OpenMOSS

HERMES: KV Cache as Hierarchical Memory for Efficient Streaming Video Understanding

Haowei Zhang[1,*] **Shudong Yang**[1,2,*] **Jinlan Fu**[1,3,†] **See-Kiong Ng**[3] **Xipeng Qiu**[1,2]

[1]Fudan University, [2]Shanghai Innovation Institute, [3]National University of Singapore

Abstract

Recent advancements in Multimodal Large Language Models (MLLMs) have demonstrated significant improvement in offline video understanding. However, extending these capabilities to streaming video inputs, remains challenging, as existing models struggle to simultaneously maintain stable understanding performance, real-time responses, and low GPU memory overhead. To address this challenge, we propose ***HERMES***, a novel training-free architecture for real-time and accurate understanding of video streams. Based on a mechanistic attention investigation, we conceptualize KV cache as a hierarchical memory framework that encapsulates video information across multiple granularities. During inference, ***HERMES*** reuses a compact KV cache, enabling efficient streaming understanding under resource constraints. Notably, ***HERMES*** requires no auxiliary computations upon the arrival of user queries, thereby guaranteeing real-time responses for continuous video stream interactions, which achieves 10× faster TTFT compared to prior SOTA. Even when reducing video tokens by up to 68% compared with uniform sampling, ***HERMES*** achieves superior or comparable accuracy across all benchmarks, with up to 11.4% gains on streaming datasets.

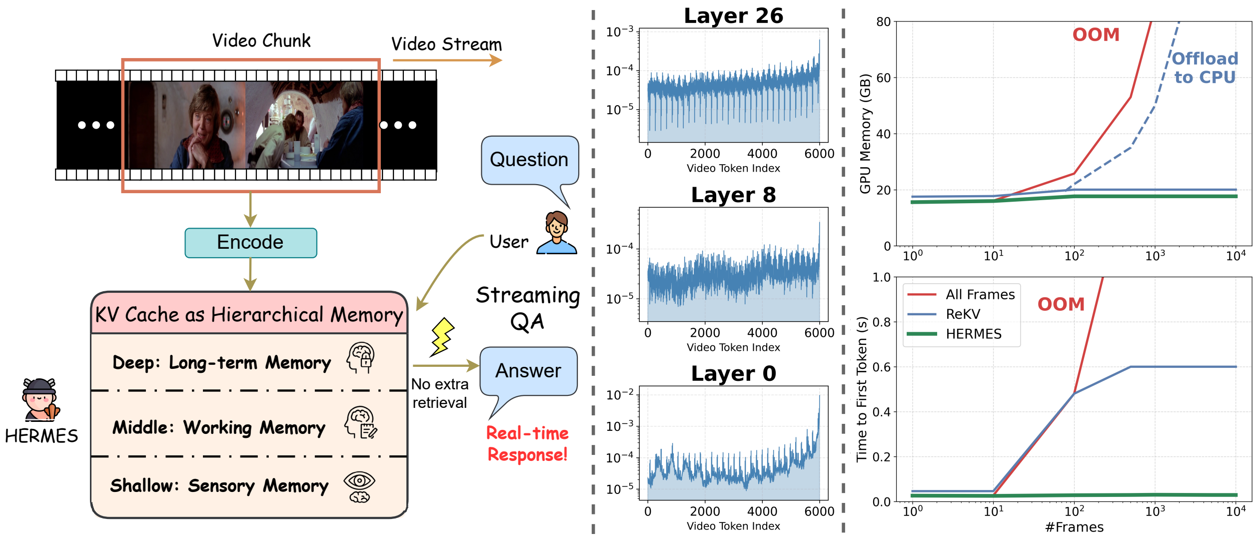
Correspondence: hwzhang25@m.fudan.edu.cn, jinlanjonna@gmail.com
Homepage: https://hermes-streaming.github.io/
Repository: https://github.com/haowei-freesky/HERMES

1 Introduction

Recent years have witnessed remarkable evolution in the capabilities of Multimodal Large Language Models (MLLMs) in video understanding tasks [4, 12, 23]. Despite the progress, the rapid emergence of real-time applications demands stable long video understanding, low-latency response, and memory-efficient deployment. However, existing MLLMs struggle to simultaneously satisfy these requirements on streaming videos. Notably, TimeChat-Online [50] observes that a large number of streaming video tokens are redundant, motivating compression methods to address these challenges. While numerous compression techniques have been proposed for offline videos [40, 44, 48], most are ill-suited for memory management in streaming scenarios, as streaming inputs are unpredictable in future frames and queries.

To adapt to streaming inputs, recent research introduces specialized memory management techniques, which

*Equal contribution.
†Corresponding author.

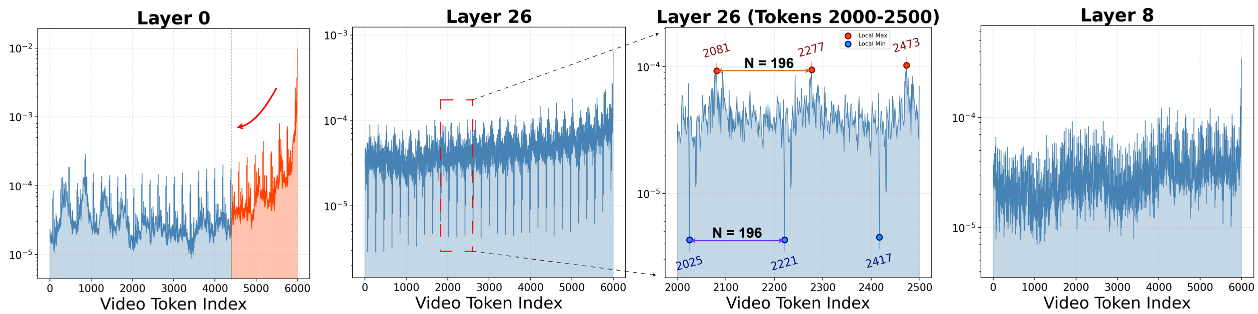


(a) ***HERMES*** **Framework** **(b)** Attention Analysis **(c)** Efficiency Test

Figure 1 **Left**: ***HERMES*** is a training-free approach for efficient streaming video understanding, enabling stable inference by reusing KV cache and performing hierarchical management of video tokens stored in KV cache. **Middle**: ***HERMES*** is based on a mechanistic investigation of the layer-wise attention preferences over hierarchical video information. **Right**: We evaluate LLaVA-OV-7B on a single A800 GPU (80 GB). As input frames increase, ***HERMES*** consistently maintains extremely low latency (TTFT < 30 ms) and stable GPU memory consumption, exhibiting no risk of OOM errors and requiring no auxiliary external computational resources.

generally fall into two paradigms: external memory and internal memory. External memory methods store video content as captions or raw vision patches in databases, and perform ad-hoc retrieval and multimodal prefilling at query time [45, 47], suffering from high latency and a lack of end-to-end cohesion. Additionally, many of these methods necessitate costly model-specific training [41, 46, 51]. In contrast, internalizing memory directly into the key-value cache (KV cache) remains underexplored, yet is crucial for low-latency responses and seamless end-to-end reasoning over stored video contexts. Moreover, KV cache naturally acts as a latent, model-intrinsic memory [19] that frequently interacts with the video stream, making it particularly suitable for training-free memory management. ReKV [13] and LiveVLM [31] are representative training-free, cache-based methods for streaming memory management. They store previous video segments in external CPU or disk and need to perform an additional retrieval when a user query arrives, which still rely on external computational resources and leads to significant latency. StreamMem [49] leverages chat template tokens to guide compression but lacks fine-grained KV management and mechanistic interpretability.

To overcome the aforementioned limitations of existing streaming video methods, we propose ***HERMES*** (KV Cache as **H**i**ER**archical **M**emory for **E**fficient **S**treaming Video Understanding), a training-free and plug-and-play approach that can be seamlessly integrated into existing MLLMs. Grounded in a mechanistic investigation of layer-wise attention shown in Fig. 1b, we conceptualize KV cache as a hierarchical memory framework that stores video information across multiple levels of granularity: shallow layers function as sensory memory, exhibiting a strong recency bias toward newly arriving frames; deep layers act as long-term memory, focusing on frame-level rhythmic anchor tokens; and middle layers serve as transitional working memory that balances recency information with frame-level semantic representations. Our method ***HERMES*** comprises three components: hierarchical KV cache management, cross-layer memory smoothing, and position re-indexing. During inference, ***HERMES*** reuses the compact KV cache and requires no auxiliary computations or external devices upon the arrival of user queries, thereby guaranteeing real-time responses. Experiments show that ***HERMES*** maintains stable and accurate performance with up to 68% fewer video tokens, while maintaining consistently low response latency and a constant GPU memory footprint.



(a) Shallow layer attention. **(b)** Deep layer attention. **(c)** Middle layer attention.

Figure 2 Visualization of the average attention weights (log scale) for user queries over video tokens in LLaVA-OV-7B with a FIFO KV cache budget of 6K video tokens per layer, averaged across 300 user video questions.

To summarize, our main contributions are as follows:

1. Grounded in a mechanistic analysis on attention visualization, we pioneer the conceptualization of KV cache as a hierarchical video memory framework across multiple granularities.
2. We propose ***HERMES***, a training-free method for streaming video understanding by reusing hierarchically managed KV cache. Despite reducing video tokens by up to 68%, ***HERMES*** achieves competitive accuracy, with gains of up to 11.4% on streaming benchmarks.
3. ***HERMES*** exhibits outstanding efficiency in streaming scenarios. Compared to the prior training-free SOTA method, it achieves up to a 10× speedup in latency. With a constant, compact GPU memory footprint and no auxiliary computation at query time, ***HERMES*** ensures consistently low-latency responses.

2 Layer-wise Preference for Hierarchical Streaming Video Information

Sliding Window is a standard paradigm for streaming video processing by incrementally encoding the continuous video stream chunk by chunk. When KV cache reaches the pre-defined memory budget, token eviction is triggered, and deciding which tokens to keep is crucial for stable understanding. Existing methods [13, 46, 49] rely on coarse-grained eviction strategies such as FIFO uniformly across all layers, overlooking layer-wise attention preferences.

To fill this gap, we conduct a mechanistic investigation of attention preferences in MLLM decoder layers, revealing how layers specialize in storing multiple-granularity video memory. To derive generalized insights, we randomly sample 100 video-question pairs from each of the short (62s[3] - 141s), medium (251s - 1,092s) and long (1,795s - 3,579s) duration subsets of the VideoMME benchmark [16] to cover diverse video durations and user queries. The video samples are uniformly sampled at 0.5 fps and subsequently fed into LLaVA-OV-7B in a streaming chunk-wise manner, with each chunk containing 8 frames. LLaVA-OV-7B consists of 28 decoder layers, and each video frame is uniformly encoded into 196 visual tokens. During the prefilling stage for video tokens, we maintain a constant budget $|M|$ of 6K video tokens per KV cache layer. After each eviction step, the positional indices of tokens per KV cache layer are re-indexing to contiguous $[0, |M|)$.

