Revisiting Multimodal KV Cache Compression: A Frequency-Domain-Guided Outlier-KV-Aware Approach

Supplementary Material

Outline for Supplementary Material

Due to the page limitation for the submission paper, we present additional details and visualization results from the following aspects:

- A: More Experiments in different KV Retention Ratios
- B: More Experiments on Decoding Stage KV Retention
- C: More Ablation Experiments on Low-Pass Cutoff Factor γ
- D: More Observations
  - D.1: Frequency-domain energy distribution of the KV matrices
  - D.2: Frequency-domain energy distribution of KV matrices across different layers
  - D.3: The impact on model performance of prioritizing the removal of KV pairs under different removal ratios
- E: More Efficiency Analysis
- F: Case Study
- G: Benchmarks
  - G.1 Multi-Images Understanding Benchmarks
  - G.2 High Resolution Benchmarks
  - G.3 Video Understanding Benchmark
- H: Models
  - H.1 Qwen2.5-VL-7B-Instruct
  - H.2 LLaVA-OneVision-1.5-8B-Instruct
- I: Limitation and Future Work

A. More Experiments in different KV Retention Ratios

In this section, we present the performance of various models on MileBench under lower KV Cache retention rates in Tab. 8. Task T represents Temporal Multi-Image task. Task S represents Semantic Multi-Image task. NH represents Needle in a Haystack task. IR represents Image Retrieval task. As shown in Tab. 8, FlashCache outperforms other methods on most tasks without requiring attention score recalculation. This demonstrates that we can identify and preserve important Key-Value pairs solely based on the distribution characteristics of the Key-Value matrices.

B. More Experiments on Decoding Stage KV Retention

Following established baselines (LOOK-M, MEDA), we focus on compressing prefill KV Cache while fully retaining decoding KVs. To validate robustness in long-generation scenarios, we evaluate *FlashCache* on the MathVista(COT) benchmark. As shown in Tab.9, *FlashCache* also maintains performance advantages in reasoning-heavy task.

C. More Ablation Experiments on Low-Pass Cutoff Factor γ

Additional ablation on LLaVA-OneVision-8B (Tab.10) aligns with our Qwen results, identifying $\gamma \in [0.1, 0.2]$ as optimal and robust. Theoretically, a lower γ creates smoother *Base KV* that forces the deviation term to capture critical high-frequency features. Conversely, a high γ causes *Base KV* to overfit, making outliers indistinguishable. Thus, $\gamma \in [0.1, 0.2]$ is a robust guideline across models and datasets.

D. More Observations

In this section, we further supplement our previous observations of low-frequency concentration in the KV matrices and outlier KV phenomena. Additional experiments across more models and datasets, along with visualizations, are conducted.

D.1. Frequency-domain energy distribution of the KV matrices

We observe that the frequency-domain energy of KV matrices is predominantly concentrated at low frequency, with high frequency components occupying a relatively small proportion. To further validate this phenomenon, we conduct additional experiments across two models—Qwen2.5-VL-7B-Instruct [37] and LLaVA-OneVision-1.5-8B-Instruct [4] using eight datasets. Fig. 7 presents the results for Qwen2.5-VL-7B-Instruct, while Fig. 8 shows the results for LLaVA-OneVision-1.5-8B-Instruct. Combining the findings from both figures, we observe that the phenomenon of KV matrices frequency-domain energy concentration in the low-frequency range is significantly established.

D.2. Frequency-domain energy distribution of KV matrices across different layers

We observe clear differences in the frequency-domain energy distribution of KV matrices across layers. To further validate this phenomenon, we conduct additional experiments on two models—Qwen2.5-VL-7B-Instruct [37] and LLaVA-OneVision-1.5-8B-Instruct [4]—using eight datasets. The results for Qwen2.5-VL-7B-Instruct are

Table 8. Performance of various KV cache strategies on several MLLMs on MileBench's tasks with KV Cache retention ratio $\rho = 0.1, 0.05$. The best results are highlighted in bold.

