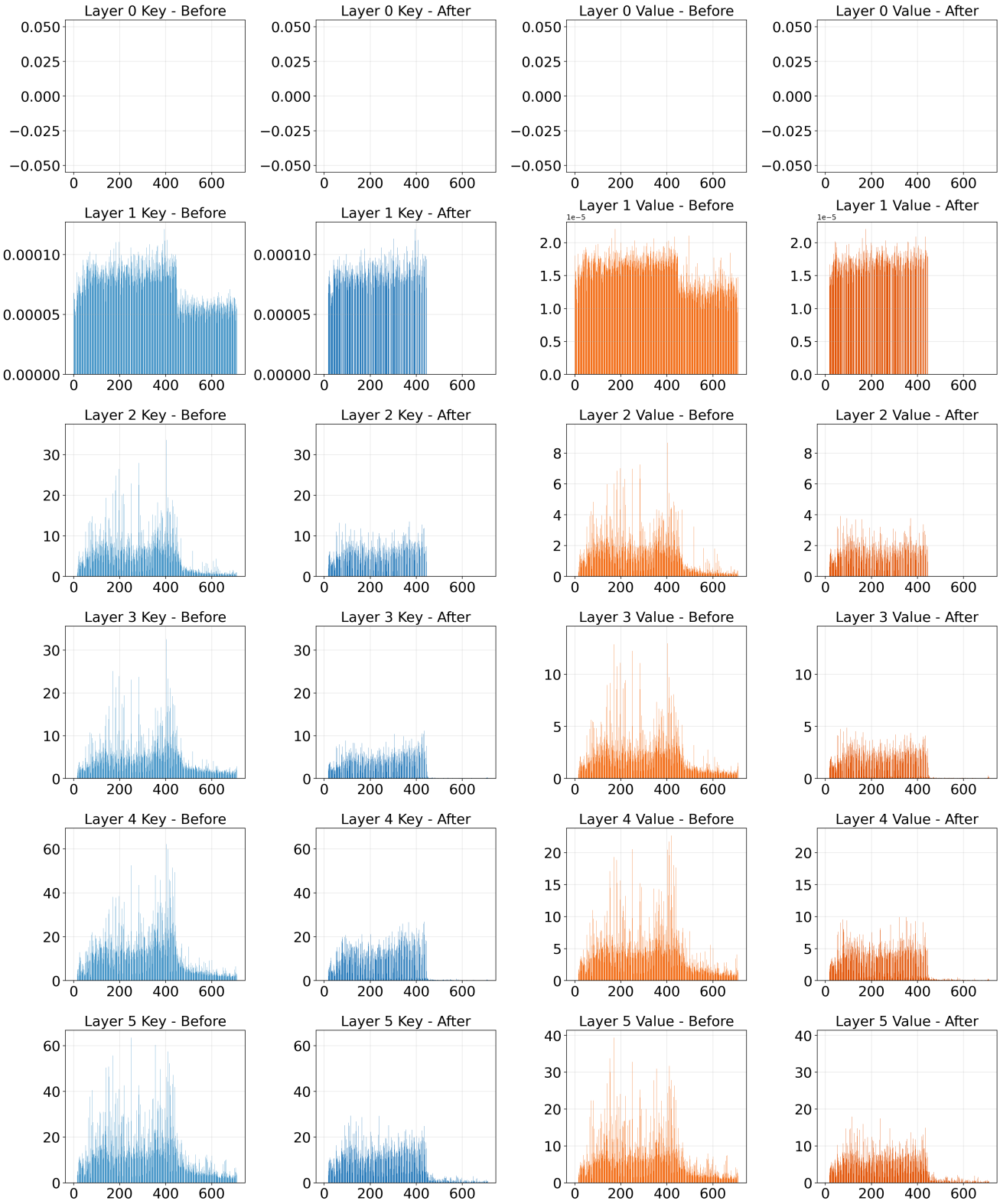

Figure 18. KV Deviation (Layer 0 – Layer 5)