Layer-wise attention visualizations over video tokens maintained in a FIFO KV cache in Fig. 2 reveal three general stages of attention preference, along with more visualization results presented in App. A:

- **Shallow Layers as Sensory Memory**: As shown in Fig. 2a, the shallow layers (e.g., layer 0) exhibit an intense recency bias, with attention sharply concentrated on the most recent visual tokens and rapidly decaying over earlier ones. This behavior aligns with the concept of Sensory Memory [2, 37]: shallow

[3]To ensure the sliding window contains 6,000 tokens, a video at 0.5 fps for LLaVA-OV must have a duration of at least $6,000/196/0.5 \approx 62s$.

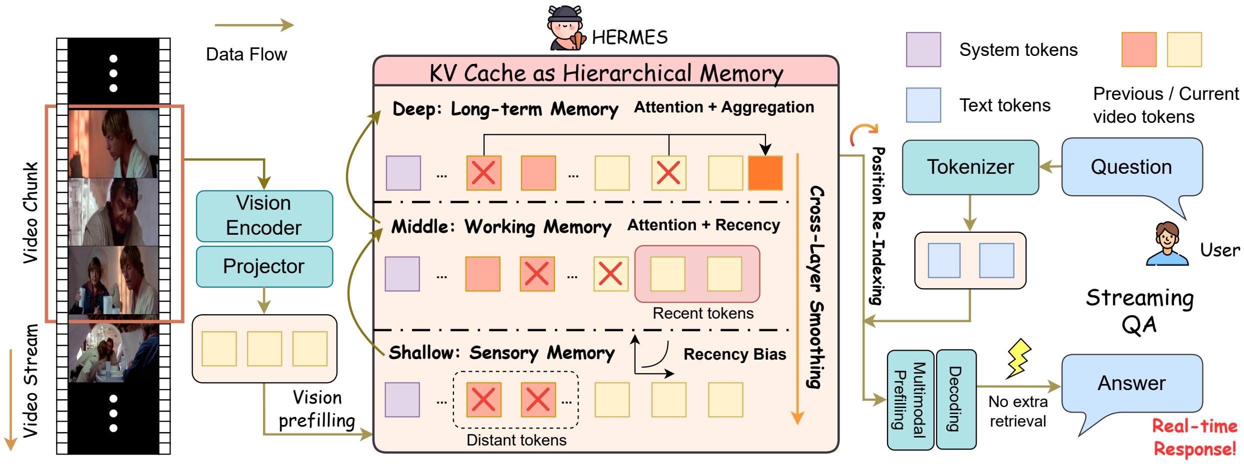


Figure 3 Overview of the ***HERMES*** architecture for streaming video QA. By implementing a hierarchical KV cache and specialized management strategies, ***HERMES*** enables real-time and accurate responses through direct cache reuse, eliminating the need for additional retrieval operations or external memory whenever users pose questions.

layers function as a short-lived buffer for the most recent visual inputs, enabling the model to quickly perceive incoming frames.

- **Deep Layers as Long-term Memory**: In deep layers (e.g., layer 26 in Fig. 2b), recency bias largely disappears. Instead, the attention pattern becomes highly sparse and rhythmic, with local extrema appearing at regular intervals. These extrema are exactly N = 196 tokens apart, matching to the number of tokens encoding a single frame in LLaVA-OV-7B. These local maxima can be regarded as frame-level "anchor tokens", summarizing the visual information of each frame. This pattern reflects Long-term Memory [2, 37]: deep layers store critical frame-level semantic representations for long-horizon understanding.
- **Middle Layers as Working Memory**: Middle layers (e.g., layer 8 in Fig. 2c) exhibit a gradual reduction in recency bias, with attention more evenly distributed across recent and earlier tokens. Simultaneously, the attention begins to transition toward the rhythmic patterns in the deep layers. This behavior corresponds to Working Memory [3, 19]: middle layers integrate recent and earlier visual information, bridging short-term sensory traces with frame-level semantic summaries.

3 HERMES

We propose ***HERMES***, a training-free framework that can be seamlessly integrated with MLLMs. As shown in Fig. 3, ***HERMES*** has three components: hierarchical KV cache management, cross-layer memory smoothing, and position re-indexing.

3.1 Hierarchical KV Cache Management

Motivated by the layer-wise attention patterns identified in Sec. 2, we design a hierarchical KV cache strategy. For each video token with KV cache index i at layer l, where i denotes its physical position in KV cache, we compute an importance score S_i^l to decide its retention:

- **Shallow Layers**: They act as sensory memory with strong recency bias. Inspired by Ebbinghaus' memory decay theory [14], we model token importance using an exponential forgetting curve based on temporal distance:

$$S_i^l = \alpha_i^l \cdot e^{-k\Delta t_i}, \Delta t_i = T - 1 - i, \tag{1}$$

where T is the total number of video tokens in the cache, $k > 0$ is the forgetting rate, α_i^l denotes the normalization factor.

- **Deep Layers**: Deep layers function as frame-level long-term memory with stable anchor tokens. Their attention distributions are sparse, and these anchor tokens consistently receive high attention across frames, making attention magnitude a reliable indicator of long-term importance. We therefore compute token importance directly from attention weights with respect to the user query. To handle unpredictable queries in streaming scenarios, we use a generic guidance prompt (see App. B) as a pseudo query. Token importance is computed as:
$$S_i^l = \alpha_i^l \cdot W_i^l, \tag{2}$$
where W_i^l denotes the attention weight of the i-th token at the layer l.

- **Middle Layers**: Middle layers serve as working memory, transitioning from recency-dominated shallow layers to attention-driven deep layers. We compute importance by interpolating recency and attention with a layer-dependent weight:
$$\omega^l = \omega_0 - \gamma \cdot \frac{l - l_{\text{short}}}{l_{\text{long}} - l_{\text{short}}}, \tag{3}$$
where l_{short} and l_{long} denote the layer indices, with $\omega_0 = 0.75$ and $\gamma = 0.6$. The importance score of token i at layer l is then computed as
$$S_i^l = (1 - \omega^l)\, A_i^l + \omega^l\, R_i^l, \tag{4}$$
where A_i^l and R_i^l denote the normalized attention weight and recency score, respectively, computed as in Eqs. (1) and (2).

3.2 Cross-Layer Memory Smoothing

Hierarchical KV cache management may introduce cross-layer inconsistency, as tokens at the same cache index can be evicted independently across layers, leading to misaligned visual memory. Since effective LLM memory relies on cross-layer interaction [6, 19, 33, 39], we address this issue with Cross-Layer Memory Smoothing.

Instead of treating video tokens at the same KV cache index as independent across layers, we propagate and smooth importance signals from deeper to shallower layers. Given raw importance scores S_i^l, the smoothed score is computed as:
$$\tilde{S}_i^l = (1 - \lambda_l) \cdot S_i^l + \lambda_l \cdot S_i^{l+1}, \tag{5}$$
$\lambda \in [0, 1]$ is the smoothing hyperparameter that controls the strength of cross-layer smoothing.

We then apply Top-K selection based on $\tilde{S}_i^l$ to maintain a fixed memory budget $|M|$ per layer:
$$\begin{aligned} \mathcal{I}_l &= \text{TopK}(\tilde{S}_l, |M|), \\ K_l &= K_l[\mathcal{I}_l], \quad V_l = V_l[\mathcal{I}_l]. \end{aligned} \tag{6}$$

To preserve long-term information, evicted tokens are aggregated into a **summary token** per layer, which compactly encodes long-term memory and is retained in the KV cache (see App. F).

3.3 Position Re-Indexing

Continuous accumulation of streaming inputs causes positional indices to exceed the model's maximum supported range, severely degrading text generation quality. To stabilize inference, we apply position re-indexing, which remaps positional indices to a contiguous range $[0, |M|)$ within the memory budget $|M|$. We design two strategies:

Lazy Re-Indexing Re-indexing is triggered only when positional indices approach the model limit, resulting in lower computational overhead. By preserving the original positional indices of recent tokens, it prevents positional drift compared to eager re-indexing, making it well suited for streaming video understanding.

Eager Re-Indexing Re-indexing is performed at each compression step, maintaining strictly contiguous RoPE indices in KV cache. While this strategy stabilizes long-range visual semantics [21, 22, 46], it leads to higher computational cost due to frequent re-indexing, making it more suitable for offline videos.

The details of re-indexing implementation for 1D RoPE (LLaVA-OV) and 3D M-RoPE (Qwen2.5-VL) are illustrated in App. E.1 and App. E.2, respectively.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate ***HERMES*** on diverse streaming and offline benchmarks. For streaming understanding, we use StreamingBench [27], OVO-Bench [25] and RVS (including RVS-Ego and EVS-Movie) [53]. For offline video evaluation, we adopt one short video dataset MVBench [24], along with two long video datasets, VideoMME [16] and Egoschema [30]. We conduct evaluation on the official dev split of Egoschema and report VideoMME results without subtitles. Our benchmark selection covers both multiple-choice and open-ended questions as QA form. The details of utilized benchmarks are demonstrated in App. D.

Models. To further verify the broad applicability of our method, we select two popular open-source MLLM series, LLaVA-OneVision (LLaVA-OV) [23] and Qwen2.5-VL [5]. Each is tested across two different parameter scales, covering a large range from 0.5B to 32B. For Qwen2.5-VL, we maintain its native dynamic resolution on video input, ensuring a fair comparison with the base model.

Implementation Details. For evaluating ***HERMES*** across all benchmarks, each video is encoded and processed chunk by chunk, with 16 frames per chunk, and sequentially prefilling the backbone LLM. Then, token compression is triggered once the predefined memory budget is exceeded.

For the layer partition, we follow the mechanistic investigations presented in Sec. 2: 10% shallow, 60% middle and 30% deep layers. A more comprehensive analysis of attention behaviors as supportive evidence can be found in Fig. 6. The cross-layer memory smoothing hyperparameter λ proposed in Sec. 3.2 is layer-dependent, with detailed configurations reported in App. C.

All evaluations are conducted using FP16 mixed precision and efficiency tests are conducted on a single A800 GPU, consistent with prior works [8, 13]. Greedy decoding is used to generate deterministic outputs. Accuracy evaluations can be completed on one H200 GPU.

4.2 Main Results

Streaming Video Understanding Extensive experiments on streaming benchmarks reveal the key findings:

(1) ***HERMES*** *outperforms on multiple-choice streaming datasets, showing exceptional real-time understanding and backward tracing capabilities*. As shown in Tab. 1, it achieves state-of-the-art performance on StreamingBench and OVO-Bench, significantly surpassing base models and training-free baselines. Built on Qwen2.5-VL-7B, ***HERMES*** reaches 79.44% and 59.21% accuracy using only 4K video tokens, improving over Qwen2.5-VL-7B by 6.13% and 6.93%, while outperforming all 7B-scale open-source online and offline models. Full results on StreamingBench and OVO-Bench are shown in Tab. 11 and Tab. 12 respectively.