Method	LLaVA-OneVision-1.5-8B				Qwen2.5-VL-7B				Qwen2.5-VL-32B			
	Task T	Task S	NH	IR	Task T	Task S	NH	IR	Task T	Task S	NH	IR
Full Cache	46.97	65.33	27.03	11.67	55.59	69.17	27.35	14.17	56.51	65.21	27.19	18.17
$\rho = 0.1$												
StreamingLLM [41]	46.81	58.40	6.72	12.5	54.71	59.23	6.56	**15.83**	57	56.15	2.5	20.00
H2O [49]	46.84	58.51	15.94	11.83	54.58	58.65	6.25	14.83	56.94	58.33	2.5	21.0
SnapKV [20]	46.97	57.62	10.47	12.67	55.58	**61.30**	7.19	15.50	57.06	58.71	16.06	18.83
LOOK-M [34]	46.84	58.51	15.94	11.83	55.08	58.54	5.78	13.67	51.91	58.99	2.81	24.67
MEDA [35]	44.23	52.76	14.06	12.17	54.24	57.05	5.47	13.67	51.31	48.45	2.97	**28.67**
FlashCache	**47.09**	**58.56**	**20.00**	**12.83**	**55.65**	60.93	**18.45**	14.5	**57.23**	**59.08**	**16.10**	20.17
$\rho = 0.05$												
StreamingLLM [41]	46.52	57.05	5.16	12.00	49.46	53.64	5.63	15.00	51.94	42.34	1.25	26.83
H2O [49]	47.01	56.99	12.50	11.50	50.47	52.77	5.31	11.33	52.31	48.44	1.25	23.67
SnapKV [20]	45.97	**57.82**	7.97	13.00	51.52	**56.33**	6.25	15.33	55.69	49.07	6.25	19.83
LOOK-M [34]	**48.01**	56.99	12.50	11.50	51.69	53.67	5.31	**16.33**	49.97	45.77	2.5	**27.83**
MEDA [35]	38.40	48.25	10.31	13.00	51.74	52.95	5.16	15.33	51.86	50.01	2.65	26.33
FlashCache	45.51	51.12	**15.56**	**13.33**	**54.55**	56.19	**10.63**	14.83	**56.25**	**53.11**	**8.29**	26.33

Table 9. Comparison on MathVista (Qwen2.5-VL-7B, $\rho = 0.1$).

Method	Full Cache	StreamLLM	H2O	SnapKV	LOOK-M
MathVista	68.2	40.8	50.9	50.9	47.3
Method	MEDA	VLCache	Quest	PyramidKV	**FlashCache**
MathVista	51.6	22.9	22.9	21.1	**52.6**

shown in Fig. 9, and those for LLaVA-OneVision-1.5-8B-Instruct are presented in Fig. 10. Taken together, these results provide strong evidence that the frequency-domain energy distribution of KV matrices varies substantially across different layers.

D.3. The impact on model performance of prioritizing the removal of KV pairs under different removal ratios

In Fig.11, the frequency-domain outlier phenomenon is consistent across diverse model architectures, parameter scales, and datasets. This evidence substantiates that our core insight is inherently general rather than model/dataset-specific.

E. More Efficiency Analysis

In this section, we further analyze the efficiency of FlashCache. Tab. 12 reports the actual KV cache memory footprint and end-to-end inference time as the input length varies, under a 10% KV retention ratio. We measure the memory footprint and the end-to-end inference time for generating 512-token outputs to obtain per-token latency across input lengths 2K, 4K, 8K, 16K, 32K, 64K. All experiments are conducted with FlashAttention on Qwen2.5-VL-7B-Instruct. The results show that FlashCache substantially reduces both the KV cache memory footprint and the end-to-end inference time, with the acceleration effect becoming increasingly pronounced as the input length grows.

As shown in Tab.11, *FlashCache*'s overhead is negligible (12.45ms at 32K). Unlike baselines that suffer from OOM due to attention re-computation, *FlashCache* eliminates attention dependency, ensuring memory stability at 64K while introducing substantially lower prefilling overhead than existing methods.

F. Case Study

To further demonstrate the advantages of our approach, we have selected several reasoning examples, as shown in Fig. 12. We compare FlashCache against state-of-the-art KV cache eviction methods: StreamingLLM [41], H2O [49], SnapKV [20], LOOK-M [34], and MEDA [35]. We conduct comparisons across different image lengths, different text lengths, and different tasks. KV Cache retention ratio is setting as 0.05. Experimental results indicate that under high compression ratios, FlashCache preserves more of the model's inherent performance charac-

Table 10. Ablation Study on low-pass filter cutoff factor γ on LLaVA-OneVision-8B.

γ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
INIAH	**27.50**	24.69	18.44	14.06	11.88	11.25	10.00	10.31	10.00
GPR1200	12.33	**12.83**	11.50	12.17	11.67	11.83	12.33	12.33	12.17

Table 11. Method Latency in per layer with 32K/64K sequence length (metric: ms).

Method	Prefilling	H2O	SnapKV	LOOK-M	MEDA	FlashCache
32K	53.63	1297.27	76.62	1290.64	OOM	**12.45**
64K	148.08	OOM	OOM	OOM	OOM	**28.12**

Table 12. More Efficiency Analysis

Metric	Cache Size(GB)		Total Time(s)	
Method	Full KV	FlashCache	Full KV	FlashCache
2K	0.11	**0.01**	19.93	**19.54**
4K	0.21	**0.02**	20.53	**19.56**
8K	0.43	**0.03**	22.12	**19.62**
16K	0.85	**0.09**	24.57	**20.28**
32K	1.71	**0.17**	30.51	**21.75**
64K	3.42	**0.34**	43.5	**25.31**

teristics. At the same time, we find that FlashCache enables the model to answer questions it previously could not. We believe FlashCache filters out noise in some input data, thereby improving the model's performance under certain conditions.

