AccKV: Towards Efficient Audio-Video LLMs Inference via Adaptive-Focusing and Cross-Calibration KV Cache Optimization

Zhonghua Jiang[1*], Kui Chen[1*], Kunxi Li[1*], Keting Yin[1], Yiyun Zhou[1], Zhaode Wang[2], Chengfei Lv[2], Shengyu Zhang[1†]

[1]Zhejiang University
[2]Alibaba Group
{jiangzhonghua, chen_kui, kunxili, yinkt, yiyunzhou, sy_zhang}@zju.edu.cn
{zhaode.wzd, chengfei.lcf}@alibaba-inc.com

Abstract

Recent advancements in Audio-Video Large Language Models (AV-LLMs) have enhanced their capabilities in tasks like audio-visual question answering and multimodal dialog systems. Video and audio introduce an extended temporal dimension, resulting in a larger key-value (KV) cache compared to static image embedding. A naive optimization strategy is to selectively focus on and retain KV caches of audio or video based on task. However, in the experiment, we observed that the attention of AV-LLMs to various modalities in the high layers is not strictly dependent on the task. In higher layers, the attention of AV-LLMs shifts more towards the video modality. In addition, we also found that directly integrating temporal KV of audio and spatial-temporal KV of video may lead to information confusion and significant performance degradation of AV-LLMs. If audio and video are processed indiscriminately, it may also lead to excessive compression or reservation of a certain modality, thereby disrupting the alignment between modalities. To address these challenges, we propose AccKV, an Adaptive-Focusing and Cross-Calibration KV cache optimization framework designed specifically for efficient AV-LLMs inference. Our method is based on layer adaptive focusing technology, selectively focusing on key modalities according to the characteristics of different layers, and enhances the recognition of heavy hitter tokens through attention redistribution. In addition, we propose a Cross-Calibration technique that first integrates inefficient KV caches within the audio and video modalities, and then aligns low-priority modality with high-priority modality to selectively evict KV cache of low-priority modality. The experimental results show that AccKV can significantly improve the computational efficiency of AV-LLMs while maintaining accuracy.

Introduction

Recently, the rapid development of Audio-Video Large Language Models (AV-LLMs) (Cheng et al. 2024; Tang et al. 2024b; Liu et al. 2025c) has significantly improved their performance in tasks such as video captioning (Wang et al. 2019; Geng et al. 2024; Liu et al. 2025a), audio-visual question answering (Sun et al. 2024; Chen et al. 2020; Wu et al. 2024), and multimodal dialog systems (Team et al. 2023; Alamri et al. 2019; Chen et al. 2025; Wang et al. 2025a; Yu et al. 2024). AV-LLMs, like VideoLLaMA2 (Cheng et al. 2024) and AVicuna (Tang et al. 2024b), realize an efficient understanding of audio and video signals by integrating spatial-temporal modeling and cross-modal alignment technology. However, as the scale of the AV-LLMs expands, the memory usage problem of key-value (KV) cache in the inference process becomes increasingly prominent, especially for audio and video data, which naturally have long sequences and high redundancy characteristics, further exacerbating the computational and storage overhead. How to efficiently manage multimodal KV cache has become the core challenge in the actual deployment of AV-LLMs.

In order to speed up the reasoning process (Feng and Zhang 2023; Xiao et al. 2025; Xie et al. 2025) of Large Language Models (LLMs) (Achiam et al. 2023; Zhou et al. 2025a; Yang et al. 2024; Zhou, Han, and Chen 2025; Liu et al. 2024b), several methods (Shi et al. 2024; Li et al. 2025a; Jiang et al. 2025a; Li et al. 2025b) have been proposed to optimize the KV cache. The KV cache scheme based on dynamic attention filtering (such as H2O (Zhang et al. 2023), SnapKV (Li et al. 2024b)) significantly reduces computing cost by maintaining a sparse set of important KV caches, but its design paradigm focuses on text scenarios and fails to adapt to heterogeneous redundancy patterns in multimodal data (Li et al. 2024a; Alamri et al. 2019; Jinghan Yu 2025). LOOK-M (Wan et al. 2024) and FastV (Chen et al. 2024) formulate KV cache evaluation and merging strategies for the redundancy of visual tokens. However, due to AV-LLMs' unique attention pattern to audio and video and the unique spatial-temporal characteristics of audio and video data, the existing KV cache strategy cannot be directly applied to AV-LLMs.

Video and audio introduce extended temporal dimensions, leading to larger key-value caches compared to static image embeddings. A simple optimization strategy is to selectively focus on and retain the KV cache of audio or video according to specific task requirements. However, we observe that:

- **Attention Convergence:** In high layers, AV-LLMs' at-

[*]These authors contributed equally.
[†]The corresponding author.

tention to each modality is not strictly dependent on specific tasks. Regardless of the task type (such as audio question and answer or video description), AV-LLMs tend to focus more on the video modality in higher layers.

- **Heterogeneous modality merging conflict:** Directly merging the KV cache of audio with temporal characteristics and video with spatial-temporal characteristics may cause information confusion, thus significantly reducing AV-LLMs performance. This is because the redundancy characteristics and distribution of different modalities are significantly different. A simple direct merging will destroy the structure of modal information and cause confusion of feature distribution.
- **Excessive compression/reservation trap:** If audio and video are processed indiscriminately, it may also lead to excessive compression or reservation of a certain modality, and destroy the balance and alignment between different modalities. In many real-world scenes, audio and visual events are synchronized (for example, voice and lip movements). AV-LLMs need to learn this synchronization to understand the content correctly.

