This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

OmniZip: Audio-Guided Dynamic Token Compression for Fast Omnimodal Large Language Models

Keda Tao[1,2,3,†], Kele Shao[1,4,2], Bohan Yu[3], Weiqiang Wang[3], Jian Liu[3,*], Huan Wang[2,*]
Zhejiang University[1], Westlake University[2], Ant Group[3], Shanghai Innovation Institute[4]
https://github.com/KD-TAO/OmniZip

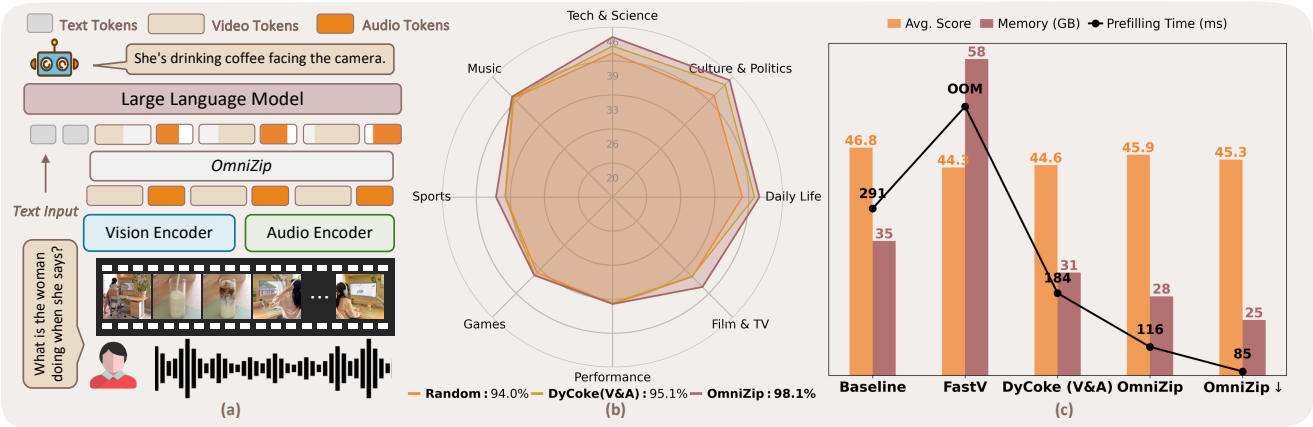

Figure 1. **(a)**: We introduce *OmniZip*, an audio-video token compression method tailored for efficient OmniLLMs. The key innovation is a "*listen-to-prune*" paradigm – utilizing *audio* to dynamically guide video token pruning, complemented by a proposed compression module. **(b)**: OmniZip achieves superior performance on various audio-video tasks on WorldSense [20], outperforming other methods. **(c)**: Efficiency and performance comparison on WorldSense with Qwen2.5-Omni [60]. OmniZip can achieve 2.51-3.42× wall-clock inference speedup (on an A6000 48G GPU), 1.4× memory reduction against other top-performing methods with almost the same performance.

Abstract

Omnimodal large language models (OmniLLMs) have attracted increasing research attention of late towards unified audio-video understanding. However, the high computational cost of processing longer joint audio-video token sequences has become a key bottleneck. Existing token compression methods have not addressed the emerging need to jointly compress multimodal tokens. To bridge this gap, we present ***OmniZip****, a training-free, audio-guided audio-visual token-compression framework that optimizes multimodal token representation and accelerates model inference. Specifically, OmniZip first identifies salient audio tokens, then computes an audio retention score for each time group to capture information density, thereby dynamically guiding video token pruning and preserving cues from audio anchors enhanced by cross-modal similarity. For each time window, OmniZip compresses the video tokens using an interleaved spatio-temporal scheme. Extensive results demonstrate the merits of OmniZip: it achieves a 3.42× inference speedup and a 1.4× memory reduction over other top-performing counterparts, while maintaining the performance of OmniLLMs without training.*

1. Introduction

Video large language models (VideoLLMs) have demonstrated strong performance in video question answering and complex scene understanding [2, 6, 25, 27, 30, 33, 49, 54, 69, 70]. Due to a video inherently containing both visual and auditory streams, recent efforts have begun to focus on *omnimodal large language models* (OmniLLMs) towards unified audio–video understanding [15, 18, 26, 29, 43, 47, 52, 60, 61, 64, 67, 72].