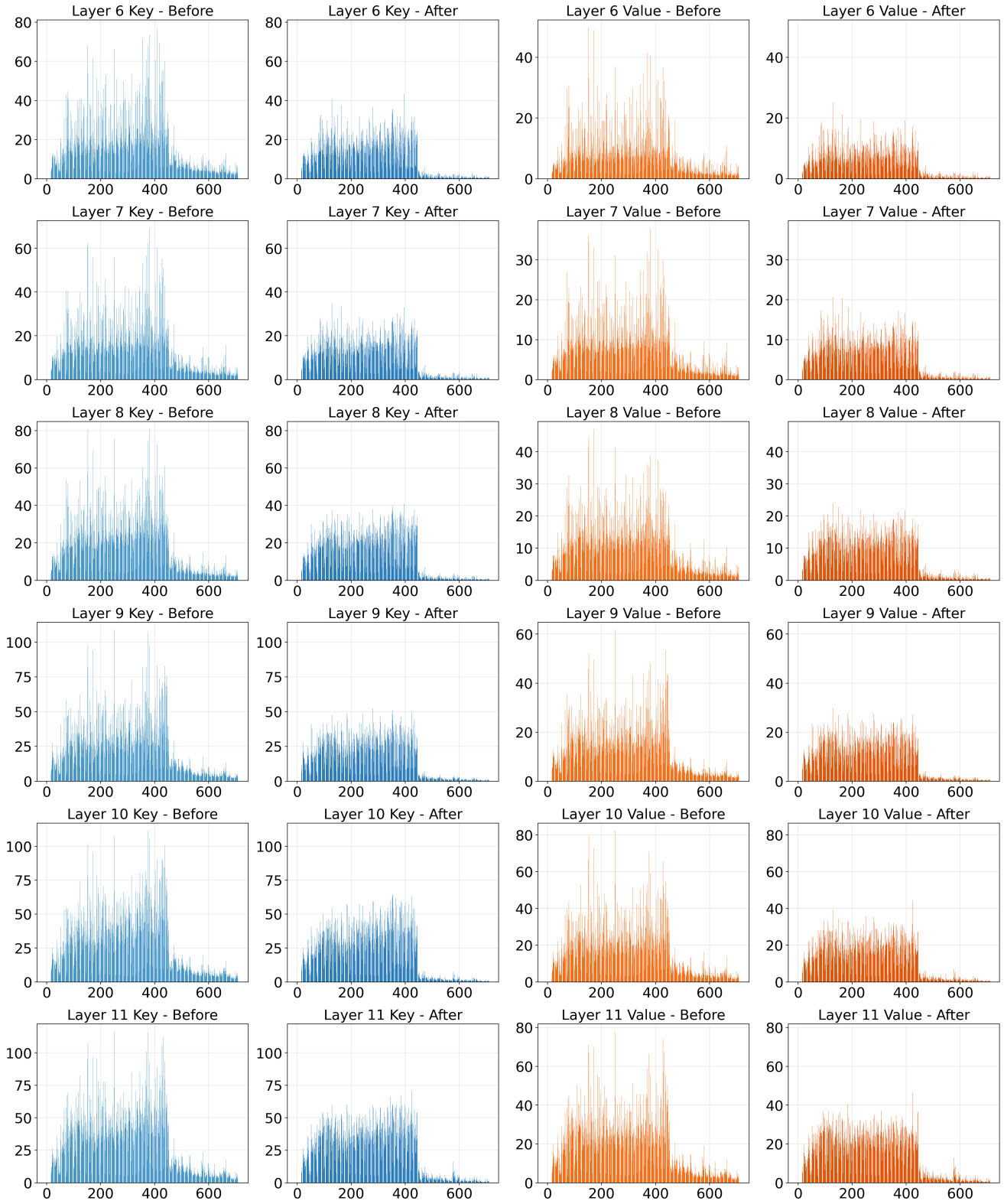

Figure 19. KV Deviation (Layer 6 – Layer 11)