To solve these problems, we propose AccKV, an Adaptive-Focusing and Cross-Calibration KV cache optimization framework. AccKV dynamically focuses on key modalities according to AV-LLMs' attention pattern in different layers through layer adaptive focusing. At the same time, Cross-Calibration technology is introduced to align the KV cache of low-priority and high-priority modality after merging the redundancy within each modality, so as to avoid information confusion. In addition, existing methods usually rely on accumulated attention scores to identify important tokens. However, because the attention matrix is a lower triangular matrix, the early token is more likely to be selected as the key token, and the accumulated score of the most recent token can never be greater than the first token. AccKV uses attention redistribution to enhance the recognition of important tokens by reallocating attention weight.

Specifically, AccKV conducts an in-depth analysis of the attention patterns of audio and video modality tokens in different layers of AV-LLMs, dynamically calculates the modality priority of audio and video in each layer, and assigns differentiated weights accordingly, effectively focusing on key modalities. At the same time, we introduced an attention redistribution mechanism that improves the accuracy of identifying important tokens by readjusting the distribution of attention scores, assigning lower weights to attention scores related to early tokens and higher weights to attention scores related to more recent tokens. In addition, AccKV adopts Cross-Calibration technology, which first merges inefficient KV caches within each modality. This step reduces unnecessary information redundancy and effectively avoids information confusion caused by directly integrating different characteristic tokens. Subsequently, a cross-modal aligner based on cosine similarity was designed, using the K state of high-priority modality as anchor points to perform threshold filtering on low-priority modality, ensuring alignment between the two modalities.

The experimental results show that AccKV can significantly reduce the computational cost of AV-LLMs while maintaining accuracy. In summarize, this paper makes the following key contributions:

- We observed a unique attention pattern in AV-LLMs and proposed a new Adaptive-Focusing and Cross-Calibration KV cache optimization framework based on it, named AccKV.
- Based on AccKV framework, we have implemented a baseline that dynamically focuses on the key modality according to the attention patterns of AV-LLMs in different layers. At the same time, Cross-Calibration technology is introduced to compress the KV cache within each modality, and then align the KV cache of low-priority and high-priority modality. In addition, we use attention redistribution techniques to enhance the recognition of important tokens.
- We conducted extensive experimental validation of the AccKV framework and tested it on two AV-LLMs (VideoLLaMA2 (Cheng et al. 2024) based on Qwen2-7B-Instruction (Yang et al. 2024) language decoder and AVicuna (Tang et al. 2024b) based on Vicuna-7B-v1.5 (Zheng et al. 2023) language decoder). The experiment covered multiple datasets, and the results showed that AccKV can achieve efficient inference while maintaining performance in different scenarios, verifying its effectiveness and universality.

Related Works

In LLMs, KV cache, as a key optimization technique, improves inference efficiency by caching previously calculated Key and Value vectors to avoid repeated calculations in the decoding stage. To address the growing computational demand (Zhou et al. 2025b; Li et al. 2025c; Wang et al. 2025b; Jiang et al. 2025b), H2O (Zhang et al. 2023) introduces a mechanism based on cumulative attention scores to identify and ignore relatively unimportant tokens by maintaining the top-k most representative KV pairs, thereby reducing computational and space complexity. However, H2O (Zhang et al. 2023) ignores the fact that the importance of tokens may change dynamically with the context. The Quest (Tang et al. 2024a) strategy adopts a page-level KV cache management method, using the inherent sparsity characteristics of the self-attention mechanism, using the Key state in each page to estimate its importance, and selecting the most important pages for precise calculation, thereby effectively reducing the computational load, but failing to reduce space complexity. The KIVI (Liu et al. 2024c) strategy focuses on the quantization characteristics of Key and Value vectors, and effectively reduces memory usage by implementing channel-level quantization on Key and token-level quantization on Value, while maintaining model performance. FastGen (Ge et al. 2023) further analyzes the functional differences of different attention heads, customizes specific compression strategies for each head, selects important tokens based on cumulative attention scores, and shows significant efficiency improvements in processing long sequences. Although the above strategies have shown powerful capabilities in plain text LLM scenarios, they do not take into ac-

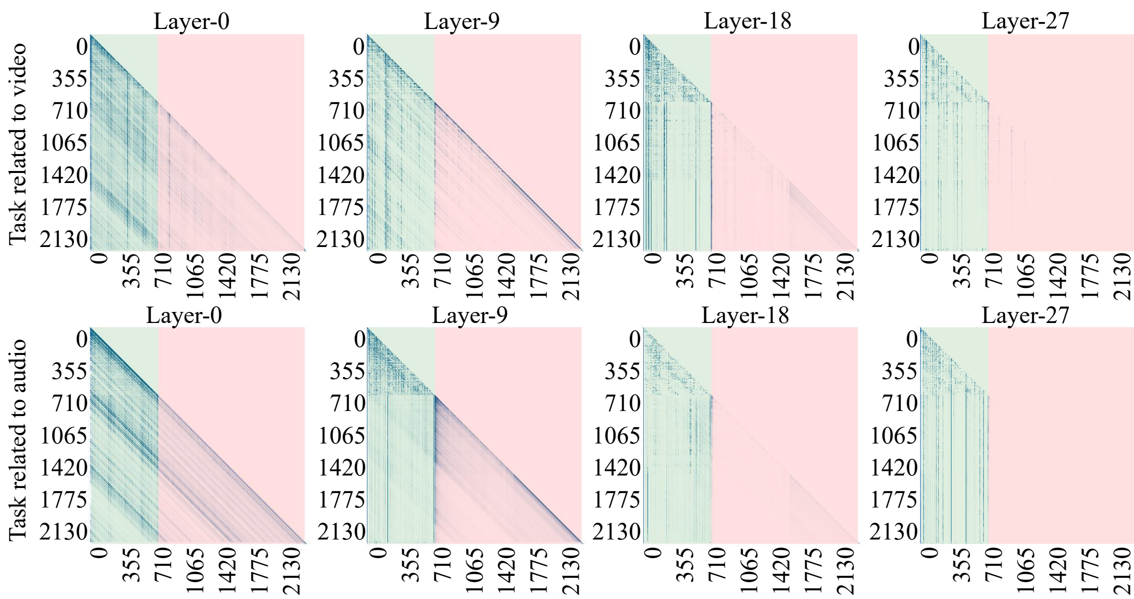

Figure 1: The attention distribution of VideoLLaMA2. Green and pink denote video and audio modalities, respectively. Blue represents attention score, and the darker the color, the higher the score. In lower layers, audio and video modalities each receive partial attention. In higher layers, the attention of AV-LLMs converge to video modality, regardless of the task type.

count the redundant characteristics of tokens (Team et al. 2023), which limits their performance in MLLMs (Zhou, Yao, and Chen 2025; Feng and Zhang 2024; Feng et al. 2025; Liu et al. 2025b,d).