G. Benchmarks

G.1. Multi-Images Understanding Benchmarks

MileBench[29]. MileBench is a multimodal long-context benchmark designed to evaluate how well Multimodal Large Language Models handle complex multi-image inputs over extended contexts. It is built from 6,440 long-context samples collected from 29 publicly available or self-constructed datasets, with each sample containing multiple images and accompanying text. In our paper, we use four task types from MileBench: Task T (Temporal Multi-Image), which focuses on understanding and reasoning over sequences of images with temporal relationships; Task S (Semantic Multi-Image), which requires integrating complementary semantic information distributed across multiple images; NH (Needle in a Haystack), which tests the model's ability to retrieve a small but crucial visual or textual detail from long, cluttered multimodal contexts; and IR (Image Retrieval), which assesses whether the model can accurately locate or identify the correct image given a textual description or multimodal query. Together, these tasks make MileBench a comprehensive tool for probing long-context multi-image comprehension, reasoning, and retrieval abilities.

MUIRBench[36]. MUIRBench is a comprehensive benchmark for robust multi-image understanding in multimodal large language models. It contains 11,264 images and 2,600 multiple-choice questions, covering 12 diverse multi-image understanding tasks (such as scene understanding, temporal ordering, and visual retrieval) and 10 categories of inter-image relations (including multiview, narrative, and complementary relations). Each example is constructed in a pairwise manner, where an answerable instance is paired with a minimally modified unanswerable variant, enabling robust evaluation of whether models truly integrate information across multiple images rather than relying on superficial cues. This design makes MUIRBench a strong stress test for multi-image reasoning and a useful complement to standard single-image vision–language benchmarks.

MMMU[46]. MMMU (Massive Multi-discipline Multimodal Understanding) is a large-scale benchmark designed to evaluate foundation models on expert-level, university and beyond problems across a wide range of real-world disciplines. It contains over 10,000 multimodal multiple-choice questions spanning fields such as medicine, law, physics, chemistry, engineering, and the humanities, with inputs that combine text, diagrams, charts, tables, and other images. Many questions involve not just a single image but multiple related figures or panels (for example several diagrams, tables, or subplots for one problem) that must be interpreted jointly, allowing the benchmark to test genuine multi-image reasoning rather than isolated image understanding. All items are carefully curated from real exam or professional materials, emphasizing high-level reasoning, domain knowledge, and cross-modal comprehension,

and are widely used to assess how close multimodal large language models are to human expert performance in complex, knowledge-intensive scenarios.

G.2. High Resolution Benchmarks

V*[39]. V* is an LLM-guided visual search mechanism designed to help multimodal large language models actively locate task-relevant regions in an image instead of passively encoding the whole frame. When paired with an MLLM, it underpins the SEAL ("Show, Search, and Tell") meta-architecture, which iteratively queries the image to refine visual grounding and reasoning. To evaluate this capability, the authors introduce V* Bench, a dedicated benchmark built from 191 high-resolution natural images (average resolution around 2246×1582) that emphasize fine-grained details and visually crowded scenes. V* Bench focuses on detailed visual grounding through attribute recognition and spatial relationship reasoning questions, making it particularly suitable for testing whether models can reliably handle high-resolution inputs and answer questions about small or easily overlooked visual elements.

HR-Bench[38]. HR-Bench is a high-resolution visual question answering benchmark designed to assess the fine-grained perception ability of multimodal large language models on ultra high-resolution images. It consists of two sub-tasks: Fine-grained Single-instance Perception (FSP), which contains 100 samples focusing on detailed attribute recognition, OCR, and visually grounded prompting within a single image, and Fine-grained Cross-instance Perception (FCP), which contains 100 samples targeting map understanding, chart analysis, and spatial relationship reasoning across multiple visual elements. HR-Bench is provided in both 8K and 4K settings: the 8K version uses images with approximately 8K-level resolution, while the 4K version is obtained by carefully annotating objects in the 8K images and cropping them into 4K regions, enabling rigorous evaluation of whether models can maintain and exploit fine details under realistic high-resolution conditions.