(2) ***HERMES*** *excels on open-ended streaming tasks, showing fine-grained temporal and spatial comprehension*. On RVS-Ego and RVS-Movie (Tab. 2), we evaluate the model answer by GPT-3.5-turbo-0125 on accuracy and score (1–5 scale), consistent with compared baselines. ***HERMES*** consistently surpasses all prior training-free methods and improves accuracy by up to 11.4% over the base model with uniformly sampled 64 frames. These extensive experiments demonstrate ***HERMES***'s strong abilities in various streaming tasks, as well as its general applicability across foundation models. Moreover, we provide case studies from RVS benchmark,

Model	#Frames	StreamingBench Real-Time	OVO-Bench Real-Time	OVO-Bench Backward	Avg.
Human	-	91.46	93.20	92.33	92.83
Proprietary MLLMs					
Gemini 1.5 pro [12]	1 fps	75.69	69.32	62.54	66.41
GPT-4o [32]	64	73.28	64.46	60.75	62.87
Claude 3.5 Sonnet [1]	20	72.44	-	-	-
Open-source Offline MLLMs					
Video-LLaMA2-7B [11]	32	49.52	-	-	-
VILA-1.5-8B [26]	14	52.32	-	-	-
Video-CCAM-14B [15]	96	53.96	-	-	-
LongVA-7B [54]	128	59.96	-	-	-
Qwen2-VL-7B [43]	64	69.04	60.65	48.58	54.62
InternVL-V2-8B [10]	16	63.72	60.73	44.00	52.37
LLaVA-NeXT-Video-32B [28]	64	66.96	-	-	-
MiniCPM-V-2.6-8B [18]	32	67.44	-	-	-
Open-source Online MLLMs					
Flash-VStream-7B [52]	-	23.23	29.86	25.35	27.61
VideoLLM-online-8B [7]	2 fps	35.99	20.79	17.73	19.26
Dispider-7B [35]	1 fps	67.63	54.55	36.06	45.31
TimeChat-Online-7B [50]	1 fps	75.36	61.90	41.70	51.80
StreamForest-7B [51]	1 fps	77.26	61.20	52.02	56.61
Training-free Offline-to-Online Methods					
LLaVA-OV-7B [23]	64	71.34	63.06	43.64	53.35
+ ReKV [13]	0.5 fps	69.22	57.33	44.16	50.75
+ LiveVLM [31]	0.5 fps	72.92	-	-	-
+ StreamKV [9]	0.5 fps	68.80	-	-	-
+ HERMES (6K tokens)	0.5 fps	72.63	65.07	48.80	56.94
+ HERMES (4K tokens)	0.5 fps	**73.23**	**66.34**	**50.20**	**58.27**
LLaVA-OV-0.5B [23]	64	59.64	49.70	34.59	42.15
+ ReKV [13]	0.5 fps	57.39	43.77	33.06	38.42
+ HERMES (6K tokens)	0.5 fps	61.04	50.34	34.75	42.55
+ HERMES (4K tokens)	0.5 fps	**62.04**	**50.72**	**34.80**	**42.76**
Qwen2.5-VL-7B [5]	1 fps	73.31	59.90	44.65	52.28
+ HERMES (6K tokens)	1 fps	78.72	68.42	48.10	58.26
+ HERMES (4K tokens)	1 fps	**79.44**	**68.98**	**49.43**	**59.21**
Qwen2.5-VL-32B [5]	1 fps	74.27	64.40	50.33	57.37
+ HERMES (6K tokens)	1 fps	**80.20**	71.93	**57.71**	**64.82**
+ HERMES (4K tokens)	1 fps	80.08	**72.37**	55.42	63.90

Table 1 Performance comparison (%) on StreamingBench and OVO-Bench. The "Avg." column reports the results of the average accuracy of real-time visual perception and backward tracing tasks.

showing finer-grained temporal (shown in Fig. 11) and spatial understanding (shown in Fig. 12) abilities of ***HERMES*** than its base model.

Offline Video Understanding The results presented in Tab. 4 demonstrate the *competitive performance of **HERMES** across multiple temporal scales on offline benchmarks*, compared to the base model and other training-free methods. Under a limited budget of video tokens, ***HERMES*** achieves performance that is better than or comparable to the corresponding base models. ***HERMES*** based on LLaVA-OV-7B surpasses the base model on long video datasets Egoschema and VideoMME, achieving 60.29% and 58.85%, respectively, and attains 56.92% accuracy on the short video dataset MVBench, which is comparable to the base model's 57.02%.

4.3 Efficiency Analysis

To evaluate the efficiency of ***HERMES***, we utilize three metrics: peak GPU memory usage, Time to First Token (TTFT), defined as the latency measured from the moment a user inputs a query to the decoding of the first output token, and Time Per Output Token (TPOT) across varying numbers of input frames. All

Model	RVS-Ego		RVS-Movie	
	Acc	Score	Acc	Score
LLaVA-OV-7B [23]	56.2	3.7	43.0	3.3
+ ReKV† [13]	63.7	4.0	54.4	3.6
+ ReKV w/o off. [13]	55.8	3.3	50.8	3.4
+ Flash-VStream [52]	57.0	4.0	53.1	3.3
+ InfiniPot-V [22]	57.9	3.5	51.4	3.5
+ StreamMem [49]	57.6	3.8	52.7	3.4
+ StreamingTOM [8]	58.3	3.9	53.2	3.5
+ HERMES (6K tokens)	**60.3**	**4.0**	**54.4**	**3.6**
+ HERMES (4K tokens)	58.3	3.9	**54.4**	**3.6**
LLaVA-OV-0.5B [23]	51.8	3.7	37.2	3.2
+ ReKV† [13]	54.7	3.9	44.6	3.4
+ HERMES (6K tokens)	**53.0**	**3.8**	**42.5**	**3.4**
+ HERMES (4K tokens)	52.7	3.8	41.7	**3.4**

Table 2 Performance on RVS-Ego and RVS-Movie. †: ReKV caches the KV states of all previously seen frames and is therefore treated as an upper bound.

Metric	Frames			
	16	64	256	512
Chunk Size: 8				
GPU Mem. / GB ↓	16.54	16.66	16.66	16.66
TTFT / ms ↓	27.01	28.41	28.44	28.41
TPOT / ms ↓	24.43	23.89	24.02	23.98
Chunk Size: 16				
GPU Mem. / GB ↓	17.46	17.66	17.66	17.66
TTFT / ms ↓	27.02	28.97	28.50	28.38
TPOT / ms ↓	24.50	23.59	23.56	53.63

Table 3 Efficiency across input frame numbers under two chunk sizes. "TTFT" denotes *Time to First Token* and "TPOT" denotes *Time Per Output Token*.

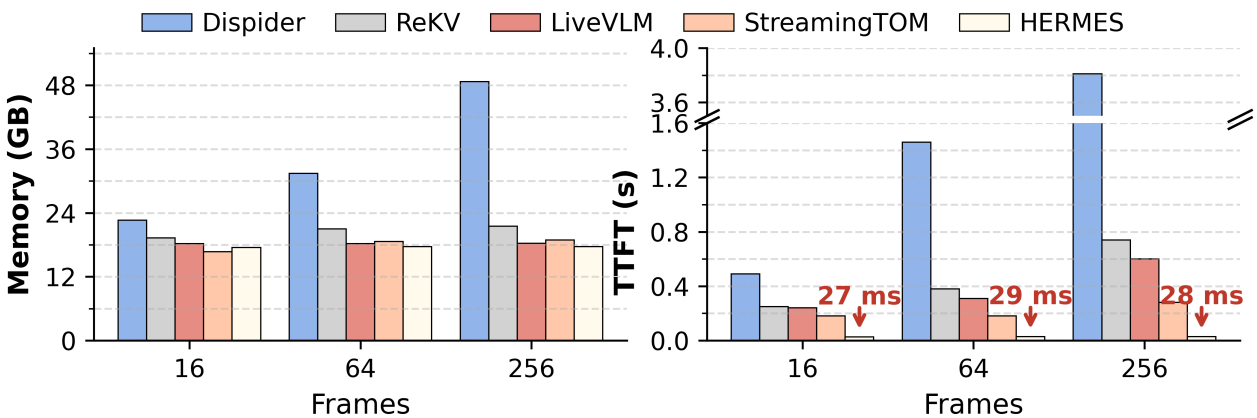


Figure 4 GPU memory and TTFT latency comparison across input frame numbers. ***HERMES*** achieves 10 × faster in TTFT compared to prior SOTA.

experiments are conducted using LLaVA-OV-7B as the base model with a 4K-token memory budget. Fig. 4 shows the comparison of memory usage and TTFT among ***HERMES*** and representative streaming methods. Unlike Dispider and LiveVLM, ***HERMES*** consistently maintains stable memory usage and TTFT as frames increase. Notably, under the 256-frame setting, ***HERMES*** achieves 1.04× reduction in peak memory compared to the prior SOTA LiveVLM, while achieving an impressive 10× speedup in TTFT over the prior SOTA StreamingTOM.

We further examine the efficiency of ***HERMES*** under varying encoded video chunk sizes, with the results shown in Tab. 3. GPU memory usage does not increase with longer video lengths due to the fixed memory budget. TTFT and TPOT remain consistently low across varying video lengths and encoding chunk sizes, confirming real-time responsiveness in practical streaming scenarios.

4.4 Ablation Study

We conduct ablation studies to evaluate the contributions of ***HERMES***'s components and hyperparameter choices, covering: (1) KV cache memory budget, (2) cross-layer memory smoothing and its hyperparameters, (3) position re-indexing strategies for streaming and offline datasets, and (4) summary tokens for long-term

Model	#Frames	MVBench	Egoschema	VideoMME Long	VideoMME Avg.
Proprietary MLLMs					
Gemini 1.5 pro [12]	1 fps	75.69	69.32	62.54	66.41
GPT-4o [32]	64	73.28	64.46	60.75	62.87
Claude 3.5 Sonnet [1]	20	72.44	-	-	-
Open-source Offline MLLMs					
Video-LLaMA2-7B [11]	32	49.52	-	-	-
VILA-1.5-8B [26]	14	52.32	-	-	-
Video-CCAM-14B [15]	96	53.96	-	-	-
LongVA-7B [54]	128	59.96	-	-	-
LLaVA-Video-7B [55]	32	58.60	57.3	-	63.30
Qwen2-VL-7B [43]	64	67.00	66.70	-	63.30
InternVL-V2-8B [10]	16	65.80	-	-	56.30
Kangaroo-7B [29]	64	64.60	-	-	-
LLaVA-NeXT-Video-32B [28]	64	66.96	-	-	-
MiniCPM-V-2.6-8B [18]	32	67.44	-	-	-
Open-source Online MLLMs					
Dispider-7B [35]	1 fps	-	55.60	-	57.20
TimeChat-Online-7B [50]	1 fps	75.36	61.90	41.70	53.22
StreamForest-7B [51]	1 fps	70.20	-	-	61.40
Training-free Offline-to-Online Methods					
LLaVA-OV-7B [23]	64	**57.02**	59.93	48.00	57.67
+ ReKV [13]	0.5 fps	56.83	**60.70**	46.89	57.74
+ HERMES (6K tokens)	0.5 fps	56.95	60.23	49.11	58.44
+ HERMES (4K tokens)	0.5 fps	56.92	60.29	**49.22**	**58.85**
Qwen2.5-VL-7B [5]	1 fps	65.00	58.47	53.89	**64.52**
+ HERMES (6K tokens)	1 fps	65.40	59.47	**54.44**	62.00
+ HERMES (4K tokens)	1 fps	**65.53**	**59.97**	53.44	60.63

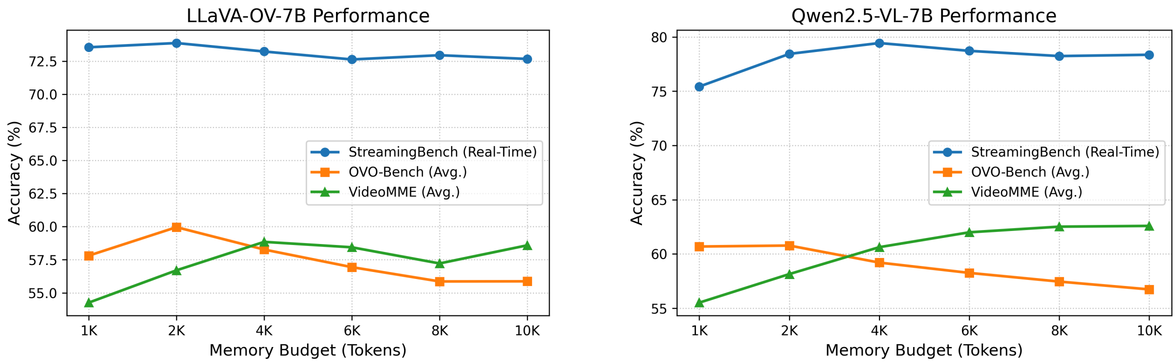
Table 4 Performance comparison (%) on offline benchmarks.

memory retention.