However, OmniLLM inference at scale remains constrained by the computational and memory bottleneck, primarily due to the prohibitively large number of audio-video tokens and the quadratic complexity of attention in large

*Corresponding authors: Huan Wang (wanghuan@westlake.edu.cn), Jian Liu(rex.lj@antgroup.com).
†Work done during internship at Ant Group.

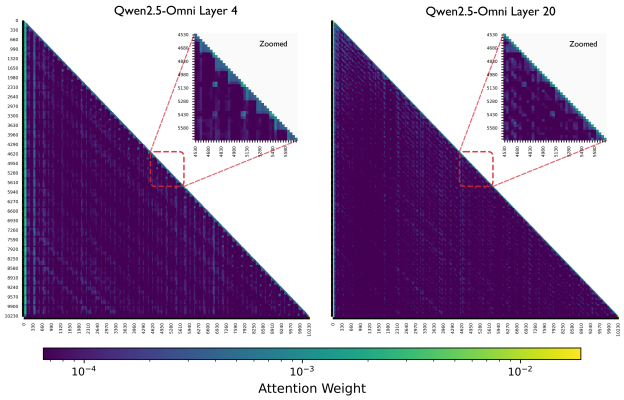

Figure 2. **Audio tokens dominate attention heatmaps.** Regular vertical bands aligned with audio-token positions indicate consistently higher attention to audio tokens, while many video tokens receive little attention, suggesting greater redundancy. Attention aggregates within time windows and decays across windows, indicating that audio and video tokens preferentially attend to short-range context within the same window. Moreover, deeper layers allocate less attention to raw audio and video tokens.

language models [39–41, 48]. Token compression techniques have been a promising approach for facilitating long-sequence inference on multimodal LLMs. Recent works have been investigating token reduction from a *purely visual* perspective [4, 5, 21, 34, 38–42, 46, 48, 59, 63, 65, 68], while for OmniLLMs, the additional *audio* tokens further inflate sequence length and are non-negligible. At the core of technical challenges, audio and video streams exhibit distinct temporal scales and varying sparsity, and the coexistence of redundancy and complementarity renders token pruning particularly sensitive and challenging. As such, *joint audio–video token compression for OmniLLMs remains underexplored so far.*

This work presents, to our knowledge, the first systematic study of reducing tokens under omnimodal inputs, and we propose *OmniZip*, an audio-guided audio-video token compression method for OmniLLMs, as shown in Fig. 1(a). Specifically, we start by performing token attention analyses. In OmniLLM, the token sequence is constructed by segmenting the audio and video streams into fixed-length time windows. Fig. 2 shows regularly recurring vertical bands at audio-token positions, indicating that attention on audio tokens is consistently greater than on video tokens, which suggests the dominance of audio inputs. A magnified view indicates predominantly intra-window attention—mutual attention between audio and video tokens within the same window is most pronounced. This pattern suggests that token compression should operate at the time-window granularity, which differentiates from prior single-modal compression strategies. A detailed analysis appears in Sec. 3.2.

Based on these analyses, OmniZip features three novel technical innovations. *First*, we identify dominant audio tokens and compute the audio retention rate for each time window, which we interpret as time-wise information density and an event-boundary prior. *Second*, windows with high retention are treated as information-dense, and the corresponding video tokens receive a lower pruning rate; conversely, information-sparse windows are assigned higher pruning rates. *Third*, to further preserve multimodal capability, we uniformly sample audio anchors and select secondary audio tokens for merging via cross-modal similarity. For video tokens, we propose an interleaved spatio-temporal token compression method that addresses temporal redundancy between frames and spatial redundancy within frames. This interleaved design suppresses redundancy while avoiding excessive reduction along any single dimension.