G.3. Video Understanding Benchmark

FAVOR-Bench[32]. FAVOR-Bench is a comprehensive benchmark for fine-grained video motion understanding that targets the ability of multimodal large language models to perceive and reason about detailed temporal dynamics rather than just coarse scene changes. It contains 1,776 videos from both ego-centric and third-person perspectives, each accompanied by structured manual motion annotations. On top of this data, FAVOR-Bench provides both close-ended and open-ended evaluation settings: for close-ended evaluation, it offers 8,184 multiple-choice question–answer pairs organized into six motion-centric sub-tasks, while for open-ended evaluation it includes a GPT-assisted caption assessment protocol together with a cost-efficient LLM-free evaluation method. Through these tasks, FAVOR-Bench enables systematic measurement of how well models recognize, localize, and describe fine-grained motions in realistic videos, and serves as a challenging testbed for improving motion-sensitive video understanding.

H. Models

H.1. Qwen2.5-VL-7B-Instruct

Qwen2.5-VL-7B-Instruct [37] is a 7-billion-parameter vision–language model from the Qwen2.5-VL family, instruction-tuned to handle a wide range of multimodal tasks with text, images, documents, charts, and video as input. It combines a strong language backbone with a visual encoder that supports dynamic resolutions and long-context processing, enabling capabilities such as document analysis, OCR, chart and layout understanding, temporal reasoning over videos, and structured output generation (for example JSON and bounding boxes for visual localization).

H.2. LLaVA-OneVision-1.5-8B-Instruct

LLaVA-OneVision-1.5-8B-Instruct [4] is an 8-billion-parameter large multimodal model in the LLaVA-OneVision-1.5 family, designed as a fully open, efficient framework for high-quality multimodal training. It is trained on native-resolution images and curated instruction data, aiming to deliver strong performance on diverse vision–language benchmarks while keeping computational costs relatively low. The model follows the LLaVA-style visual instruction tuning paradigm and is positioned as a general-purpose multimodal assistant capable of fine-grained image understanding, reasoning, and dialogue under an open-source training and evaluation pipeline.

I. Limitation and Future Work

In the future, we will continue to explore efficient compression for multimodal KV Cache without attention scores, striving to extend its application to more scenarios such as embodied intelligence and ultra-long context.