In MLLMs, the optimization of KV cache requires customized strategies for the redundant characteristics of multimodal tokens. LOOK-M (Wan et al. 2024) analyzes the attention scores in the prefilling phase, prioritizes text tokens, and dynamically adjusts the retention strategy of KV cache in the decoding phase. It achieves efficient KV cache management by merging similar tokens. MustDrop (Liu et al. 2024a) eliminates redundancy in the spatial dimension by merging adjacent tokens with high similarity and establishes a set of key tokens in the visual encoding phase, evaluates the importance of tokens through global and individual attention scores in the prefilling phase, and further compresses the KV cache in the decoding phase. PruneVid (Huang, Zhou, and Han 2024) calculates the attention scores of question tokens and image tokens, selectively prunes video features related to question tokens, and minimizes video redundancy by merging spatial-temporal tags, achieving efficient KV cache management. However, these strategies are mainly targeted at the visual modality, and do not fully consider the unique attention patterns of the AV-LLMs for the audio-visual modality and the spatial-temporal alignment characteristics of the audio-visual modality. Therefore, they cannot be applied directly to AV-LLMs.

Method

Observation

Attention convergence. In order to further analyze the attention patterns of AV-LLMs in different layers, we observed the attention distribution of each layer in detail. As shown in Figure 1, we found that in the lower layers, the audio and video modalities each received partial attention. However, at higher levels, the attention of AV-LLMs tends to converge to the video modality, regardless of the task type (such as audio Q&A or video description). This indicates that AV-LLMs tend to pay more attention to video modality in higher layers. Through this observation, we can better understand the internal working mechanism of AV-LLMs and guide the design of more effective KV cache management strategies.

In addition, we have verified this phenomenon through further experiments. The results in Table 1 show that evicting the KV cache of video tokens in the high layer of the model results in significant performance degradation for both audio and video related tasks. In contrast, evicting the audio token from the KV cache has a relatively small impact on the performance of AV-LLMs. This proves the existence of attention convergence in AV-LLMs.

Algorithm 1: AccKV Algorithm.

Require: Attention score matrix $A \in \mathbb{R}^{l \times l}$, KV Cache K and V, the budget size of cache k, threshold τ.
Ensure: Updated KV cache.
1: Apply Attention Redistribution: $A_{i,j} \leftarrow \frac{i+1}{l-j} * A_{i,j}, \ \forall i, j \in [0, l-1]$
2: Calculate modal priority weights W_v, W_a using Eq. (3) to (4).
3: Adjust attention scores based on W_v, W_a.
4: Select top-k tokens with highest accumulated attention scores using Eq. (6) to (10).
5: Merge unimportant tokens within the same modality using Eq. (11) and (14).
6: Perform cross-modal alignment using cosine similarity as described in Eq. (15) to (18).
7: Evict KV cache of low-priority modality base on threshold τ.

	Full Cache	Evict Audio	Evict Video
Audio Task	0.700490	0.691272	0.523281
Video Task	0.541667	0.525415	0.356390

Table 1: The impact of evicting video and audio KV cache in higher layers on the performance of AV-LLMs.

	Full Cache	Evict	Merge
Audio Task	0.700490	0.651984	0.166571
Video Task	0.541677	0.474414	0.120523

Table 2: The impact of heterogeneous modal merging on the performance of AV-LLMs. Evict: Evict the unimportant KV cache in audio modality and the unimportant KV cache in video modality. Merge: Merge the unimportant KV cache in the audio modality with the unimportant KV cache in the video modality.

	Full Cache	Evict All Audio	Evict All Video
Audio Task	0.700490	0.018601	0.56708
Video Task	0.541667	0.490000	0.09015

Table 3: The impact of evicting all audio KV cache or all video KV cache on AV-LLMs performance.

Heterogeneous modality merging conflict. In experiment, we observed that the LOOK-M (Wan et al. 2024) method performed poorly in audio and video scenes. After in-depth analysis, we believe that the main reason for this phenomenon is that LOOK-M (Wan et al. 2024) directly merges the audio KV cache with temporal characteristics with the video KV cache with spatial-temporal characteristics. This direct merging strategy fails to fully consider the significant differences in redundancy characteristics and distributions between different modalities, thereby disrupting the structure of modal information, leading to confusion in feature distributions, and ultimately significantly reducing model performance.

To verify this hypothesis, we designed and conducted a small-scale experiment. In this experiment, we attempted to merge the unimportant KV cache in the audio modality with the unimportant KV cache in the video modality. The results in Table 2 indicate that this merging operation indeed leads to a significant decrease in the performance of AV-LLMs, thus verifying our hypothesis.

Excessive compression/reservation trap. If audio and video are processed indiscriminately, it may also lead to excessive compression or reservation of a certain modality, and destroy the balance and alignment between different modalities. In many real-world scenarios, audio and visual events are synchronized (such as speech and lip movements). AV-LLMs need to learn this synchronicity in order to correctly understand the content.