Empirically, OmniZip demonstrates strong performance on audio-video understanding tasks, significantly outperforming single-modality token compression methods. As shown in Fig. 1 (c), OmniZip achieves a 2.51× to 3.42× inference speedup on Qwen2.5-Omni-7B [60], all while exhibiting the lowest memory consumption (reducing the GPU memory footprint by 10G) and maintaining the highest accuracy. Crucially, OmniZip is *training-free*.

Our contributions in this work are summarized as follows:

- This work presents, to our knowledge, the first analysis of how audio-video tokens can be pruned to reduce computational overhead in omnimodal settings, and proposes OmniZip, a novel, training-free audio-video token compression framework for OmniLLMs to accelerate inference.
- We propose an audio-guided token compression method, complemented by a proposed video token compression module, to aggressively prune audio-video tokens while preserving cross-modal semantic and temporal alignment.
- Experimental results on several audio-video understanding benchmarks show that OmniZip can compress audio-video tokens while maintaining high inference accuracy, significantly improving inference speed, and reducing memory overhead.

2. Related Work

2.1. Omnimodal Large Language Models

To achieve a more human-like multimodal interaction experience, OmniLLMs have emerged. By leveraging multimodal data, they learn richer contextual information and achieve a deeper understanding of inter-modal relationships [15, 18, 29, 43, 44, 47, 52, 58, 60, 61, 64, 72]. In video understanding tasks, compared to VideoLLMs, OmniLLMs can additionally consider audio information alongside visual data, enabling more realistic answers and a more comprehensive understanding. Recent work, such as Qwen2.5-Omni [60], introduced an end-to-end model capable of perceiving all modalities. While InteractiveOmni [52] has en-

abled multi-round audio-video conversations, significant recent work [1, 61, 64, 67] has further advanced state-of-the-art omnimodal understanding capabilities. However, the large number of multimodal tokens introduced by video and audio inputs significantly impedes the practical deployment and application of OmniLLMs. Balancing model performance and computational efficiency remains a significant challenge. Thus, developing efficient methods to simplify the derivation of tokenized audio-video information is essential.

2.2. Token Compression

Recent research has focused on token compression to enhance the inference efficiency of multimodal large language models. This approach is highly effective as multimodal inputs often contain significant redundancies, such as image [3, 4, 38, 46, 59, 63, 65, 68], video [5, 21, 39, 41, 42, 48], and audio [23, 28, 32, 44]. A key advantage is that these methods can be applied as a tuning-free, post-processing technique. These methods operate by first establishing a metric to evaluate token importance, followed by corresponding compression operations [40]. While token compression methods for single modalities have been widely studied, their application to the omnimodal setting has not yet been explored. Furthermore, current mainstream methods typically depend on accessing the attention matrices from either the video encoder or the LLM [19, 39, 48, 59, 65]. This dependency is often incompatible with modern optimizations such as FlashAttention [7, 8], necessitating the materialization of the full attention matrix. In conjunction with ultra-long visual token sequences, this readily leads to Out-of-Memory (OOM) errors. Therefore, such methods exhibit poor scalability to larger, more advanced models. Considering the inherent coupling of video and audio, we conduct the first exploration of token compression for the combined audio-video understanding task, aiming to facilitate the practical deployment of OmniLLMs.

3. Proposed Method

In this section, we first describe the overall architecture of OmniLLMs (Sec. 3.1), and then present the analyses based on the token attention distributions (Sec. 3.2). Next, we detail our proposed method, OmniZip (Sec. 3.3). Then, the ISTC module for video-token pruning is introduced. Fig. 3 illustrates the overall architecture. Finally, we further remark on the design concept of our method in Sec. 3.5.