Memory Budget To investigate the impact of memory budget on understanding performance, we conduct ablations by varying the memory budget $|M|$ from 1K to 10K. As shown in Fig. 5a, for ***HERMES*** built upon LLaVA-OV-7B, the performance on both streaming and offline datasets stabilizes once memory budget reaches 4K. Notably, streaming datasets can tolerate a smaller memory budget. In contrast, the performance on long offline datasets degrades significantly when the memory budget is below 4K. The ablation on Qwen2.5-VL-7B is provided in Fig. 5b, yielding conclusions consistent with those on LLaVA-OV-7B.

Cross-Layer Memory Smoothing In Tab. 5, we evaluate variants without the proposed cross-layer memory smoothing mechanism, as well as alternative hyperparameter configurations. All these variants exhibit degraded performance on the VideoMME benchmark, demonstrating both the critical role of memory smoothing and the effectiveness of our chosen hyperparameter settings.

Position Re-Indexing Strategies For all streaming evaluations, we adopt the lazy position re-indexing strategy, while we use the eager re-indexing strategy for offline evaluations. Ablation studies in Tab. 7 and Tab. 8 show



(a) Performance comparison of LLaVA-OV-7B across different memory budgets.

(b) Performance Comparison of Qwen2.5-VL-7B across Different Memory Budgets.

Hyperparameter			VideoMME			
λ_{deep}	λ_{mid}	$\lambda_{shallow}$	Short	Medium	Long	**Avg.**
0	0	0	69.67	51.11	43.44	54.74
0.5	0	0	69.67	51.44	43.56	54.89
0	0.5	0	70.89	54.78	46.44	57.37
0	0	0.5	70.89	54.44	47.00	57.44
0.5	0.5	0.5	**71.78**	54.78	47.33	57.96
0.4	0.3	0.1	71.33	**54.89**	**49.11**	**58.44**

Table 5 Ablation on different cross-layer memory smoothing hyperparameter λ.

Model	**Aggregation**	VideoMME			
		Short	Medium	Long	**Avg.**
LLaVA-OV-7B	-	69.89	55.11	48.00	57.67
+ HERMES	w/o	71.33	54.78	47.78	57.96
+ HERMES	w/	**71.33**	54.89	**49.11**	**58.44**

Table 6 Ablation on summary tokens in deep layers. The gray row is our default setting in all experiments.

the effectiveness of these strategies in their respective scenarios.

Summary Tokens in Deep Layers In Sec. 3.2, we aggregate the evicted tokens in each deep layer into one summary token at each compression step. The results in Tab. 6 indicate that these summary tokens effectively preserve long-term memory, leading to improved performance on VideoMME.

5 Related Work

Streaming Video Understanding Existing MLLMs [4, 5, 12, 23] are primarily designed for pre-defined offline videos and struggle with continuous streaming videos. While some prior works have adapted existing offline MLLMs to online settings [46, 50, 51], they rely on costly model-specific training. Training-free streaming methods, such as ReKV [13] and LiveVLM [31], prefill offload KV cache to external devices. At user query time, they retrieve the full KV cache and reconstruct it on the GPU, incurring high latency and overall memory usage. In contrast, StreamMem [49] heuristically reuses KV cache, but lacks fine-grained KV cache management and interpretability. Unlike prior training-free methods, ***HERMES*** is grounded in a systematic attention analysis with improved interpretability and reliability.

KV Cache Compression for Video Input Numerous KV cache compression techniques have been proposed for offline video understanding [40, 42, 44, 48], but most of these methods are poorly suited for streaming scenarios due to the unpredictable future frames and user queries [8]. Existing online KV cache compression paradigms [8, 13, 31, 49] largely overlook the inherently hierarchical storage structure of the KV cache. ***HERMES*** addresses this gap by introducing a hierarchical KV cache management strategy, which enables fine-grained memory utilization and low-latency responses.

Model	Re-Indexing	StrBench	OVO-Bench		
		Real-Time	Real-Time	Backward	Avg.
LLaVA-OV-7B	-	71.34	63.06	43.64	53.35
+ HERMES	lazy	**72.63**	**65.07**	**48.80**	**56.94**
+ HERMES	eager	72.30	64.91	47.21	56.06

Table 7 Ablation on different re-indexing strategies on streaming benchmarks. The gray row represents our default setting in all evaluations for streaming benchmarks. "StrBench" represents *StreamingBench*.

Model	Re-Indexing	VideoMME			
		Short	Medium	Long	Avg.
LLaVA-OV-7B	-	69.89	55.11	48.00	57.67
+ HERMES	lazy	69.67	51.67	43.44	54.93
+ HERMES	eager	**71.33**	54.89	**49.11**	**58.44**

Table 8 Ablation on different re-indexing strategies on offline benchmark VideoMME. The gray row represents our default setting in all evaluations for offline benchmarks.

6 Conclusion

This paper proposes ***HERMES***, a training-free framework for efficient streaming video understanding. Guided by mechanistic attention analysis, we conceptualizes KV cache as a hierarchical video memory system across multiple granularities. By introducing a cross-layer memory smoothing and position re-indexing, ***HERMES*** further enhances the understanding performance for long streaming input. Extensive experiments demonstrate that ***HERMES*** delivers accurate performance under continuously growing video streams, while consistently maintaining extremely low response latency and compact GPU memory usage, making it well suited for real-world streaming deployment.

References

[1] Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

[2] R.C. Atkinson and R.M. Shiffrin. Human memory: A proposed system and its control processes, 1968. ISSN 0079-7421. URL https://www.sciencedirect.com/science/article/pii/S0079742108604223.

[3] Alan D. Baddeley and Graham Hitch. Working memory, 1974. ISSN 0079-7421. URL https://www.sciencedirect.com/science/article/pii/S0079742108604521.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[6] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time, 2024. URL https://arxiv.org/abs/2501.00663.

[7] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video, 2024. URL https://arxiv.org/abs/2406.11816.

[8] Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding, 2025. URL https://arxiv.org/abs/2510.18269.

[9] Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, Ming Lu, and Shanghang Zhang. Streamkv: Streaming video question-answering with segment-based kv cache retrieval and compression, 2025. URL https://arxiv.org/abs/2511.07278.

[10] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024. URL https://arxiv.org/abs/2404.16821.

[11] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024. URL https://arxiv.org/abs/2406.07476.

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian,

Vlado Galić, Wenhu Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoee Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchette, Jinmeng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmann, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhindal, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian, Wu, Guangda Lai, Paul Collins, Alex Haig, Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Laurent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roee Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayana Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasangi, Congchao Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun Yan, Gary Leung, Vikas Sindhwani, Parker Barnes, Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar Bhumihar, Sivan

Eiger, Gui Citovsky, Ben Withbroe, Zhang Li, Siyang Xue, Niccolò Dal Santo, Georgi Stoyanov, Yves Raimond, Steven Zheng, Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel Saputro, Adnan Ozturel, Ganesh Mallya, Kushal Majmundar, Ross West, Paul Caron, Jinliang Wei, Lluis Castrejon, Sharad Vikram, Deepak Ramachandran, Nikhil Dhawan, Jiho Park, Sara Smoot, George van den Driessche, Yochai Blau, Chase Malik, Wei Liang, Roy Hirsch, Cicero Nogueira dos Santos, Eugene Weinstein, Aäron van den Oord, Sid Lall, Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale Webster, Ali Elqursh, Aedan Pope, Georges Rotival, David Raposo, Wanzheng Zhu, Jeff Dean, Sami Alabed, Dustin Tran, Arushi Gupta, Zach Gleicher, Jessica Austin, Edouard Rosseel, Megh Umekar, Dipanjan Das, Yinghao Sun, Kai Chen, Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu, Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav Kanagal Shamanna, Felix Fischer, Josef Dean, Karan Gill, Ross Mcilroy, Abhishek Bhowmick, Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir Magay, Jie Tan, Dhriti Varma, Christian Walder, Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike Kwong, Ionel Gog, Chiyuan Zhang, Sander Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu Zheng, Dessie Petrova, Yann Dauphin, Yuchuan Liu, Ido Kessler, Steven Hand, Chris Duvarney, Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra Cordell, Xin Liu, Jonathan Mallinson, Pankaj Joshi, Andy Crawford, Arun Suggala, Steve Chien, Nick Fernando, Mariella Sanchez-Vargas, Duncan Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn Shan, Terry Thurk, Robin Strudel, Paul Voigtlaender, Piyush Patil, Tim Dozat, Ali Khodaei, Sahil Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang, Brian Roark, Chaitra Hegde, Tianli Ding, Angelos Filos, Zhongru Wu, André Susano Pinto, Shuang Liu, Saarthak Khanna, Aditya Pandey, Siobhan Mcloughlin, Qiujia Li, Sam Haves, Allan Zhou, Elena Buchatskaya, Isabel Leal, Peter de Boursac, Nami Akazawa, Nina Anderson, Terry Chen, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértes, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Popat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Shaham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izzard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Põder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman, Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, Ellie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa, Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy, Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi

Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Mitchelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Riachi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin, Julia Pawar, Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei, Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Abbas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirnschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maron, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Ćevid, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwag, Gil Fidel, Luheng He, Mu Cai, Yanzhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut, Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patraucean, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Pantofaru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yinlam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Chu, Alexandre Senges, Tianxiao Shen, Mukund Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo, Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato, Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand Iyer, Marcella Valentine, András György, Arun Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan Clement, Weize Kong, Drew Garmon, Ishaan Watts, Kush Bhatia, Khyatti Gupta, Matt

Miecnikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos, Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen Ju, Andy Forbes, Jesse Emond, Kushal Chauhan, Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Moufarek, Daniel Murphy, Maya Meng, Nitish Gupta, Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan, Tiberiu Sosea, James Swirhun, Linda Friso, Behnam Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder, Edouard Yvinec, Aroonalok Pyne, Victor Carbune, Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang, Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang, Dhruv Shah, Arjun Akula, Max Dylla, Ashyana Kachra, Weicheng Kuo, Tingting Zou, Lily Wang, Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon, Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia Ponomareva, Rory Blevins, Lawrence Moore, Weijun Wang, Phil Chen, Martin Scholz, Artur Dwornik, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvashi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesołowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik Paulus, Clement Farabet, Alex Wertheim, Bing Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu, Shubham Agrawal, Lily Lin, XiangHai Sheng, Peter Sung, Tyler Breland-King, Christina Butterfield, Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth Salesky, Federico Lebron, Jonni Kanerva, Michela Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten Peter, Sarmishta Velury, David Kao, Jay Hoover, Anna Bortsova, Colton Bishop, Shoshana Jakobovits, Alessandro Agostini, Alekh Agarwal, Chang Liu, Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex Greve, Anirudh GP, Jake Marcus, Le Hou, Tom Duerig, Rivka Moroshko, Dave Lacey, Andy Davis, Julien Amelot, Guohui Wang, Frank Kim, Theofilos Strinopoulos, Hui Wan, Charline Le Lan, Shankar Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sutton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau, Billy Porter, David Ross, Sujeevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Ţifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nooteboom, Jai Gupta, Loren Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau, Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko, Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander, Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behbahani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren Munkhdalai, Dana Alon, Geoff Bacon, Efren Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony, Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun, Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri, Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang, Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, Naveen Kumar, Maulik Shah, Matt Lawlor, Austin Waters, Mark Graham, Rhys May, Sabela Ramos, Sandra Lefdal, Zeynep Cankara, Nacho Cano, Brendan O'Donoghue, Jed Borovik, Frederick Liu, Jordan Grimstad, Mahmoud Alnahlawi, Katerina Tsihlas, Tom Hudson, Nikolai Grigorev, Yiling Jia, Terry Huang, Tobenna Peter Igwe, Sergei Lebedev, Xiaodan Tang, Igor

Krivokon, Frankie Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar Vashishth, Yu Liang, Balaji Venkatraman, Chenjie Gu, Anastasios Kementsietsidis, Chen Zhu, Junehyuk Jung, Yunfei Bai, Mohammad Javad Hosseini, Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal Awasthi, Adi Mayrav Gilady, Zelda Mariet, Ramy Eskander, Haiguang Li, Hexiang Hu, Guillermo Garrido, Philippe Schlattner, George Zhang, Rohun Saxena, Petar Dević, Kritika Muralidharan, Ashwin Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich, Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling Huang, Stephen Spencer, Alice Chen, James Cohan, Junjie Wang, Jonathan Tompson, Junru Wu, Ruba Haroun, Haiqiong Li, Blanca Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, Dorsa Sadigh, Jeremiah Willcock, Tingnan Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungyong Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzębski, Mimi Ly, Mohamed Elhawaty, Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen, George Tucker, Jiahao Cai, Zuguang Yang, Cindy Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria Krakovna, Hongrae Lee, AJ Maschinot, Sébastien Cevey, HyunJeong Choe, Aijun Bai, Hansa Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, Alex Hofer, Hado van Hasselt, Svetlana Grant, Yuri Chervonyi, David Silver, Andrew Hogue, Ayushi Agarwal, Kathie Wang, Preeti Singh, Four Flynn, Josh Lipschultz, Robert David, Lizzetth Bellot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory Shaw, Alex Siegman, Lin Yang, Shravya Shetty, Subhrajit Roy, Yunting Song, Wojciech Stokowiec, Ryan Burnell, Omkar Savant, Robert Busa-Fekete, Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip Lippe, Mikhail Dektiarev, Zach Behrman, Fabian Mentzer, Kelvin Nguyen, Meng Wei, Siddharth Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan, Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakyia, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin, Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui, Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinit Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilia Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andačić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson, Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana

Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg, Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen, Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han Fu, Alex Pak, Prakhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marcáurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Villela, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras, Josh Woodward, Alessandro Bissacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Maggie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha, Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpati, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshev, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Hennigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pierson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vassigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petrini, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohananey, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila, Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga, Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Lacasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaiche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leichner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evci, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Killam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Artiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Cesar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang, George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme Tubone, Ravin Kumar, Jason Baldridge, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Pierigiovanni, Jasmine George, Richard Green,

Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholz. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

[13] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval, 2025. URL https://arxiv.org/abs/2503.00540.

[14] Hermann Ebbinghaus. Memory: A contribution to experimental psychology. Annals of neurosciences, 20(4):155, 2013.

[15] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos, 2024. URL https://arxiv.org/abs/2408.14023.

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. URL https://arxiv.org/abs/2405.21075.

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video, 2022. URL https://arxiv.org/abs/2110.07058.

[18] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

[19] Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, Zhenrong Cheng, Xuanbo Fan, Jiaxin Guo, Xinlei Yu, Zhenhong Zhou, Zewen Hu, Jiahao Huo, Junhao Wang, Yuwei Niu, Yu Wang, Zhenfei Yin, Xiaobin Hu, Yue Liao, Qiankun Li, Kun Wang, Wangchunshu Zhou, Yixin Liu, Dawei Cheng, Qi Zhang, Tao Gui, Shirui Pan, Yan Zhang, Philip Torr, Zhicheng Dou, Ji-Rong Wen, Xuanjing Huang, Yu-Gang Jiang, and Shuicheng Yan. Memory in the age of ai agents, 2025. URL https://arxiv.org/abs/2512.13564.

[20] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding, 2020. URL https://arxiv.org/abs/2007.10937.

[21] Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot: Infinite context processing on memory-constrained llms, 2024. URL https://arxiv.org/abs/2410.01518.

[22] Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding, 2025. URL https://arxiv.org/abs/2506.15745.

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

[24] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2024. URL https://arxiv.org/abs/2311.17005.

[25] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. Ovo-bench: How far is your video-llms from real-world online video understanding?, 2025. URL https://arxiv.org/abs/2501.05510.

[26] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024. URL https://arxiv.org/abs/2312.07533.

[27] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding, 2024. URL https://arxiv.org/abs/2411.03628.

[28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[29] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input, 2024. URL https://arxiv.org/abs/2408.15542.

[30] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. URL https://arxiv.org/abs/2308.09126.

[31] Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval, 2025. URL https://arxiv.org/abs/2505.15269.

[32] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce

Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

[33] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

[34] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023. URL https://arxiv.org/abs/2305.13786.

[35] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction, 2025. URL https://arxiv.org/abs/2501.03218.

[36] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis, 2016. URL https://arxiv.org/abs/1604.02808.

[37] Lianlei Shan, Shixian Luo, Zezhou Zhu, Yu Yuan, and Yong Wu. Cognitive memory in large language models, 2025. URL https://arxiv.org/abs/2504.02441.

[38] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding, 2024. URL https://arxiv.org/abs/2410.17434.

[39] Haoran Sun and Shaoning Zeng. Hierarchical memory for high-efficiency long-term reasoning in llm agents, 2025. URL https://arxiv.org/abs/2507.22925.

[40] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models, 2025. URL https://arxiv.org/abs/2411.15024.

[41] Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant, 2025. URL https://arxiv.org/abs/2505.05467.

[42] Han Wang, Yuxiang Nie, Yongjie Ye, Deng GuanYu, Yanjie Wang, Shuai Li, Haiyang Yu, Jinghui Lu, and Can Huang. Dynamic-vlm: Simple dynamic visual token compression for videollm, 2024. URL https://arxiv.org/abs/2412.09530.

[43] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

[44] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos, 2025. URL https://arxiv.org/abs/2405.19209.

[45] Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge, 2025. URL https://arxiv.org/abs/2501.13468.

[46] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams, 2025. URL https://arxiv.org/abs/2510.09608.

[47] Haolin Yang, Feilong Tang, Lingxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Yifan Lu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, Zongyuan Ge, and Imran Razzak. Streamagent: Towards anticipatory agents for streaming video understanding, 2025. URL https://arxiv.org/abs/2508.01875.

[48] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models, 2024. URL https://arxiv.org/abs/2412.04467.

[49] Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding, 2025. URL https://arxiv.org/abs/2508.15717.

[50] Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, Lingpeng Kong, Qi Liu, Yuanxing Zhang, and Xu Sun. Timechat-online: 80% visual tokens are naturally redundant in streaming videos, 2025. URL https://arxiv.org/abs/2504.17343.

[51] Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. Streamforest: Efficient online video understanding with persistent event memory, 2025. URL https://arxiv.org/abs/2509.24871.

[52] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024. URL https://arxiv.org/abs/2406.08085.

[53] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024. URL https://arxiv.org/abs/2406.08085.

[54] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision, 2024. URL https://arxiv.org/abs/2406.16852.

[55] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025. URL https://arxiv.org/abs/2410.02713.

Appendix

Appendix Contents

A More Attention Visualization

We provide more detailed attention visualization in Fig. 6 under different sliding window sizes, showing that the observed attention patterns consistently hold across varying window lengths, thus confirming the generality of the findings in Sec. 2.

B Guidance Prompt

The following two figures show the local and global guidance prompt with and without conversation history to guide the token compression, respectively. For the deep layers, since they primarily focus on frame-level global semantic information, we employ a global guidance prompt as a pseudo-query to extract attention weights of video tokens. In contrast, the middle layers lie in a transition between recency-biased attention and global semantic focus. Therefore, we adopt a hybrid guidance strategy, in which the local guidance prompt and the global guidance prompt are concatenated into a single prompt string to jointly guide the token compression.

C Configuration of Cross-Layer Memory Smoothing

Given that long-term memory tends to remain relatively stable, while short-term memory focuses on diverse perception, we set different λ for different layer stages:

$$\lambda_l = \begin{cases} 0.1, & \text{if } l \in \mathcal{L}_{shallow} \\ 0.3, & \text{if } l \in \mathcal{L}_{middle} \\ 0.4, & \text{if } l \in \mathcal{L}_{deep} \end{cases} \tag{7}$$

The ablation study Tab. 5 shows the effectiveness of this hyperparameter choice.

D Details of evaluated benchmarks

Table 9 Key statistics of the streaming benchmarks. In the "Type" column, "MC" denotes multiple-choice questions, while "OE" denotes open-ended questions. In the "Benchmark" column, "rt" denotes real-time understanding subset, while "bw" denotes backward tracing subset.

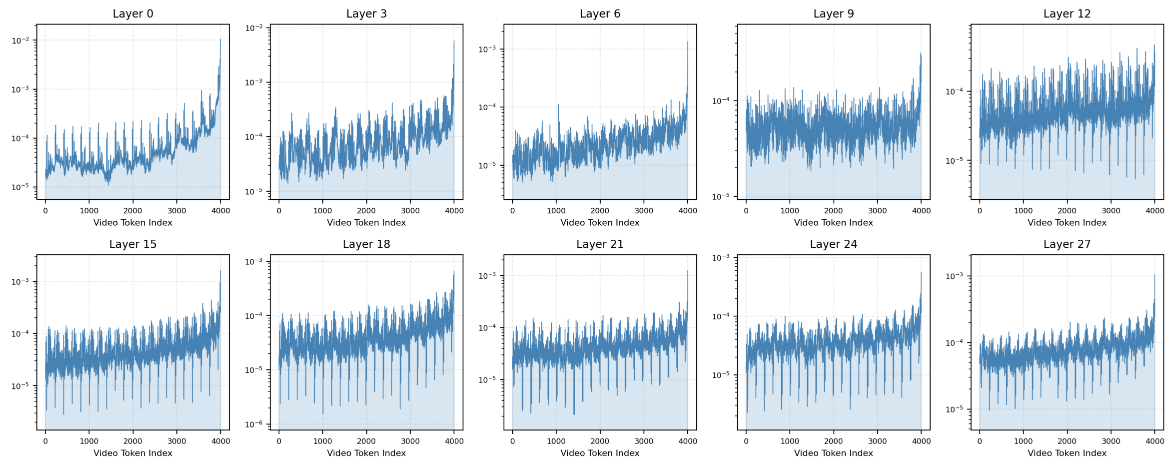
Benchmark	Duration	#Videos	#QA	Type
StreamingBench$_{rt}$	10.1min	500	2,500	MC
OVO-Bench$_{bw}$	5.9 min	275	631	MC
OVO-Bench$_{rt}$	8.8 min	237	837	MC
RVS-Ego	60 min	10	1,465	OE
RVS-Movie	30 min	22	1,905	OE

Table 10 Key statistics of the offline benchmarks. In the "Type" column, "MC" denotes multiple-choice questions.