We designed an experiment to verify the importance of modal alignment: directly evict all audio KV cache or all video KV cache, and observe the changes in AV-LLMs performance. Table 3 shows that even if all KV caches of modalities that are not directly related to the task are

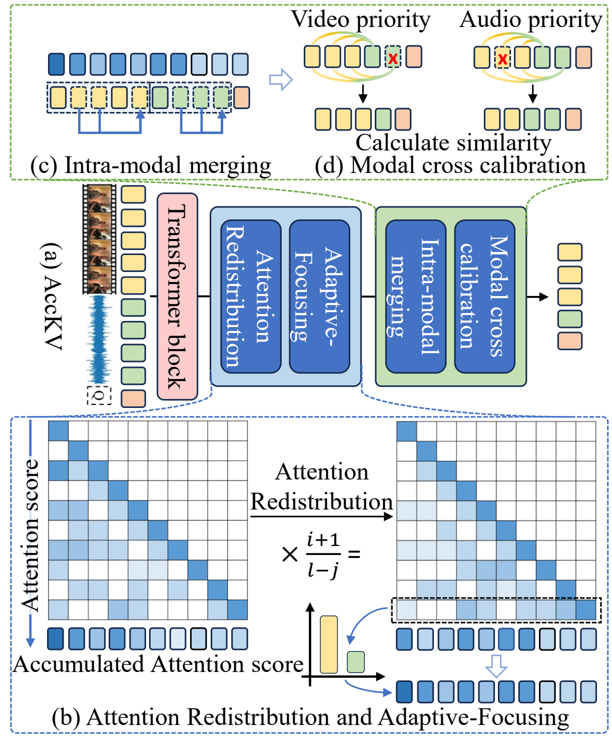

Figure 2: (a) Overview of AccKV. AccKV uses (b) attention redistribution to alleviate attention bias and adaptive focusing to identify important modalities, thereby avoiding misjudgment of KV cache importance. In addition, AccKV first (c) performs KV cache compression within each modality, and then (d) aligns low priority modality with high priority modality to maintain information structure integrity and reduce information confusion.

evicted, the performance of AV-LLMs will decrease. This result shows that the balance and alignment between modalities are critical to AV-LLMs performance, and excessive compression or retention of information of a certain modality will destroy this balance, thus affecting the AV-LLMs' ability to understand and process multimodal information.

AccKV

In this section, we propose AccKV, an Adaptive-Focusing and Cross-Calibration KV cache optimization framework. AccKV adaptively focuses on key modalities based on the attention patterns of AV-LLMs in different layers. AccKV introduces Cross-Calibration technology to compress redundancy within the modality, and then aligns the KV cache of low priority and high priority modality to avoid information confusion. In addition, AccKV adopts attention redistribution to enhance the recognition of important tokens.

Attention Redistribution. Most existing methods rely on accumulating attention scores to identify important tokens. However, due to the fact that the attention matrix is a lower triangular matrix, this mechanism tends to make early appearing tokens easier to identify as key tokens. To address this limitation, AccKV has introduced an attention reallocation that dynamically adjusts attention weights to more accurately identify the true important tokens.

Specifically, for a given attention matrix $A \in R^{l \times l}$, AccKV assigns lower weights to attention scores related to early tokens and higher weights to attention scores related to more recent tokens:

$$A_{i,j} = \frac{i+1}{l-j} * A_{i,j}, \tag{1}$$

$$A = softmax(\frac{QK^T}{\sqrt{d_k}}), \tag{2}$$

where l represents the length of the token sequence, Q denotes the query matrix of the input tokens, d_k is the dimension of K.

Adaptive-Focusing. Based on the observation results, we found that AV-LLMs exhibit different attention patterns in different layers. At the lower level, both audio and video modalities receive partial attention weights. It is worth noting that in higher levels, regardless of the type of task being performed, AV-LLMs tend to give more attention weight to video modalities. The AccKV framework dynamically calculates the modal priority of audio and video in different layers based on the attention patterns of audio and video modalities in AV-LLMs, and assigns differentiated weights accordingly, effectively focusing on key modalities.

Specifically, in each layer, AccKV calculates the average accumulated attention score of the text token at the end of the sequence with respect to the previously appearing video and audio tokens:

$$\bar{S}^v = \frac{\sum_{j=I_0^v}^{j<I_1^v} \sum_{i=I_0^t}^{i<l} A_{i,j}}{I_1^v - I_0^v}, \quad \bar{S}^a = \frac{\sum_{j=I_0^a}^{j<I_1^a} \sum_{i=I_0^t}^{i<l} A_{i,j}}{I_1^a - I_0^a}, \tag{3}$$

where the index range of the video token is $[I_0^v, I_1^v)$, the index range of the audio token is $[I_0^a, I_1^a)$, and the index range of the text token is $[I_0^t, l)$. Then we normalize $\bar{S}^v$ and $\bar{S}^a$ to obtain the modal priority of video and audio, which is used as the focus weight:

$$W^v = \frac{\bar{S}^v}{\bar{S}^v + \bar{S}^a}, \quad W^a = \frac{\bar{S}^a}{\bar{S}^v + \bar{S}^a}. \tag{4}$$

In order to focus AV-LLMs on the modality that this layer focuses on, AccKV weights the attention scores of video and audio:

$$A_{:,j} = \begin{cases} W^v * A_{:,j} & \text{if } I_0^v \leq j < I_0^v \\ W^a * A_{:,j} & \text{if } I_0^a \leq j < I_0^a. \end{cases} \tag{5}$$