3.1. Background on OmniLLM

OmniLLMs aim to ingest a full range of modalities together with human-provided prompts to form a unified audio–video understanding. Such models typically comprise a vision encoder, an audio encoder, a projector, and an LLM backbone. Given a video, we first decompose it into individual video frames clip $X_{\text{vid}} \in \mathbb{R}^{T\times H\times W\times 3}$ and audio segments X_{aud} sampled at fixed rates, where T is the number of frames after sampling. The vision encoder g_v and audio encoder g_a convert the raw video clip and audio clip into a sequence of token embeddings:

$$\mathbf{Z}_{\text{v}} = g_v(X_{\text{vid}}), \quad \mathbf{Z}_{\text{a}} = g_a(X_{\text{aud}}), \tag{1}$$

where $\mathbf{Z}_{\text{v}} \in \mathbb{R}^{N_v\times D}$, $\mathbf{Z}_{\text{a}} \in \mathbb{R}^{N_a\times D}$, N_a and N_v are the number of audio tokens and video tokens, respectively. Then, the projector maps audio-video tokens into the LLM's embedding space, enabling the model to process multimodal inputs effectively. Typically, a video yields 10–20k tokens (audio and video), severely constraining efficient deployment.

Furthermore, the stitching for audio–video tokens is organized by fixed-length time windows, as shown in Fig. 3. The audio and video streams are segmented into multiple windows of equal duration. Within each window, co-temporal multimodal tokens are aligned and concatenated into a cross-modal block; the blocks are then concatenated chronologically to form a long token sequence and fed to the LLM. The LLM jointly aligns video, audio, and textual representations to generate a response.

3.2. Token Attention Analysis

To characterize redundancy and attention patterns in audio and video tokens during inference, we visualize the attention distribution, as shown in Fig. 2. First, most tokens receive a low attention score, and the attention to both video and audio tokens decreases with layer depth, indicating that judicious token pruning can preserve model reasoning while reducing memory usage and accelerating inference.

Then, we investigate how to design an effective token compression strategy. First, we observe regularly recurring bright bands in the attention heatmap. Cross-referencing with token indices shows that these bands align with audio tokens in each time window. This indicates that audio tokens are consistently assigned greater attention than video tokens across layers, whereas large regions of video tokens exhibit significantly lower attention scores, suggesting substantial redundancy and the dominant role of audio tokens in the inference process. Magnified views reveal block-structured local attention: tokens cluster strongly within the same time window but decay rapidly across windows, indicating a strong locality for short-range temporal dependence. This motivates us to design OmniZip to perform token pruning separately within each time window.

Building on these observations, we design an audio-guided dynamic compression strategy for audio-video tokens. Specifically, after selecting retained audio tokens, we treat per-window audio retention as a proxy for information density and event-boundary likelihood, and we dynamically allocate the video pruning rate for each time window accordingly, while constraining the video compression to exceed the audio compression. This reduces the number of tokens