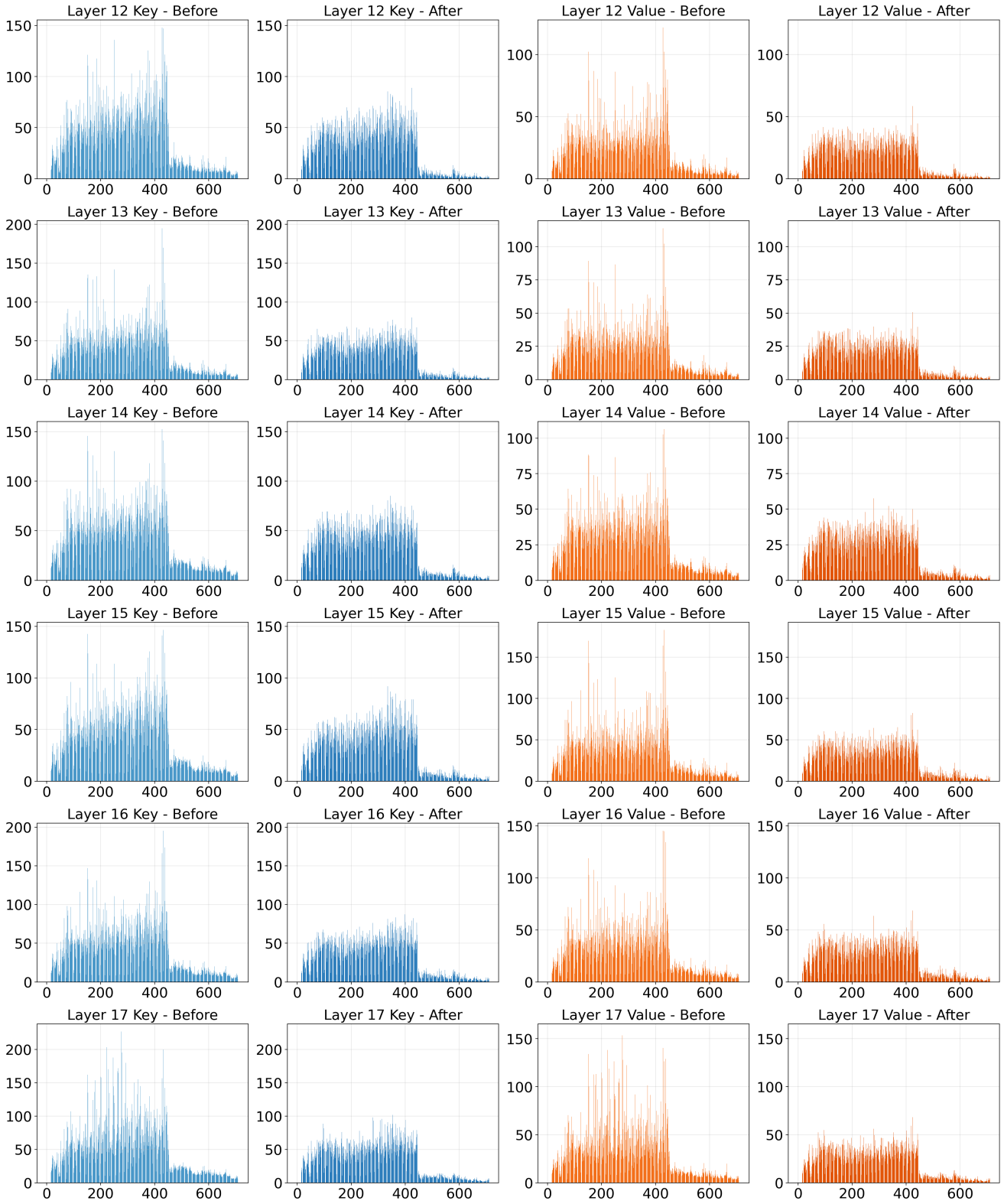

Figure 20. KV Deviation (Layer 12 – Layer 17)