Cross-Calibration. The merging of KV cache can compress redundancy while preserving some information. However, directly merging the KV cache of audio with temporal characteristics and video with spatial-temporal characteristics may cause information confusion and affect model performance. Based on this observation, we propose merging inefficient KV caches within each modality. Firstly, we calculate the accumulated attention score of the text token at the end of the sequence for the previously appearing video and audio tokens:

$$S^v = \sum_{i=I_0^t}^{i<l} A_{i,j}, I_0^v \leq j < I_1^v. \tag{6}$$

Assuming we need to preserve the KV cache of k multimodal tokens, we will select these tokens based on the accumulated attention score. The index of the k multimodal tokens with the highest accumulated attention score is:

$$I_{top} = top_k([S^v, S^a], k). \tag{7}$$

The number of reserved video and audio tokens are:

$$N^v = Min(Count(\{I | I \in I_{top}, I_0^v \leq I < I_1^v\}), 1), \tag{8}$$

$$N^a = Min(Count(\{I | I \in I_{top}, I_0^a \leq I < I_1^a\}), 1). \tag{9}$$

Then, we merge unimportant tokens within each modality. The index of the retained multimodal tokens are:

$$I_{top}^v = top_k([S^v], N^v), \quad I_{top}^a = top_k([S^a], N^a). \tag{10}$$

The index of the multimodal tokens that need to be merged is:

$$I_{merge}^v = \{I | I \notin I_{top}^v, I_0^v \leq I < I_1^v\}, \tag{11}$$

$$I_{merge}^a = \{I | I \notin I_{top}^a, I_0^a \leq I < I_1^a\}. \tag{12}$$

Within each modality, we use the average merging method to handle the KV cache that needs to be merged:

$$K_{merge}^v = Avg(K[I_{merge}^v]), V_{merge}^v = Avg(V[I_{merge}^v]), \tag{13}$$

$$K_{merge}^a = Avg(K[I_{merge}^a]), V_{merge}^a = Avg(V[I_{merge}^a]). \tag{14}$$

Subsequently, we introduce a cross-modal alignment mechanism based on cosine similarity, using the K vectors of high-priority modality as anchors to perform threshold filtering on low-priority modality. Specifically, assuming that the video modality has a higher priority in a certain layer, i.e. $W^v > W^a$, we will calculate the cosine similarity of each K vector of the audio modality relative to all K vectors of the video modality:

$$\begin{aligned} Sim = & Cos\_Sim([K[I_{top}^a], K_{merge}^a], [K[I_{top}^v], K_{merge}^v]) \\ & \in R^{(N^a+1) \times (N^v+1)}. \end{aligned} \tag{15}$$

Next, we average the cosine similarity between the K vectors of each audio token and the K vectors of all video tokens:

$$\overline{Sim} = Avg(Sim, dim = -1) \in R^{(N^a+1)}. \tag{16}$$

We evict the KV cache of low-priority modality with average similarity below the set threshold τ, ensuring alignment with high-priority modality. The last remaining KV cache is:

$$K = [K[I_{top}^v], K_{merge}^v, filter(K[I_{top}^a], K_{merge}^a), K[I_0^t, l]], \tag{17}$$

$$V = [V[I_{top}^v], V_{merge}^v, filter(V[I_{top}^a], V_{merge}^a), V[I_0^t, l]]. \tag{18}$$

When the priority of the audio modality is higher, we adopt the same processing method.

Experiments

Setting

MVBench (Li et al. 2024a) is a comprehensive and challenging multimodal video understanding benchmark that covers 20 complex video understanding tasks. In our experiment,

	AS	AP	UA	OI	OS	AL	ST	AC	SC	CO	AVSD
Full Cache	0.763425	0.782715	0.736952	0.80300	0.893333	0.737667	0.881910	0.515000	0.641925	0.58	0.479462
20% Cache Budget											
H2O	0.678517	0.757552	0.624750	0.769631	0.856584	0.724232	0.821194	0.490000	0.610161	0.451667	0.352680
SnapKV	0.711629	0.735932	0.642784	0.762555	0.868453	0.721337	0.838955	0.485833	0.602885	0.469944	0.424881
LOOK-M	0.001667	0.001429	0.012052	0.017500	0.008333	0.008000	0.004345	0.000000	0.005000	0.035000	0.161588
FastV	0.608172	0.643370	0.674452	0.629312	0.800072	0.653714	0.796022	0.452917	0.567033	0.490263	0.422403
AccKV (τ=0.9)	**0.748626**	**0.76603**	0.683374	**0.866752**	0.866752	0.736667	0.854456	**0.506667**	**0.641558**	**0.550000**	**0.425179**
AccKV (τ=0.6)	0.743455	0.760238	**0.695836**	0.785690	**0.869406**	**0.737667**	**0.86098**	0.503333	0.632954	0.538750	0.423846
10% Cache Budget											
H2O	0.571982	0.585202	0.508345	0.550764	0.726125	0.601895	0.705956	0.358417	0.458830	0.458830	0.332905
SnapKV	0.644221	0.664302	0.556322	0.660142	0.731768	0.653572	0.789548	0.383639	0.516122	0.364548	0.325341
LOOK-M	0.001667	0.001111	0.012052	0.000000	0.004167	0.000000	0.004345	0.000000	0.005000	0.025000	0.160928
FastV	0.587647	0.659611	**0.703651**	0.640790	0.790262	0.638909	0.804543	0.407239	0.546424	0.462627	0.338811
AccKV (τ=0.9)	0.642198	**0.681243**	0.642874	**0.707762**	0.785710	**0.663565**	0.829384	**0.432333**	**0.602918**	0.497167	**0.390516**
AccKV (τ=0.6)	**0.649873**	0.675981	0.631588	0.694291	**0.803389**	0.660920	**0.830027**	0.421917	0.593929	**0.501500**	0.385812

Table 4: VideoLLaMA2: Performance of eviction strategy on MVBench and AVSD. The best results are highlighted in bold. The second result is highlighted with an underline.

we evaluated 10 tasks that included both audio and video, including: Action Sequence (AS), Action Prediction (AP), Unexpected Action (UA), Object Interaction (OI), Object Shuffle (OS), Action Localization (AL), Scene Transition (ST), Action Count (AC), State Change (SC) and Character Order (CO). In addition, to further validate the effectiveness of the algorithm, we also conducted experiments using the Audio Visual Scene Aware Dialogue (AVSD) (Alamri et al. 2019) dataset. The AVSD (Alamri et al. 2019) dataset focuses on dialog understanding and provides rich audio-visual information. We use ROUGE as the evaluation metric