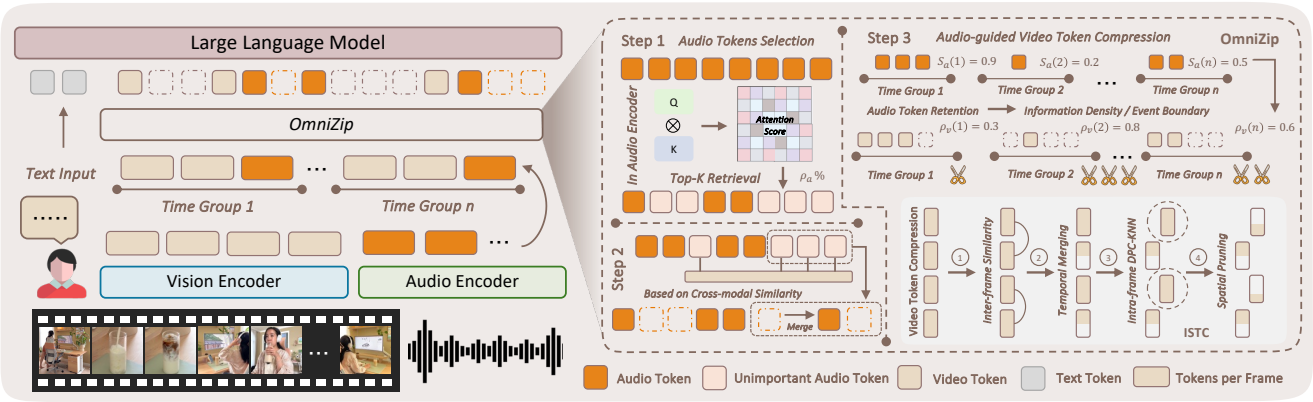

Figure 3. **Detailed overview of our OmniZip method.** First, OmniZip computes an audio retention rate derived from dominant audio tokens to determine a dynamic pruning rate for the corresponding video tokens. Next, to preserve multimodal information, we uniformly sample audio anchors and merge with non-anchor tokens selected via cross-modal similarity. Finally, video tokens undergo interleaved spatio-temporal compression (ISTC), which alternately reduces temporal redundancy by merging cross-frame tokens and spatial redundancy by pruning intra-frame tokens. ρ_a is the compression ratio of the audio token, $S_a(i)$ and $\rho_v(i)$ are the audio token retention ratio and video token compression ratio, in each time group, respectively.

processed downstream while preserving performance, substantially lowering computational and memory costs.

3.3. Our Method: OmniZip

OmniZip is a *training-free*, inference-time compressor that selects and restructures audio–video tokens before feeding them to the LLM. As shown in Fig. 3, it proceeds window-by-window and contains three stages: (i) audio token selection, (ii) audio anchor consolidation, and (iii) audio–guided dynamic video compression. Let the t-th time window contain n_a audio tokens and n_v video tokens, with embeddings $H_a^{t,i}, H_v^{t,j}$ after the projectors.

Audio Token Selection. We filter audio tokens based on the attention distribution produced by the audio encoder. Specifically, we use the last layer of the audio encoder g_a and compute the attention matrix:

$$A = \mathrm{Softmax}(QK^T/\sqrt{d}) \in \mathbb{R}^{B\times N_a\times N_a}, \tag{2}$$

where $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{N_a\times d}$ are the query and key matrices for the audio tokens, and d is the state dimension. We quantify token importance as the mean attention each audio token receives from all other audio tokens, yielding a per-token score vector $a_{avg} \in R^{B\times N_a}$. Tokens with larger mean-attention scores are considered more salient. Because many models pool audio tokens, we apply the same average-pooling operation to a_{avg} to maintain alignment with the pooled audio indices, producing an importance map. Finally, we select the audio features with the highest attention scores ($\rho_a\%$) as the representative and information-dense tokens, while treating other tokens as non-significant.

Audio Anchor Consolidation. Considering the importance and pruning sensitivity of audio tokens, we merge a subset of non-salient tokens, thereby preserving semantic salience while maintaining context coverage. Specifically, for each time window, we uniformly sample anchors from the non-salient audio tokens. To maintain multimodal consistency, we evaluate candidates using cross-modal similarity between audio and video tokens:

$$\mathbf{S_{cross}} = \hat{H}_a\hat{H}_v^{\top}, S_{ij} = \hat{h}_{ai}^{\top}\hat{h}_{vj} \in [-1, 1], \tag{3}$$

where $\hat{H}_a$, $\hat{H}_v$ denote the normalized audio token and video token sequences, respectively:

$$\hat{H} = \mathrm{Diag}\left(\sqrt{\mathrm{diag}(HH^{\top})} + \varepsilon\right)^{-1} H, \varepsilon = 10^{-6}. \tag{4}$$

Then, we select the top-$\mathcal{G}$ audio tokens most related to the paired video segment and merge them into the anchor, where $\mathcal{G}$ is the number of merging tokens for each anchor. Finally, the remaining non-salient tokens are discarded.