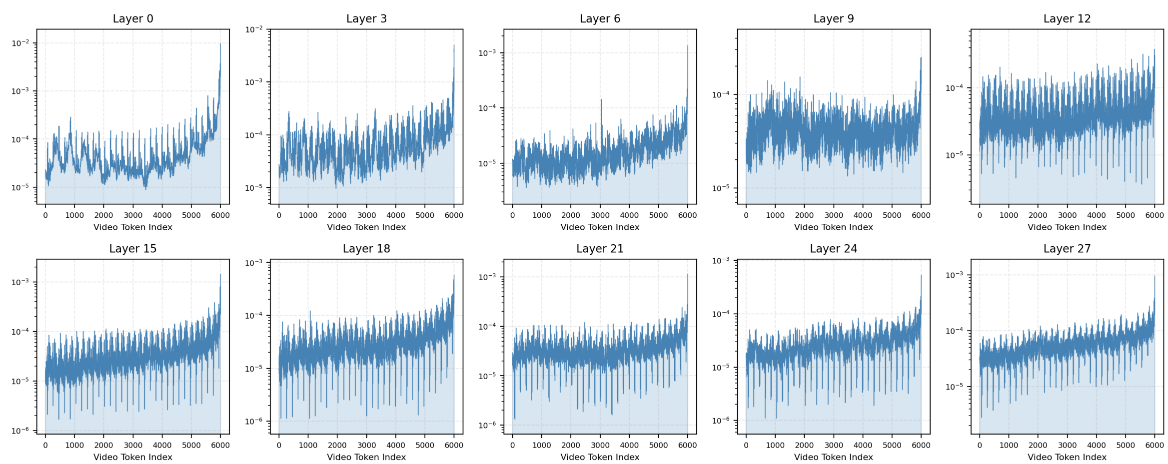
Benchmark	Duration	#Videos	#QA	Type
MVBench	16 s	3,641	4,000	MC
Egoschema	3 min	5,063	5,063	MC
VideoMME	17 min	900	2,700	MC

D.1 Streaming Benchmarks

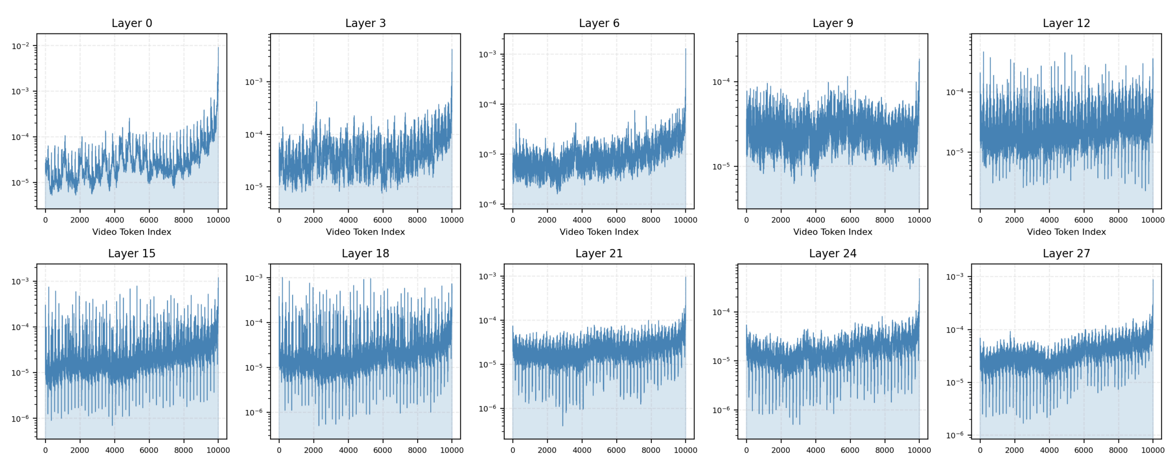
- **StreamingBench** [27] assesses the streaming video understanding capabilities of MLLMs. It evaluates three core aspects: real-time visual understanding, omni-source understanding, and contextual understanding. The Real-Time Visual Understanding subset is the most extensive component, featuring 2,500 questions across 500 videos. It covers 10 tasks, such as object perception and causal reasoning. In this paper, we focus on the Real-Time Visual Understanding subset for evaluation.
- **OVO-Bench** [25] evaluates the online reasoning and temporal awareness of MLLMs, featuring 644 videos with approximately 2,800 fine-grained multiple-choice QA pairs. It organizes 12 tasks into three distinct categories, which are real-time visual perception, backward tracing, and forward active responding. Given



(a) Sliding window of 4,000 video tokens



(b) Sliding window of 6,000 video tokens



(c) Sliding window of 10,000 video tokens

Figure 6 Visualization of the average attention weights of video tokens in LLaVA-OV-7B under different sliding window sizes.

Find recent details related to: {last_conv}. Describe the current scene in detail, focusing on specific objects, fine-grained actions, and spatial relationships.

Figure 7 Local guidance prompt to guide the token compression if conversation history exists. ”last_conv” refers to the last user query and the corresponding model answer from the conversation history.

Describe the current scene in detail, focusing on specific objects, fine-grained actions, and spatial relationships.

Figure 8 Local guidance prompt to guide the token compression if there is no conversation history.

Context summary: {last_conv}. Summarize the video narrative, identifying main characters, key events, timeline changes, and the overall theme.

Figure 9 Global guidance prompt to guide the token compression if conversation history exists. ”last_conv” refers to the last user query and the corresponding model answer from the conversation history.

Summarize the video narrative, identifying main characters, key events, timeline changes, and the overall theme.

Figure 10 Global guidance prompt to guide the token compression if there is no conversation history.

that we do not focus on the proactive responding ability of MLLMs in this paper, we exclusively utilize the real-time perception and the backward tracing subsets.

- **RVS-Ego** and **RVS-Movie** [53] are designed to evaluate the real-time understanding capabilities of models in online streaming scenarios. The datasets consist of 10 long ego-centric videos from the Ego4D dataset [17] and 22 long movie clips from the MovieNet dataset [20] dataset, totaling over 21 hours of video content.

D.2 Offline Benchmarks

- **MVBench** [24] systematically evaluates the temporal understanding capabilities of MLLMs. It utilizes a novel static-to-dynamic method to define 20 distinct temporal tasks, such as action sequence and moving direction, which cannot be effectively solved with a single frame. The videos are collected from a wide range of datasets, including NTU RGB+D [36], Perception [34], etc.
- **Egoschema** [30] is a diagnostic benchmark designed to assess long-form video understanding abilities. Derived from Ego4D [17], it consists of over 5,000 human-curated multiple-choice QA pairs associated with egocentric video clips.
- **VideoMME** [16] is a full-spectrum, multimodal benchmark designed for the comprehensive evaluation of MLLMs in video analysis. It comprises 900 manually curated videos spanning six primary domains and diverse durations to assess temporal adaptability. The dataset features 2,700 high-quality QA pairs that necessitate processing multimodal inputs, including video frames, subtitles, and audio.

E Details of Position Re-Indexing

Inspired by StreamingVLM’s strategy of managing positional stability in streaming scenarios [46], we adopt a unified left-compaction re-indexing scheme to eliminate positional gaps introduced by KV-cache pruning while preserving the semantic anchoring of the system prompt. Concretely, system text tokens are kept fixed to provide a stable textual anchor, whereas retained video tokens are re-indexed in a left-compact manner and placed contiguously after the static prefix. To reuse cached key states without re-computation, we further apply a delta-based rotary correction that compensates for the positional displacement.

E.1 Re-indexing for LLaVA-OV (1D RoPE)

LLaVA-OV employs standard 1D RoPE, where each token is associated with a scalar positional index p. Therefore, we perform left-compaction of the 1D indices: the system prefix positions remain unchanged, while the retained positions of video tokens are reassigned to form a dense contiguous segment immediately following the fixed prefix.

Let offset denote the length of the system prompt prefix tokens, and let

$$\mathcal{P} = \{p_0 < p_1 < \cdots < p_{N-1}\}$$

be the sorted set of retained video token positions (excluding the fixed prefix). For a retained video token originally at position $p_{\mathrm{old}} \in \mathcal{P}$, its compacted 1D position is defined as

$$p_{\mathrm{new}} = \mathrm{offset} + \mathrm{rank}_{\mathcal{P}}(p_{\mathrm{old}}) . \tag{8}$$

This mapping removes gaps while preserving the original temporal ordering along the stream, and ensures that the video region occupies a dense range directly after the static text region.

To align cached key states with the updated positions, we avoid re-generating keys and instead apply a rotary delta correction induced by the positional shift. For a cached key vector $\mathbf{k}_{\mathrm{old}}$ associated with position p_{old} and remapped to p_{new}, we compute

$$\mathbf{k}_{\mathrm{new}} = \mathbf{k}_{\mathrm{old}} \odot \mathrm{RotaryDelta}(p_{\mathrm{old}}, p_{\mathrm{new}}) , \tag{9}$$

where the relative phase shift is

$$\mathrm{RotaryDelta}(p_{\mathrm{old}}, p_{\mathrm{new}}) = e^{i(p_{\mathrm{new}} - p_{\mathrm{old}})\boldsymbol{\theta}}, \tag{10}$$

and $\boldsymbol{\theta}$ denotes the RoPE frequency vector. This update preserves the correctness of attention under the new indexing while enabling direct reuse of the cached KV states.

E.2 Re-indexing for Qwen2.5-VL (3D M-RoPE)

For Qwen2.5-VL, video tokens are indexed by a 3D M-RoPE coordinate $\mathbf{p} = (p^{(t)}, p^{(h)}, p^{(w)})$, covering temporal and spatial dimensions. After pruning, the retained video tokens typically occupy sparse coordinates along each dimension $d \in \{t, h, w\}$. To eliminate the gaps without disturbing the monotonic ordering, we apply dimension-wise left-compaction independently along each axis, while keeping the system token prefix fixed.