To comprehensively verify the effectiveness of AccKV, we conducted extensive experiments on two AV-LLMs equipped with different language decoders: VideoLLaMA2 (Cheng et al. 2024) based on Qwen2-7B-Instruct (Yang et al. 2024) language decoder and AVicuna (Tang et al. 2024b) based on Vicuna-7B-v1.5 (Zheng et al. 2023) language decoder. Through this approach, we are able to evaluate the performance of AccKV under different architectures and decoding mechanisms. In the experiment, we compared four representative KV caching strategies. Among them, H2O (Zhang et al. 2023) and SnapKV (Li et al. 2024b) focus on KV cache optimization for LLMs. FastV (Chen et al. 2024) and LOOK-M (Wan et al. 2024) are specifically optimized for KV caching in pure visual scenes. We conducted experiments on a NVIDIA A100 with 40GB memory.

Experiment Results

In Table 4 and 5, we conducted a comprehensive comparative analysis of AccKV with various eviction strategies in audio and video understanding scenarios. The experimental results clearly demonstrate the effectiveness of AccKV in managing KV cache in memory constrained environments, while maintaining high performance in various tasks. Specifically, AccKV achieved up to 80%-90% reduction in memory usage, with only a slight decrease in performance compared to a fully cached solution. This demonstrates the ability of AccKV to significantly reduce memory usage with minimal performance cost.

Compared with baseline eviction methods, AccKV consistently outperforms other methods on most datasets. It is worth noting that AccKV exhibits significant superiority over text-based unimodal KV exclusion methods. This discovery reveals a key limitation of text unimodal KV eviction methods: they often overlook unique redundant patterns in multimodal data, resulting in incorrect KV cache eviction and erroneous response generation. In addition, AccKV also performed well compared to LOOK-M and FastV. LOOK-M, as a method specifically designed for multimodal scenes, significantly reduces model performance by directly merging audio KV cache with temporal characteristics and video KV cache with spatial-temporal characteristics, resulting in information confusion. FastV failed to fully consider the unique attention pattern of AV-LLMs in the high layers, and indiscriminately evicted the audio and video KV cache, which may lead to excessive compression and disruption of inter-modal alignment. In contrast, AccKV uses layer-adaptive focusing technology to dynamically adjust modal weights at each layer based on its unique attention pattern. At the same time, AccKV effectively avoids information confusion and ensures alignment between modalities through modal internal merging and cross calibration mechanisms. AccKV maintains the performance of AV-LLMs while compressing KV cache. This method performs well in complex audio and video modal scenes, verifying its effectiveness and universality in multimodal scenes.

Influence of Various Cache Budgets

To evaluate the effectiveness of AccKV under different cache budget conditions, we conducted experiments on VideoLLaMA2 and AVicuna. The experimental results are presented in Table 4 and 5, respectively. As the cache budget decreases, the performance of other KV cache evic-

	AS	AP	UA	OI	OS	AL	ST	AC	SC	CO	AVSD
Full Cache	0.557046	0.409944	0.434683	0.632619	0.563191	0.43307	0.656051	0.236181	0.505397	0.356611	0.321997
160 tokens Cache Budget											
H2O	0.102427	0.142147	0.157986	0.203755	0.238484	0.021622	0.214895	0.055501	0.172033	0.006933	0.046019
SnapKV	0.277837	0.390677	0.167643	0.489307	0.336405	0.005714	0.229411	0.182790	0.274575	0.109262	0.274924
LOOK-M	0.069215	0.095429	0.161076	0.108192	0.202830	0.089147	0.246255	0.037798	0.108152	0.000625	0.019048
FastV	0.290436	0.308278	0.164896	0.497083	0.424673	0.044266	0.267676	0.098631	0.312610	0.043869	0.303037
AccKV (τ=0.9)	0.560320	**0.421473**	0.433202	0.629119	0.551913	**0.447873**	0.677995	0.215736	**0.505354**	**0.338278**	0.319018
AccKV (τ=0.6)	**0.567582**	0.420181	**0.435402**	**0.637786**	**0.55907**	0.445151	**0.680395**	**0.240208**	0.498489	0.334302	**0.321268**
120 tokens Cache Budget											
H2O	0.015563	0.005615	0.086318	0.027991	0.174468	0.027225	0.286768	0.036946	0.089592	0.001000	0.035033
SnapKV	0.255297	0.341656	0.350181	0.350181	0.065491	0.065142	0.240582	0.146657	0.173491	0.033500	0.303612
LOOK-M	0.010662	0.006688	0.051570	0.020160	0.168745	0.014839	0.238061	0.115274	0.056398	0.002000	0.213519
FastV	0.230038	0.277104	0.164116	0.342501	0.097991	0.144145	0.247935	0.108969	0.180847	0.068095	0.213519
AccKV (τ=0.9)	**0.556088**	**0.421328**	0.421478	0.634119	0.549646	**0.433463**	0.665016	**0.230069**	**0.510921**	0.358382	0.315302
AccKV (τ=0.6)	0.555433	0.418847	**0.433581**	**0.636119**	**0.561166**	0.428963	**0.667789**	0.229931	0.510109	**0.360734**	**0.319977**

Table 5: AVicuna: Performance of eviction strategy on MVBench and AVSD. The best results are highlighted in bold. The second result is highlighted with an underline.

A-R	A-F	C-C	AVSD
✔	✔	✔	**0.425179**
✔	✔	✘	0.364630
✔	✘	✘	0.342471
✘	✘	✘	0.334883

Table 6: Ablation study. A-R: Attention Redistribution, A-F: Adaptive-Focusing, C-C: Cross-Calibration. The best results are highlighted in **bold**.

tion strategies has declined significantly. In contrast, AccKV has demonstrated its outstanding robustness and efficiency. Specifically, under extremely strict memory constraints, such as VideoLLaMA2 retaining only 10% of its cache budget and AVicuna retaining only 120 tokens of KV cache, AccKV can still maintain stable performance. This indicates that AccKV can accurately identify and retain key information in KV cache, thereby minimizing the risk of information loss while significantly reducing memory usage.