Audio-Guided Video Token Compression. In prior single-modal token-pruning work, it is hard to assess whether key information and events occur between frames [39, 41, 48]. However, in OmniLLMs, introducing audio tokens is both challenging and beneficial. We set the total video token pruning ratio as ρ_v. After filtering audio tokens, we map scores back to time windows and compute a per-window audio-retention score $S_a(i) \in [0, 1]$, and i is the index of the time group. Windows with high retention are deemed significant—providing information-dense, event-boundary cues. We dynamically prune video tokens: high-saliency windows are pruned conservatively, whereas low-saliency windows are pruned more aggressively. Thus, we get the initial ratios $\rho_v'(i)$:

$$\rho_v'(i) = \rho_{max} - (\rho_{max} - \rho_{min}) \cdot S_a(i), \tag{5}$$

where ρ_{max} and ρ_{min} are the upper and lower limits of the pruning rate set to prevent excessive pruning. These initial

Method	Settings		AVUTBench							VideoMME	ShortVid-Bench	Avg.
	Retained Ratio	FLOPs Ratio	EL	OR	OM	IE	CC	CM	Avg.	wo	Avg. Score	
Qwen2.5-Omni-7B												
Full Tokens	100%	100%	38.2	67.8	59.6	85.6	44.1	66.7	64.5	66.0	70.5	100%
Random	55%	48%	38.2	64.9	55.6	80.1	34.7	65.0	61.0	65.4	68.3	96.9%
FastV	50%	54%	34.1	64.3	57.1	77.6	36.4	56.4	58.4	-	68.0	94.3%
DyCoke (V&A)	50%	44%	**38.8**	**67.2**	**58.2**	81.9	39.0	62.4	62.0	65.5	68.5	97.5%
OmniZip (Ours)	45%	39%	38.4	**67.2**	56.9	**85.3**	**42.4**	**66.0**	**63.0**	**66.3**	**69.9**	**99.1%**
Random	40%	34%	31.7	58.5	53.3	74.9	43.2	59.0	56.9	65.0	67.7	94.3%
FastV	35%	42%	24.1	60.7	54.3	81.6	40.7	58.3	57.8	-	67.9	93.8%
DyCoke (V&A)	35%	29%	32.9	62.1	**54.9**	74.5	39.0	58.3	57.4	65.2	68.0	94.7%
OmniZip (Ours)	35%	29%	**34.1**	**67.5**	54.6	**83.7**	**42.4**	**61.2**	**61.0**	**66.1**	**69.0**	**97.6%**
Qwen2.5-Omni-3B												
Full Tokens	100%	100%	32.9	65.3	58.4	85.0	44.1	62.6	62.2	62.6	69.4	100%
Random	55%	45%	31.7	59.2	55.4	77.3	**44.9**	**62.1**	58.7	61.1	67.9	96.6%
FastV	50%	49%	27.1	57.0	56.3	80.5	42.3	60.1	55.9	-	68.0	95.7%
DyCoke (V&A)	50%	40%	31.9	64.3	57.3	82.2	40.7	61.3	60.7	61.6	67.4	97.7%
OmniZip (Ours)	45%	36%	**32.4**	**65.0**	**57.7**	**84.9**	41.5	61.4	**61.3**	**62.8**	**68.5**	**99.2%**
Random	40%	31%	28.2	60.8	54.9	73.1	42.3	**61.6**	57.5	60.6	67.0	95.4%
FastV	35%	37%	24.2	60.8	54.3	81.6	40.7	58.3	57.7	-	67.7	96.9%
DyCoke (V&A)	35%	26%	32.9	62.1	54.9	74.5	38.9	58.3	57.4	61.0	67.5	95.7%
OmniZip (Ours)	35%	26%	**28.8**	**63.1**	**58.2**	**84.0**	**42.4**	60.4	**60.1**	**62.7**	**68.0**	**98.3%**

Table 1. **Comparison of different methods on omnimodal (audio & video) QA benchmarks**. The **best** result among token pruning methods for each metric is in bold, and the second-best is underlined. The '-' symbol indicates that FastV fails to execute due to an Out-of-Memory (OOM) error, and we also ignore its value when calculating the average score. The 'DyCoke (V&A)' label denotes the application of its TTM module [48] to both audio and video tokens.

ratios $\rho'_v(i)$ are then algorithmically normalized to ensure the final rates ρ_v strictly adhere to the global pruning budget. Overall, audio pruning remains more conservative, while video pruning is time-adaptive. This audio-guided strategy preserves key frames and temporal-alignment cues without additional training, while substantially reducing the total token count and inference overhead.