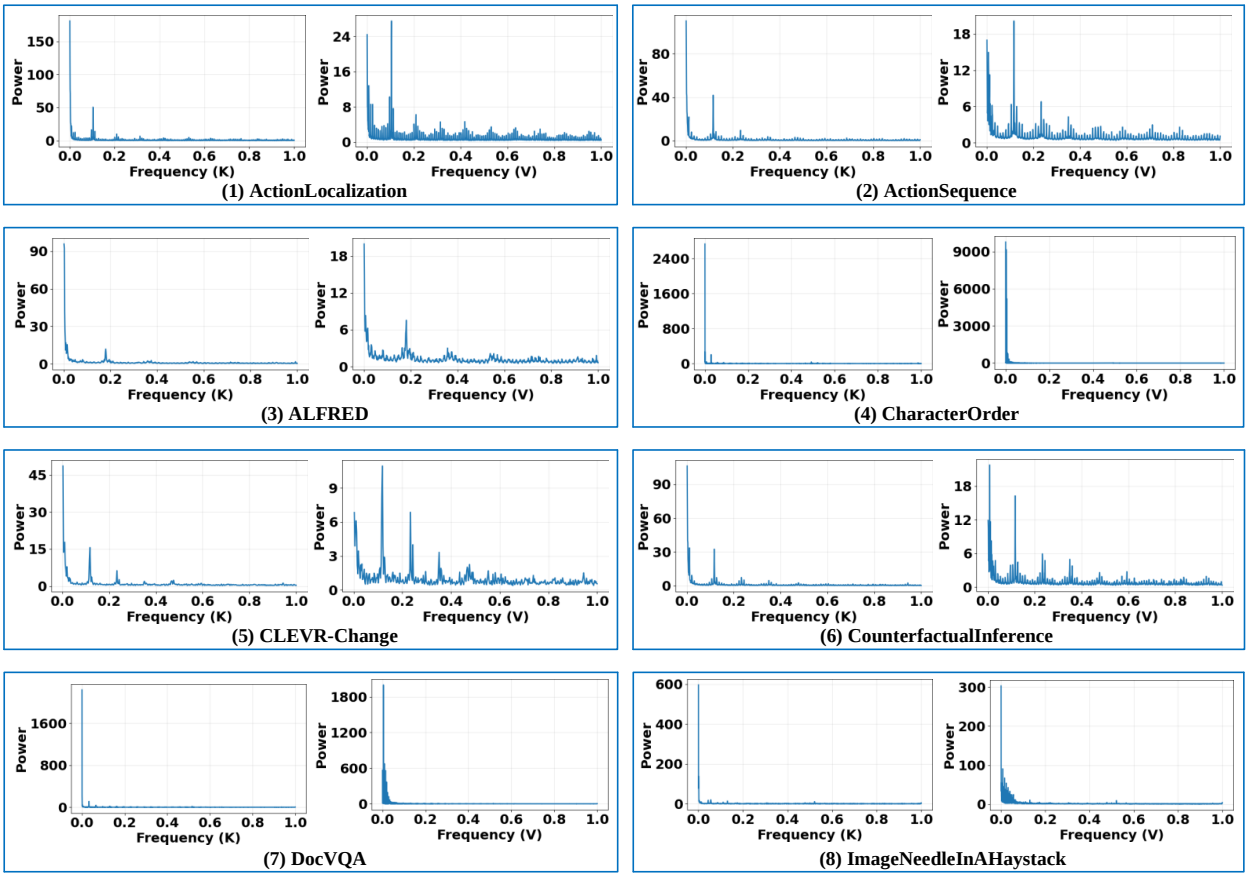

Figure 7. Frequency-domain energy distribution of the KV matrices. Experiments are conducted on Qwen2.5-VL-7B-Instruct.

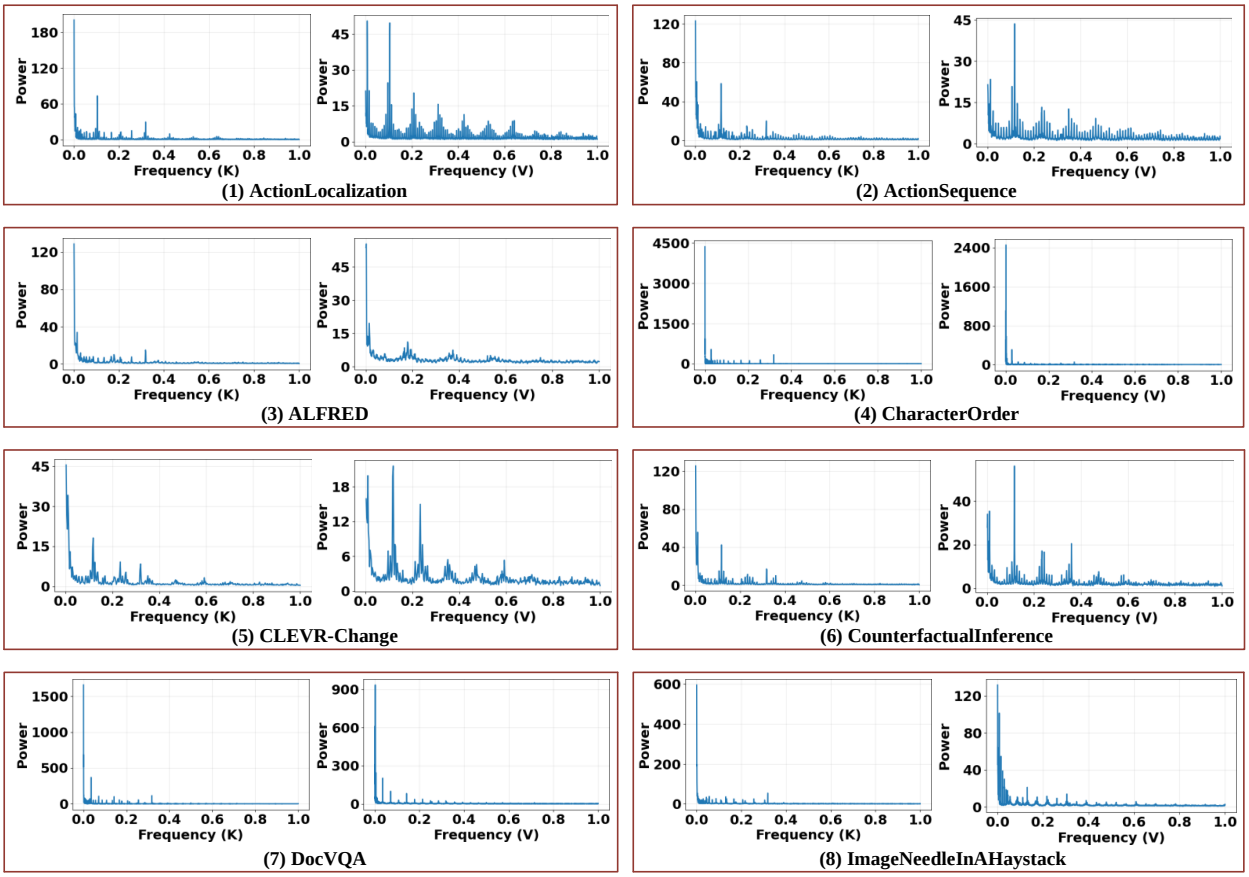

Figure 8. Frequency-domain energy distribution of the KV matrices. Experiments are conducted on LLaVA-OneVision-1.5-8B-Instruct.