Ablation Study

To evaluate the individual contributions of each component in our proposed AccKV framework, we conducted an ablation study. Specifically, we examined the impact of three key modules: Attention Redistribution (A-R), Adaptive-Focusing (A-F), and Cross-Calibration (C-C). The results are summarized in Table 6. The result shows the performance of the model on the AVSD dataset when different combinations of these modules are enabled or disabled. When all three modules (A-R, A-F, and C-C) are active, the model achieves the highest performance.

These results highlight the importance of each module of AccKV. Attention Redistribution helps in more accurately identifying important tokens by adjusting attention weights dynamically. Adaptive-Focusing ensures that the model can dynamically focus on key modalities based on the layer's attention patterns. Cross-Calibration integrates redundant information within each modality and aligns low-priority and high-priority modality effectively, avoiding information confusion. Together, these modules work synergistically to improve the model's efficiency and performance.

Conclusion

This study first observed some unique phenomena in AV-LLMs that are crucial for improving model inference efficiency. Specifically, we found attention convergence phenomenon, where in higher layers, AV-LLMs' attention to different modalities does not strictly depend on specific tasks. Regardless of the task type, AV-LLMs tend to focus more on the video modality in higher layers. At the same time, directly merging audio KV cache with temporal characteristics and video KV cache with spatial-temporal characteristics may lead to information confusion, significantly reducing the performance of AV-LLMs. If audio and video are processed indiscriminately, it may lead to over compression or reservation of a certain modality, and destroy the balance and synchronization between different modalities. Based on the above findings, we propose AccKV, which is specifically designed to improve the computational efficiency in the inference process of AV-LLMs. AccKV dynamically selects the key modality based on the attention patterns of different layers of AV-LLMs through Adaptive-Focusing technology, and uses Cross-Calibration technology to integrate redundant information internally before aligning low-priority and high-priority modality to avoid information confusion. In addition, AccKV introduces an attention reallocation mechanism to more accurately identify important tokens by adjusting attention weights. The experimental results show that AccKV can significantly reduce the computational cost of AV-LLMs while maintaining model accuracy. This verifies the effectiveness and universality of AccKV.

Acknowledgments

This work was supported by the Key Research and Development Program of Zhejiang Province (No. 2025C01026), the National Natural Science Foundation of China (No. 62402429, U24A20326, 62441236), the Ningbo Yongjiang Talent Introduction Programme (2023A-397-G). Additionally, this work was partially supported by Young Elite Scientists Sponsorship Program by CAST (2024QNRC001). The author gratefully acknowledges the support of Zhejiang University Education Foundation Qizhen Scholar Foundation.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Lee, S.; Anderson, P.; Essa, I.; Parikh, D.; Batra, D.; Cherian, A.; Marks, T. K.; and Hori, C. 2019. Audio-Visual Scene-Aware Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725. IEEE.

Chen, J.; Hu, J.; Wang, G.; Jiang, Z.; Zhou, T.; Chen, Z.; and Lv, C. 2025. TaoAvatar: Real-Time Lifelike Full-Body Talking Avatars for Augmented Reality via 3D Gaussian Splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 10723–10734. Computer Vision Foundation / IEEE.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 19–35. Springer.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Feng, Z.; Guo, Q.; Xiao, X.; Xu, R.; Yang, M.; and Zhang, S. 2025. Unified Video Generation via Next-Set Prediction in Continuous Domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19427–19438.

Feng, Z.; and Zhang, S. 2023. Efficient vision transformer via token merger. *IEEE Transactions on Image Processing*, 32: 4156–4169.

Feng, Z.; and Zhang, S. 2024. Evolved Hierarchical Masking for Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ge, S.; Zhang, Y.; Liu, L.; Zhang, M.; Han, J.; and Gao, J. 2023. Model tells you what to discard: Adaptive kv cache compression for llms. *arXiv preprint arXiv:2310.01801*.

Geng, T.; Zhang, J.; Wang, Q.; Wang, T.; Duan, J.; and Zheng, F. 2024. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. *arXiv preprint arXiv:2411.19772*.

Huang, X.; Zhou, H.; and Han, K. 2024. PruneVid: Visual Token Pruning for Efficient Video Large Language Models. *arXiv preprint arXiv:2412.16117*.

Jiang, Z.; Li, K.; Zhou, Y.; Liu, S.; Wang, Z.; lv, C.; and Zhang, S. 2025a. PureKV: Plug-and-Play KV Cache Optimization with Spatial-Temporal Sparse Attention for Vision-Language Large Models. arXiv:2510.25600.

Jiang, Z.; Xu, J.; Zhang, S.; Shen, T.; Li, J.; Kuang, K.; Cai, H.; and Wu, F. 2025b. FedCFA: Alleviating Simpson's Paradox in Model Aggregation with Counterfactual Federated Learning. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 17662–17670.

Jinghan Yu, Z. M. Y. M. K. L. Y. W. D. L. X. M. J. L., Junhao Xiao. 2025. MILD: Multi-Layer Diffusion Strategy for Complex and Precise Multi-IP Aware Human Erasing. arXiv:2508.06543.

Li, K.; Jiang, Z.; Shen, Z.; Wang, Z.; Lv, C.; Zhang, S.; Wu, F.; and Wu, F. 2025a. MadaKV: Adaptive Modality-Perception KV Cache Eviction for Efficient Multimodal Long-Context Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 13306–13318. Association for Computational Linguistics.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, K.; Xiong, Y.; Jiang, Z.; Zhou, Y.; Wang, Z.; Lv, C.; and Zhang, S. 2025b. FlowMM: Cross-Modal Information Flow Guided KV Cache Merging for Efficient Multimodal Context Inference. arXiv:2511.05534.