Let

$$\mathcal{P}^{(d)} = \{p_0^{(d)} < p_1^{(d)} < \cdots < p_{N_d-1}^{(d)}\}$$

denote the sorted set of retained coordinates along dimension d. For a token originally located at $p_{\mathrm{old}}^{(d)} \in \mathcal{P}^{(d)}$, its compacted coordinate is defined by its rank within $\mathcal{P}^{(d)}$, shifted by the fixed prefix offset:

$$p_{\mathrm{new}}^{(d)} = \mathrm{offset} + \mathrm{rank}_{\mathcal{P}^{(d)}}\left(p_{\mathrm{old}}^{(d)}\right), \qquad d \in \{t, h, w\}. \tag{11}$$

This procedure yields a dense and contiguous (t, h, w) grid for the video tokens placed immediately after the static text region, thereby ensuring positional continuity while preserving the distinct semantic roles of temporal and spatial indices.

As in the 1D case, we reuse cached keys by applying a M-RoPE correction. Given a key $\mathbf{k}_{\mathrm{old}}$ associated with

$$\mathbf{p}_{\mathrm{old}} = (p_{\mathrm{old}}^{(t)}, p_{\mathrm{old}}^{(h)}, p_{\mathrm{old}}^{(w)})$$

and remapped to

$$\mathbf{p}_{\text{new}} = (p_{\text{new}}^{(t)}, p_{\text{new}}^{(h)}, p_{\text{new}}^{(w)}),$$

the corrected key is obtained as

$$\mathbf{k}_{\text{new}} = \mathbf{k}_{\text{old}} \odot \mathrm{RotaryDelta}(\mathbf{p}_{\text{old}}, \mathbf{p}_{\text{new}})\,, \tag{12}$$

with the relative phase shift:

$$\mathrm{RotaryDelta}(\mathbf{p}_{\text{old}}, \mathbf{p}_{\text{new}}) = \underset{d\in\{t,h,w\}}{\mathrm{Concat}} \left(e^{i(p_{\text{new}}^{(d)} - p_{\text{old}}^{(d)})\boldsymbol{\theta}^{(d)}} \right), \tag{13}$$

where Concat denotes the concatenation operation along the channel dimension, and $\boldsymbol{\theta}^{(d)}$ represents the rotary frequency vector corresponding to the channel section allocated for dimension d.

F Algorithm of Summary Tokens

Algorithm 1 Summary Token Aggregation

Require: K_p, V_p: Pruned KV tensors from visual tokens; P_p: Original position indices of pruned tokens; t: Target position index for the summary token.
Ensure: k_{sum}, v_{sum}: Single aggregated summary token cache.

Step 1: Aggregate Value
\# Simple spatial mean
$v_{sum} \leftarrow \text{Mean}(V_p)$

Step 2: Aggregate Key
\# Phase alignment before pooling
$\Delta\theta \leftarrow \text{RotaryDelta}(P_p \rightarrow t)$
\# Calculate rotation shift from P_p to t
$K_{aligned} \leftarrow \text{ApplyDelta}(K_p, \Delta\theta)$
\# Align all keys to the same phase
$k_{sum} \leftarrow \text{Mean}(K_{aligned})$

Step 3: Update KV Cache
$K_{new} \leftarrow \text{Concat}([K_{kept}, k_{sum}])$
$V_{new} \leftarrow \text{Concat}([V_{kept}, v_{sum}])$

return K_{new}, V_{new}

G Full Performances

G.1 StreamingBench

G.2 OVO-Bench

H Case Study

We provide six representative case study examples from RVS-Ego and RVS-Movie to demonstrate the advantages of ***HERMES*** compared to the foundation model LLaVA-OV-7B. During the understanding of streaming long videos, ***HERMES*** exhibits significantly finer-grained temporal (shown in Fig. 11) and spatial understanding Fig. 12 capabilities than its corresponding foundation model.

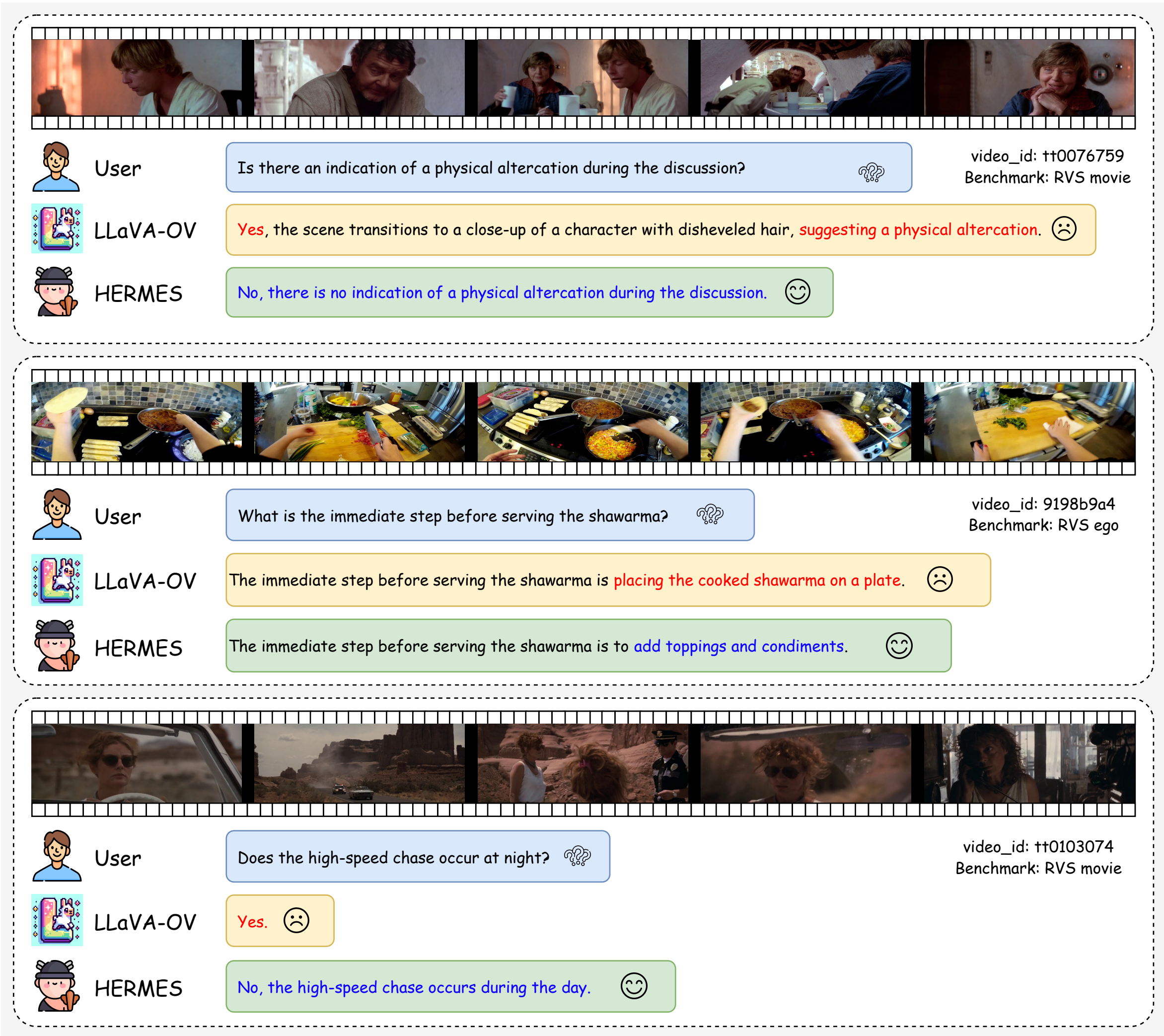


Figure 11 Cases demonstrating the superior fine-grained temporal understanding capability of ***HERMES*** relative to the LLaVA-OV-7B base model.

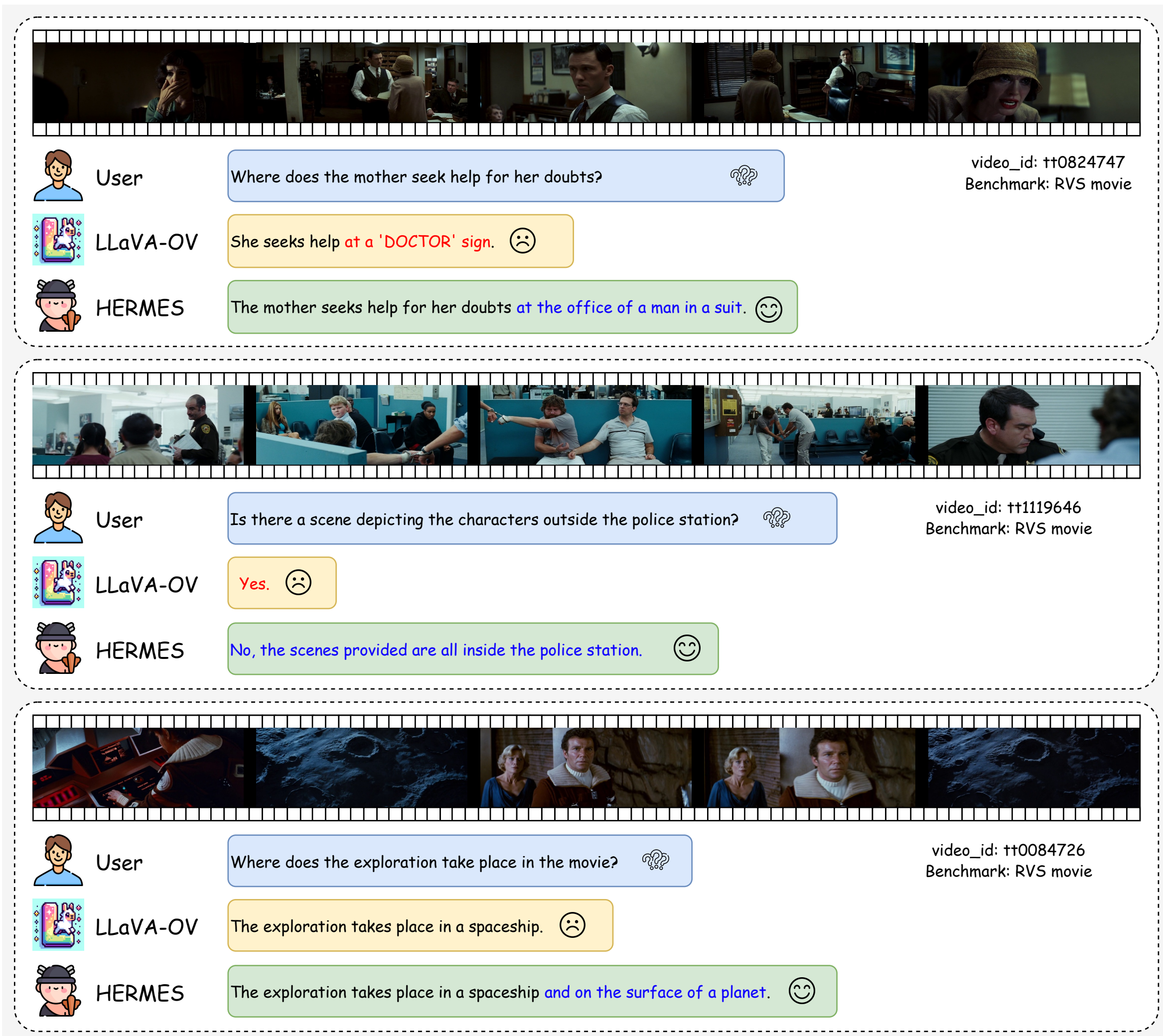


Figure 12 Cases demonstrating the superior fine-grained spatial understanding capability of ***HERMES*** relative to the LLaVA-OV-7B base model.