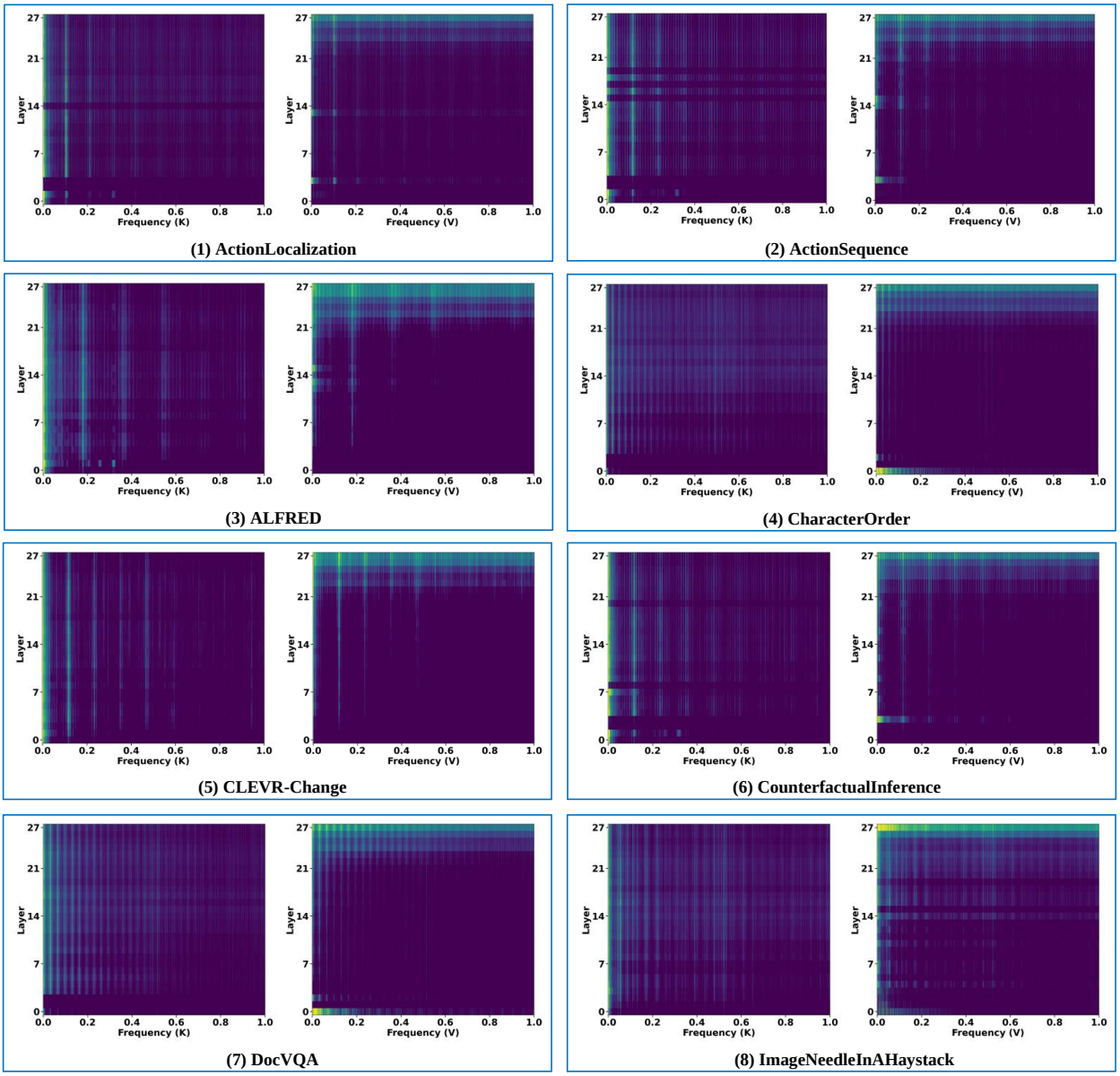

Figure 9. Frequency-domain energy distribution of KV matrices across different layers. Experiments are conducted on Qwen2.5-VL-7B-Instruct.