Li, K.; Zhan, T.; Fu, K.; Zhang, S.; Kuang, K.; Li, J.; Zhao, Z.; Wu, F.; and Wu, F. 2025c. Mergenet: Knowledge migration across heterogeneous models, tasks, and modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 4824–4832.

Li, Y.; Huang, Y.; Yang, B.; Venkitesh, B.; Locatelli, A.; Ye, H.; Cai, T.; Lewis, P.; and Chen, D. 2024b. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37: 22947–22970.

Liu, J.; Chen, S.; He, X.; Guo, L.; Zhu, X.; Wang, W.; and Tang, J. 2025a. VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset. arXiv:2304.08345.

Liu, T.; Shi, L.; Hong, R.; Hu, Y.; Yin, Q.; and Zhang, L. 2024a. Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model. *arXiv preprint arXiv:2411.10803*.

Liu, Y.; Hu, X.; Zhang, S.; Chen, J.; Wu, F.; and Wu, F. 2024b. Fine-Grained Guidance for Retrievers: Leveraging LLMs' Feedback in Retrieval-Augmented Generation. *arXiv preprint arXiv:2411.03957*.

Liu, Y.; Li, P.; Wei, Z.; Xie, C.; Hu, X.; Xu, X.; Zhang, S.; Han, X.; Yang, H.; and Wu, F. 2025b. InfiGUIAgent: A Multimodal Generalist GUI Agent with Native Reasoning and Reflection. In *ICML 2025 Workshop on Computer Use Agents*.

Liu, Y.; Li, P.; Xie, C.; Hu, X.; Han, X.; Zhang, S.; Yang, H.; and Wu, F. 2025c. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. *arXiv preprint arXiv:2504.14239*.

Liu, Z.; Liu, Y.; Zhu, G.; Xie, C.; Li, Z.; Yuan, J.; Wang, X.; Li, Q.; Cheung, S.-C.; Zhang, S.; et al. 2025d. Infi-MMR: Curriculum-based Unlocking Multimodal Reasoning via Phased Reinforcement Learning in Multimodal Small Language Models. *arXiv preprint arXiv:2505.23091*.

Liu, Z.; Yuan, J.; Jin, H.; Zhong, S.; Xu, Z.; Braverman, V.; Chen, B.; and Hu, X. 2024c. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*.

Shi, L.; Zhang, H.; Yao, Y.; Li, Z.; and Zhao, H. 2024. Keep the cost down: A review on methods to optimize llm's kv-cache consumption. *arXiv preprint arXiv:2407.18003*.

Sun, G.; Yu, W.; Tang, C.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; Wang, Y.; and Zhang, C. 2024. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*.

Tang, J.; Zhao, Y.; Zhu, K.; Xiao, G.; Kasikci, B.; and Han, S. 2024a. Quest: Query-Aware Sparsity for Efficient Long-Context LLM Inference. arXiv:2406.10774.

Tang, Y.; Shimada, D.; Bi, J.; Feng, M.; Hua, H.; and Xu, C. 2024b. Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding. *arXiv preprint arXiv:2403.16276*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wan, Z.; Wu, Z.; Liu, C.; Huang, J.; Zhu, Z.; Jin, P.; Wang, L.; and Yuan, L. 2024. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. *arXiv preprint arXiv:2406.18139*.

Wang, X.; Du, Z.; Xu, H.; Yin, S.; Han, Y.; Zhu, J.; Zhang, K.; and Liu, Q. 2025a. Personalized Visual Content Generation in Conversational Systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.-F.; and Wang, W. Y. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4581–4591.

Wang, X.; Wu, L.; Li, Z.; Yuan, H.; Shen, S.; Xu, H.; Su, Y.; and Lei, C. 2025b. Mitigating redundancy in deep recommender systems: A field importance distribution perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 1515–1526.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2024. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Xiao, J.; Chen, Y.; Feng, X.; Wang, R.; and Wu, Z. 2025. RecNet: Optimization for Dense Object Detection in Retail Scenarios Based on View Rectification. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Xie, C.; Cai, S.; Wang, W.; Li, P.; Sang, Z.; Yang, K.; Zhang, Y.; Li, Z.; Zhu, G.; Liu, Z.; et al. 2025. Infir: Crafting effective small language models and multimodal small language models in reasoning. *arXiv preprint arXiv:2502.11573*.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yu, H.; Qu, Z.; Yu, Q.; Chen, J.; Jiang, Z.; Chen, Z.; Zhang, S.; Xu, J.; Wu, F.; Lv, C.; and Yu, G. 2024. GaussianTalker: Speaker-specific Talking Head Synthesis via 3D Gaussian Splatting. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 3548–3557.

Zhang, Z.; Sheng, Y.; Zhou, T.; Chen, T.; Zheng, L.; Cai, R.; Song, Z.; Tian, Y.; Ré, C.; Barrett, C.; et al. 2023. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36: 34661–34710.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.

Zhou, Y.; Han, W.; and Chen, J. 2025. Revisiting applicable and comprehensive knowledge tracing in large-scale data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 235–251. Springer.

Zhou, Y.; Lv, Z.; Zhang, S.; and Chen, J. 2025a. Cuff-KT: Tackling Learners' Real-time Learning Pattern Adjustment via Tuning-Free Knowledge State Guided Model Updating. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 4144–4155.

Zhou, Y.; Lv, Z.; Zhang, S.; and Chen, J. 2025b. Disentangled knowledge tracing for alleviating cognitive bias. In *Proceedings of the ACM on Web Conference 2025*, 2633–2645.

Zhou, Y.; Yao, C.; and Chen, J. 2025. Cola: Collaborative low-rank adaptation. *arXiv preprint arXiv:2505.15471*.