Table 11 Accuracy comparison (%) on StreamingBench focusing on *Real-Time Visual Understanding* tasks. Real-Time Visual Understanding tasks consists of Object Perception (OP), Causal Reasoning (CR), Clips Summarization (CS), Attribute Perception (ATP), Event Understanding (EU), Text-Rich Understanding (TR), Prospective Reasoning (PR), Spatial Understanding (SU), Action Perception (ACP), and Counting (CT).

Model	#Frames	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Avg.
Human	-	89.47	92.00	93.60	91.47	95.65	92.52	88.00	88.75	89.74	91.30	91.46
Proprietary MLLMs												
Gemini 1.5 pro [12]	1 fps	79.02	80.47	83.54	79.67	80.00	84.74	77.78	64.23	71.95	48.70	75.69
GPT-4o [32]	64	77.11	80.47	83.91	76.47	70.19	83.80	66.67	62.19	69.12	49.22	73.28
Claude 3.5 Sonnet [1]	20	73.33	80.47	84.09	82.02	75.39	79.53	61.11	61.79	69.32	43.09	72.44
Open-source Offline MLLMs												
Video-LLaMA2-7B [11]	32	55.86	55.47	57.41	58.17	52.80	43.61	39.81	42.68	45.61	35.23	49.52
VILA-1.5-8B [26]	14	53.68	49.22	70.98	56.86	53.42	53.89	54.63	48.78	50.14	17.62	52.32
Video-CCAM-14B [15]	96	56.40	57.81	65.30	62.75	64.60	51.40	42.59	47.97	49.58	31.61	53.96
LongVA-7B [54]	128	70.03	63.28	61.20	70.92	62.73	59.50	61.11	53.66	54.67	34.72	59.96
InternVL-V2-8B [10]	16	68.12	60.94	69.40	77.12	67.70	62.93	59.26	53.25	54.96	56.48	63.72
Kangaroo-7B [29]	64	71.12	84.38	70.66	73.20	67.08	61.68	56.48	55.69	62.04	38.86	64.60
LLaVA-NeXT-Video-32B [28]	64	78.20	70.31	73.82	76.80	63.35	69.78	57.41	56.10	64.31	38.86	66.96
MiniCPM-V-2.6-8B [18]	32	71.93	71.09	77.92	75.82	64.60	65.73	70.37	56.10	62.32	53.37	67.44
Open-source Online MLLMs												
Flash-VStream-7B [52]	-	25.89	43.57	24.91	23.87	27.33	13.08	18.52	25.20	23.87	48.70	23.23
VideoLLM-online-8B [7]	2 fps	39.07	40.06	34.49	31.05	45.96	32.40	31.48	34.16	42.49	27.89	35.99
Dispider-7B [35]	1 fps	74.92	75.53	74.10	73.08	74.44	59.92	76.14	62.91	62.16	45.80	67.63
TimeChat-Online-7B [50]	1 fps	80.22	82.03	79.50	83.33	76.10	78.50	78.70	64.63	69.60	57.98	75.36
StreamForest-7B [51]	1 fps	83.11	82.81	82.65	84.26	77.50	78.19	76.85	69.11	75.64	54.40	77.26
Training-free Offline-to-Online Methods												
LLaVA-OV-7B [23]	32	78.75	78.12	80.76	**81.19**	71.70	72.59	72.22	63.82	66.01	38.34	71.34
+ ReKV [13]	0.5 fps	76.02	81.25	77.92	76.90	66.04	66.04	69.44	60.98	64.31	49.22	69.22
+ LiveVLM [31]	0.5 fps	**81.47**	78.13	83.28	79.08	69.57	**74.14**	**75.00**	**69.11**	67.71	40.41	72.92
+ StreamKV [9]	0.5 fps	73.80	77.30	85.90	77.50	**73.30**	63.90	69.40	61.40	63.20	35.80	68.80
+ HERMES (6K tokens)	0.5 fps	77.93	**82.03**	86.12	**81.19**	66.04	73.52	74.07	63.01	67.71	**45.08**	72.63
+ HERMES (4K tokens)	0.5 fps	79.02	81.25	**87.70**	80.20	69.18	71.96	73.15	66.26	**69.41**	43.52	**73.23**
LLaVA-OV-0.5B [23]	32	71.39	57.81	65.93	69.64	69.18	55.76	57.41	**52.85**	62.04	16.58	59.64
+ ReKV [13]	0.5 fps	65.12	60.16	66.56	66.01	66.67	52.96	57.41	48.37	60.34	18.13	57.39
+ HERMES (6K tokens)	0.5 fps	71.93	60.16	69.09	71.29	68.55	57.32	**60.19**	51.22	**63.74**	**19,69**	61.04
+ HERMES (4K tokens)	0.5 fps	**72.21**	**61.72**	**70.98**	**72.94**	**72.33**	**57.94**	**60.19**	**52.85**	**63.74**	19.17	**62.04**
Qwen2.5-VL-7B [5]	1 fps	77.93	76.56	78.55	80.86	76.73	76.95	80.56	65.45	65.72	**52.85**	73.31
+ HERMES (6K tokens)	0.5 fps	83.38	78.91	86.12	87.13	**78.62**	**86.60**	**84.26**	74.80	71.39	46.63	78.72
+ HERMES (4K tokens)	0.5 fps	**83.65**	**81.25**	**88.01**	**87.46**	76.73	**86.60**	82.41	**76.02**	**73.94**	46.63	**79.44**
Qwen2.5-VL-32B [5]	1 fps	76.29	79.69	78.55	**83.50**	76.10	79.44	80.56	61.38	68.27	**59.07**	74.27
+ HERMES (6K tokens)	0.5 fps	**84.47**	79.69	**87.70**	83.17	**81.76**	**88.16**	86.11	74.80	**77.62**	49.22	**80.20**
+ HERMES (4K tokens)	0.5 fps	83.92	**80.47**	**87.70**	**83.50**	80.50	**88.16**	**87.04**	**75.20**	77.34	48.19	80.08

Table 12 Accuracy comparison (%) on OVO-Bench focusing on *Real-Time Visual Perception* and *Backward Tracing* tasks. Real-Time Visual Perception tasks consist of Optical Character Recognition (OCR), Action Recognition (ACR), Attribute Recognition (ATR), Spatial Understanding (STU), Future Prediction (FPD), Object Recognition (OJR). Backward Tracing tasks consists of Episodic Memory (EPM), Action Sequence Identification (ASI), Hallucination Detection (HLD).

Model	**#Frames**	**Real-Time Visual Perception**							**Backward Tracing**				**Overall**
		OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	Avg.
Human	-	93.96	92.57	94.83	92.70	91.09	94.02	93.20	92.59	93.02	91.37	92.33	92.77
Proprietary MLLMs													
Gemini 1.5 Pro [12]	1fps	85.91	66.97	79.31	58.43	63.37	61.96	69.32	58.59	76.35	52.64	62.54	65.93
GPT-4o [32]	64	69.80	64.22	71.55	51.12	70.30	59.78	64.46	57.91	75.68	48.66	60.75	62.61
Open-source Offline MLLMs													
LLaVA-Video-7B [55]	64	69.80	59.63	66.38	50.56	72.28	61.41	63.34	51.18	64.19	9.68	41.68	52.51
Qwen2-VL-7B [43]	64	69.13	53.21	63.79	50.56	66.34	60.87	60.65	44.44	66.89	34.41	48.58	54.62
InternVL2-8B [10]	64	68.46	58.72	68.97	44.94	67.33	55.98	60.73	43.10	61.49	27.41	44.00	52.37
LongVU-7B [38]	1fps	55.70	49.54	59.48	48.31	68.32	63.04	57.40	43.10	66.22	9.14	39.49	48.45
Open-source Online MLLMs													
VideoLLM-online-8B [7]	2fps	8.05	23.85	12.07	14.04	45.54	21.20	20.79	22.22	18.80	12.18	17.73	19.26
Flash-VStream-7B [52]	1fps	25.50	32.11	29.31	33.71	29.70	28.80	29.86	36.36	33.78	5.91	25.35	27.61
Dispider-7B [35]	1fps	57.72	49.54	62.07	44.94	61.39	51.63	54.55	48.48	55.41	4.30	36.06	45.31
TimeChat-Online-7B [50]	1fps	75.20	46.80	70.70	47.80	69.30	61.40	61.90	55.90	59.50	9.70	41.70	51.80
StreamForest-7B [51]	1fps	68.46	53.21	71.55	47.75	65.35	60.87	61.20	58.92	64.86	32.26	52.02	56.61
Training-free Offline-to-Online Methods													
LLaVA-OV-7B [23]	32	67.79	55.05	72.41	48.31	72.28	62.50	63.06	57.24	55.41	18.28	43.64	53.35
+ ReKV [13]	0.5 fps	52.35	54.13	69.83	43.26	67.33	57.07	57.33	57.58	56.08	18.82	44.16	50.75
+ HERMES (6K tokens	0.5 fps	**72.48**	**62.39**	69.83	47.75	**73.27**	64.67	65.07	**61.28**	58.78	26.34	48.80	56.94
+ HERMES (4K tokens	0.5 fps	**72.48**	**62.39**	**74.14**	**50.56**	**73.27**	**65.22**	**66.34**	60.61	**61.49**	**28.49**	**50.20**	**58.27**
LLaVA-OV-0.5B [23]	32	53.69	53.21	48.28	**33.71**	**60.40**	**48.91**	49.70	46.13	45.27	**12.37**	34.59	42.15
+ ReKV [13]	0.5 fps	41.61	44.95	50.00	29.78	**60.40**	35.87	43.77	46.13	43.92	9.14	33.06	38.42
+ HERMES (6K tokens	0.5 fps	**57.05**	**49.54**	55.17	32.58	**60.40**	47.28	50.34	**47.81**	47.30	9.14	34.75	42.55
+ HERMES (4K tokens	0.5 fps	56.38	47.71	**56.90**	32.02	62.38	**48.91**	**50.72**	**47.81**	**47.97**	8.60	**34.80**	**42.76**
Qwen2.5-VL-7B [5]	1fps	67.79	55.05	67.24	42.13	66.34	60.87	59.90	**51.52**	58.78	23.66	44.65	52.28
+ HERMES (6K tokens	0.5 fps	**85.91**	60.55	**74.14**	52.81	70.30	**66.85**	68.42	49.49	61.49	33.33	48.10	58.26
+ HERMES (4K tokens	0.5 fps	85.23	**64.22**	71.55	**53.37**	**74.26**	65.22	**68.98**	48.48	**62.16**	**37.63**	**49.43**	**59.21**
Qwen2.5-VL-32B [5]	1fps	77.18	58.72	68.10	50.56	**74.26**	57.61	64.40	**58.59**	62.84	29.57	50.33	57.37
+ HERMES (6K tokens	0.5 fps	87.25	**66.06**	**74.14**	57.30	71.29	75.54	71.93	55.56	**70.27**	47.31	**57.71**	**64.82**
+ HERMES (4K tokens	0.5 fps	**88.59**	65.14	**74.14**	**58.99**	71.29	**76.09**	**72.37**	52.19	66.22	**47.85**	55.42	63.90