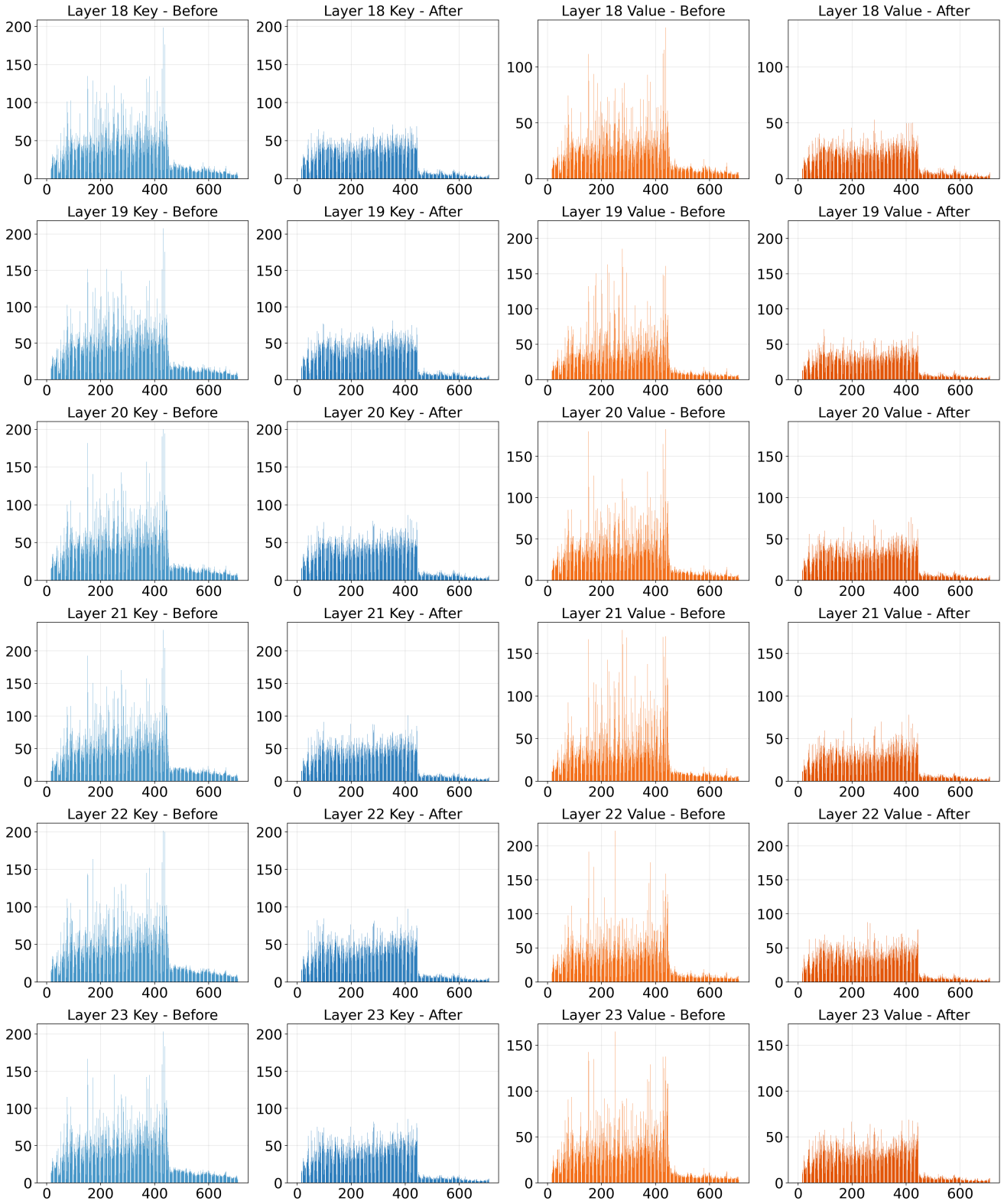

Figure 21. KV Deviation (Layer 18 – Layer 23)