3.4. ISTC Block

In this section, we describe the interleaved spatio-temporal compression (ISTC) module used in OmniZip. Video token pruning is performed independently within each time window, and we set the minimum processing unit to four frames. We interleave temporal-spatial redundancy evaluation for each frame and apply the corresponding strategies to compress tokens. As shown in Fig. 3, we first compute cosine similarity between same-position tokens in adjacent frames:

$$\mathbf{S_{vid}} = \cos(\theta) = \frac{h_v^i \cdot h_v^j}{\|h_v^i\| \|h_v^j\|}, \tag{6}$$

and use $\mathbf{S_{vid}}$ to estimate temporal redundancy and prune tokens in frames 2 and 4 with high similarity. For tokens in frames 1 and 3, we apply cluster-based pruning via density-peak clustering with k-nearest neighbors (DPC-KNN) [10]. For each video token h_v^i, we compute each token's local density ρ_i and its distance δ_i to the nearest higher-density token, yielding the final density score $\delta_i \times \rho_i$.

$$\rho_i = \exp\left(-\frac{1}{k} \sum_{h_v^j \in \mathrm{kNN}(h_v^i)} d(h_v^i, h_v^j)^2\right), \tag{7}$$

$$\delta_i = \begin{cases} \max\limits_{j \neq i} d(h_v^i, h_v^j), & \text{if } \rho_i = \max\limits_k \rho_k, \\ \min\limits_{j:\ \rho_j > \rho_i} d(h_v^i, h_v^j), & \text{otherwise.} \end{cases}, \tag{8}$$

where $d(\cdot)$ is the euclidean distance. We prune tokens based on the density score, retaining salient video tokens and discarding spatially redundant ones.

3.5. Further Remarks on Our Method Design

In this section, we analyzed the common limitations in prior work and further remark on our method design. To our knowledge, OmniZip is the first token-compression framework for OmniLLMs in the audio–video understanding setting. In its design, we align with current developments in multimodal large language models and incorporate insights from prior work. First, our method does not require accessing attention-score matrices inside the LLM, enabling compatibility with FlashAttention [7, 8] without incurring additional compute or memory overhead [4, 19, 39]. It also preserves multi-round dialogue capability and remains compatible with other inference frameworks. Second, because most mainstream models now adopt ViT-based visual encoders, methods such as VisionZip can trigger GPU memory overflow when extracting attention-score matrices [39, 65]; our approach avoids this issue. By contrast, the audio encoder is comparatively lightweight. Finally, the additional runtime cost of token pruning is a common concern: OmniZip's pruning step takes less than 40 ms, making it lightweight and not slowing inference.

Method	Retained Ratio	FLOPs (T)	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.
Qwen2.5-Omni-7B											
Full Tokens	100%	73.2	52.4	50.1	48.5	44.6	43.8	41.6	41.6	47.3	46.8
Random	55%	35.5	47.1	47.0	44.4	41.2	40.0	40.1	40.1	46.3	43.6
FastV	50%	39.3	48.8	47.4	44.2	44.1	**41.2**	38.3	40.0	46.6	44.3
DyCoke (V&A)	50%	31.9	48.4	49.9	46.7	41.4	39.9	**40.8**	40.2	46.5	44.6
OmniZip (Ours)	45%	28.3	**49.4**	**51.1**	**45.6**	**43.9**	40.1	**40.8**	41.9	**46.7**	**45.9**
OmniZip (Ours)	35%	21.4	48.3	49.5	**47.6**	42.5	40.1	40.2	**42.3**	46.3	45.3
Qwen2.5-Omni-3B											
Full Tokens	100%	37.4	51.5	50.8	45.0	45.4	43.8	42.5	44.2	46.1	46.4
Random	55%	17.0	48.2	46.3	40.7	41.4	38.6	40.0	41.8	**43.4**	42.8
FastV	50%	18.2	50.0	**50.5**	**44.1**	43.0	**40.5**	41.6	41.8	42.1	44.4
DyCoke (V&A)	50%	15.1	48.1	48.5	42.3	43.3	39.7	**43.4**	42.1	43.0	44.0
OmniZip (Ours)	45%	13.3	**50.1**	**50.5**	43.9	45.6	**40.5**	40.8	**43.7**	43.1	**45.2**
OmniZip (Ours)	35%	9.9	48.8	48.9	41.8	**46.4**	39.8	42.5	42.6	43.1	44.3