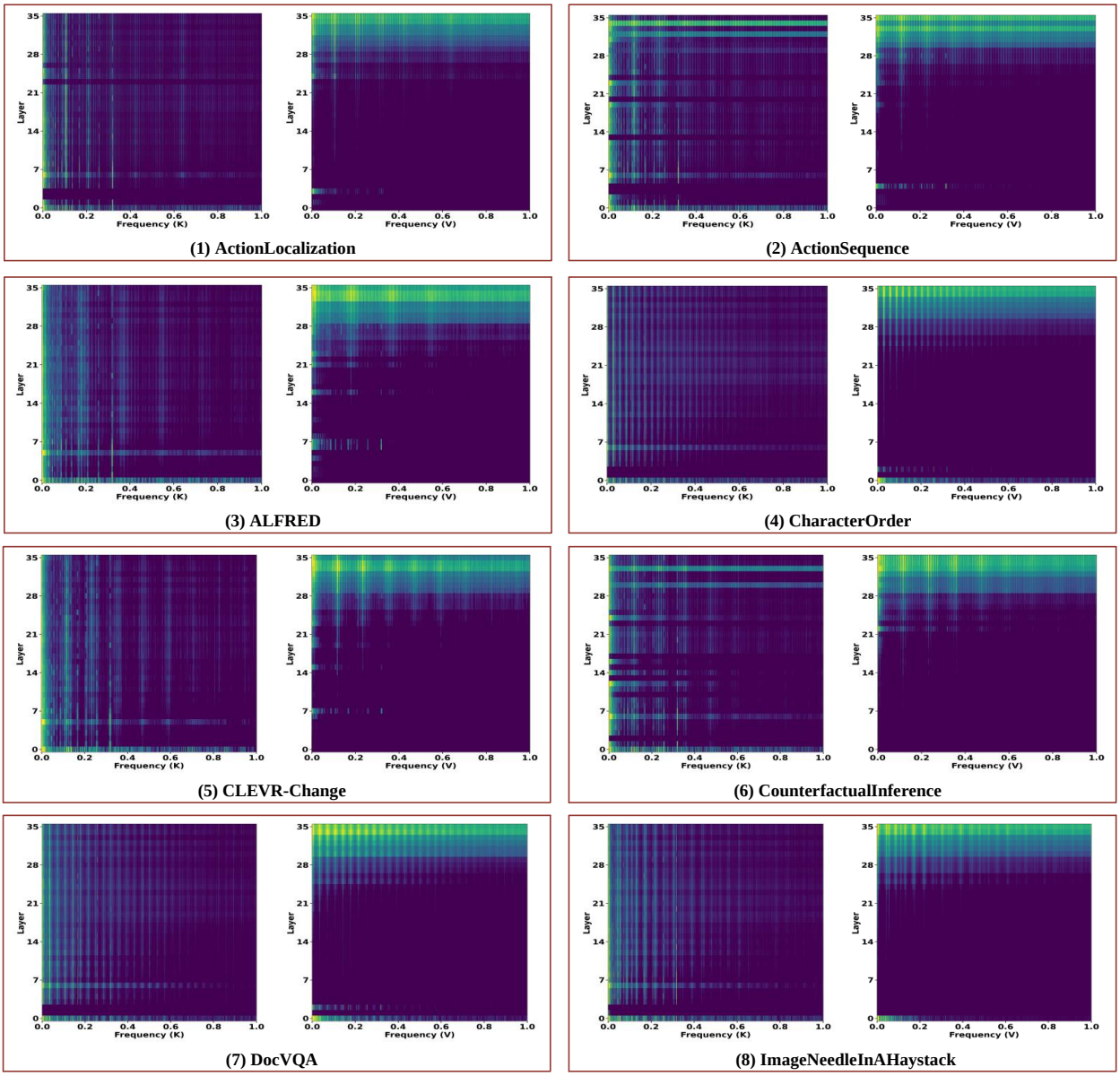

Figure 10. Frequency-domain energy distribution of KV matrices across different layers. Experiments are conducted on LLaVA-OneVision-1.5-8B-Instruct.