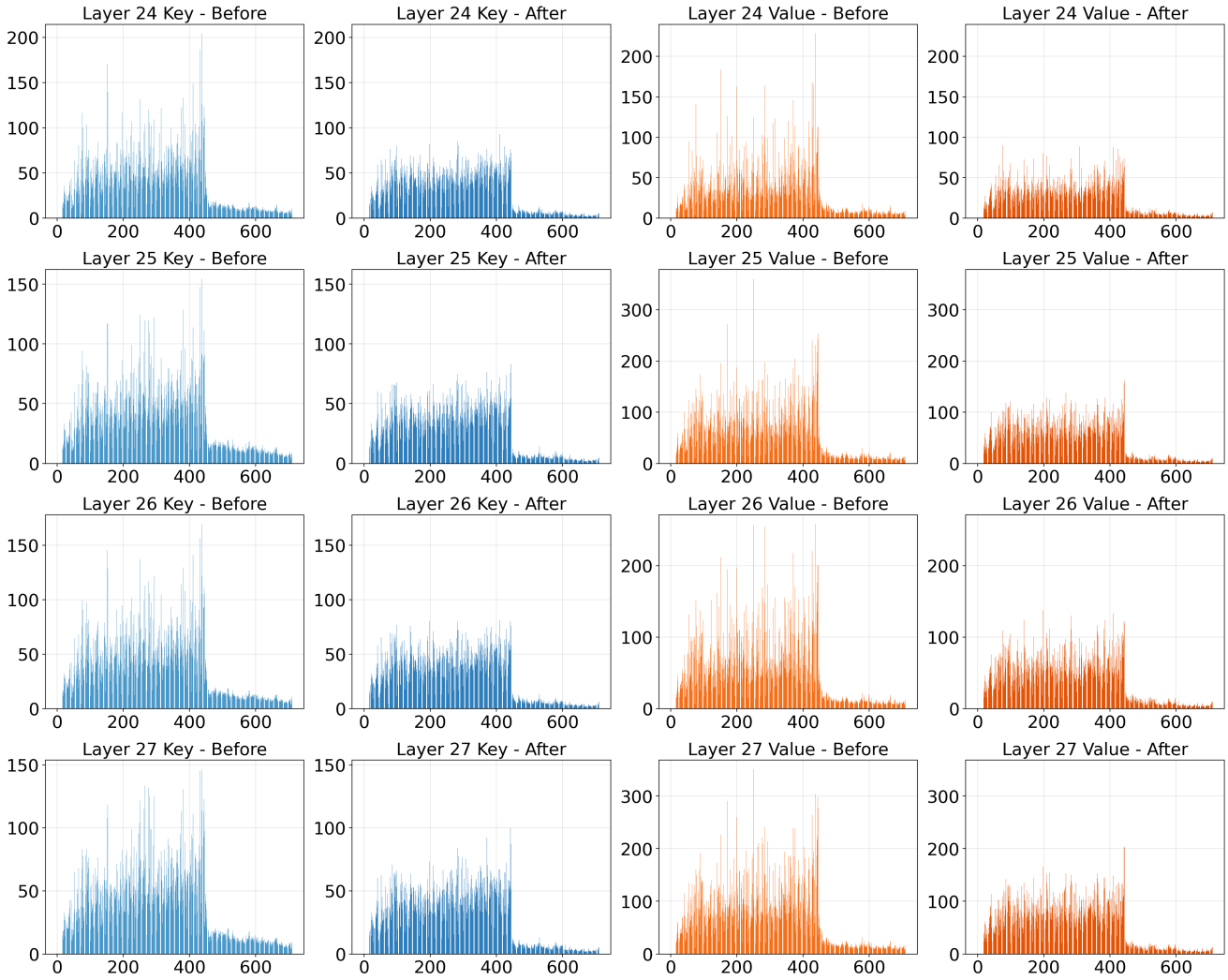

Figure 22. KV Deviation (Layer 24 – Layer 27)