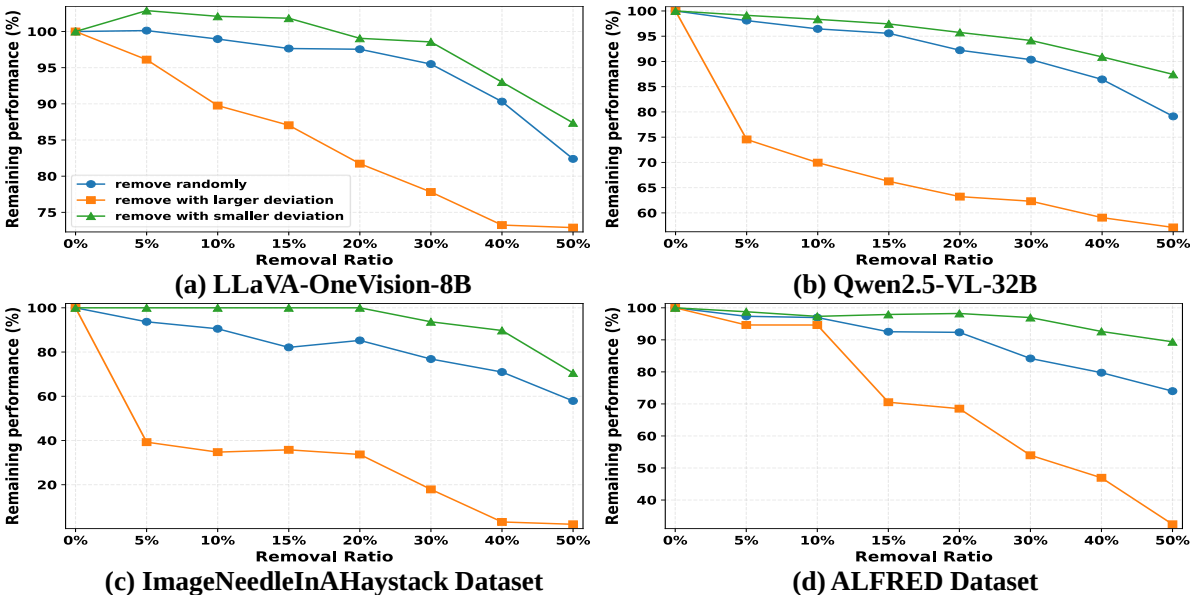

Figure 11. More observation of ”Outlier KV”.

Question: *What is the special magic San Francisco number in the images?* **Ground Truth:** *6687401*

Qwen2.5VL-7B: *The special magic San Francisco number is 6687401.* √

StreamingLLM: *The special magic San Francisco number is 668740.* ×
H2O: *The special magic San Francisco number is 666.* ×
SnapKV: *The special magic San Francisco number is 6874321.* ×
LOOK-M: *The special magic San Francisco number in the images is 666.* ×
MEDA: *The special magic San Francisco number in the images is 6666666666666666666.* ×
FlashCache: *The special magic San Francisco number is 6687401.* √

Question: *Given a primary objective and your current surroundings, use your previous decisions and perspectives to determine your next move. Your Main Goal: Heat a loaf of bread, put it on the counter. Step Details: Step#1: Move to the white table. Step#2: Pick up the bread in front of the tomato. Step#3: Bring the bread to the microwave. Current Step: "*

Ground Truth: *Heat the bread in the microwave.*

Qwen2.5VL-7B: *Put the bread in the microwave.* √

StreamingLLM: *Put the food into the microwave.* ×
H2O: *Put the bread into the oven.* ×
SnapKV: *Put the bread into the oven.* ×
LOOK-M: *Put the microwave on the counter.* ×
MEDA: *Put the bottle on the table.* ×
FlashCache: *Put the bread in the microwave.* √

Question: *What has changed from the first image to the second?*

Ground Truth: *the brown metallic cylinder is missing.*

Qwen2.5VL-7B: *the small blue metal cylinder that is behind the large purple thing changed its location.* ×

StreamingLLM: *the black object in the lower right corner of the image is no longer there.* ×
H2O: *the blue cylinder is in a different location.* ×
SnapKV: *the gray object in front of the white object is no longer there.* ×
LOOK-M: *the blue cylinder is no longer there.* ×
MEDA: *the blue cylinder is no longer there.* ×
FlashCache: *the brown metallic object is missing in the second image.* √

Question: *Upon reviewing several images, your task is to extract the necessary information and answer the given question. You must choose your answer from the Choice List. Image Caption #1: Phallus-impudicus-stinkmorchel-b Common Stinkhorn, Phallus impudicus, Family: Phallaceae, Location: Southern Germany, Ulm Image Caption #2: Brown-gilled Mushrooms - Possibly Orange pholiota. - geograph.org.uk - 279281 Brown-gilled Mushrooms - Possibly Orange pholiota. A very common mushroom found in dense clusters at the base of stumps or old trunks of decideous trees. Question: Between the Common Stinkhorn and Brown-gilled Mushrooms, which of them have a stalk overall thicker than its cap? Choice list: A. The Brown-gilled Mushroom has a cap that is, overall, thicker than its stalk. B. The Common Stinkhorn and Brown-gilled Mushrooms have a stalk and cap that are roughly the same thickness. C. The Common Stinkhorn has a stalk that is, overall, thicker than its cap. D. The Brown-gilled Mushroom has a stalk that is, overall, thicker than its cap. Your answer is:*

Ground Truth: *The Common Stinkhorn has a stalk that is, overall, thicker than its cap.*

Qwen2.5VL-7B: *C. The Common Stinkhorn has a stalk that is, overall, thicker than its cap.* √

StreamingLLM: *C. The Brown-gilled Mushroom has a stalk that is, overall, thicker than its cap.* ×
H2O: *C* ×
SnapKV: *C. The Brown-gilled Mushroom has a stalk that is, overall, thicker than its cap.* ×
LOOK-M: *C. The Brown-gilled Mushroom has a stalk that is, overall, thicker than its cap.* ×
MEDA: *C. The Brown-gilled Mushroom has a stalk that is, overall, thicker than its cap.* ×
FlashCache: *C. The Common Stinkhorn has a stalk that is, overall, thicker than its cap.* √

Figure 12. Case Study