Table 2. **Comparison of different methods on the WorldSense benchmark**. The **best** result among token pruning methods for each metric is in bold, and the second-best is underlined. The FLOPs calculation considers only the multimodal tokens originating from audio and video inputs. FastV failed to run on the 7B model due to an OOM error on an A6000 GPU, so we evaluated its performance on a single H100 (80G) GPU.

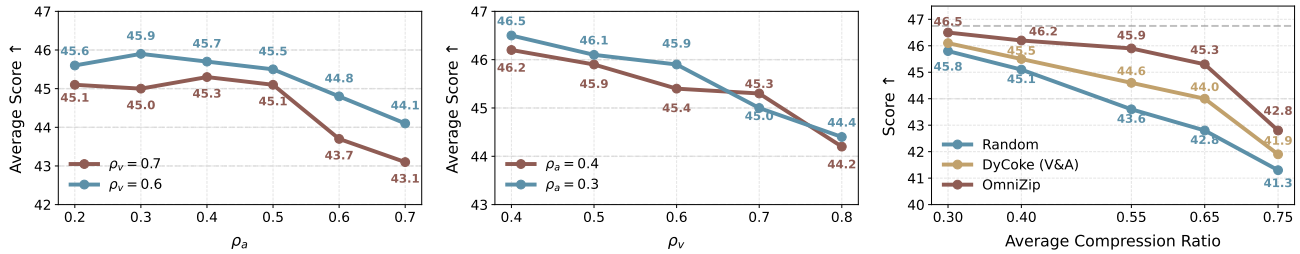

Figure 4. **Ablation study on ρ_a and ρ_v.** All experiments illustrated in the figure were carried out on the Qwen2.5-Omni-7B model and the WorldSense benchmark. **Left and Middle:** We separately analyze the influence of varying ρ_a and ρ_v on model performance. In general, excessive pruning of either modality negatively impacts model performance. However, an appropriate balance of audio and video token pruning achieves the best effect. **Right:** Performance of our method *vs.* other methods in different compression ratios.

4. Experimental Results

4.1. Evaluation Setups and Implementation Details

Benchmarks. We evaluate the performance of OmniLLMs using established audio-video understanding benchmarks: AVUT [66], VideoMME [16], ShortVid-Bench [18], and WorldSense [20]. Among these benchmarks, VideoMME is widely used for pure video-understanding evaluations, and including audio can improve accuracy. AVUT is an *audio-centric* video understanding benchmark focusing on six tasks: event localization (EL), object matching (OM), OCR matching (OR), information extraction (IE), content counting (CC), and character matching (CM). WorldSense assesses models' ability to understand over audio and video across eight domains jointly. ShortVid-Bench evaluates the ability of models to understand real-world short videos.

Comparison Methods. Given the absence of token pruning methods specifically designed for the omnimodal setting, we select representative prior methods from single-modal domains for adaptation and comparative analyses. FastV [4], during its prefill stage, utilizes the attention score matrix of the L-th layer to evaluate token relevance, subsequently pruning tokens. DyCoke [48] represents the first dynamic token compression strategy proposed for VideoLLMs. We employ its first-stage TTM module to process video and audio tokens. Furthermore, we implement a random pruning as a control group to provide a rigorous comparative analysis.

Implementation Details. We implement the proposed OmniZip on the Qwen2.5-Omni (7B and 3B) models using NVIDIA A6000 (48GB) GPUs [60]. To set pruning ratios across methods, we use the overall FLOPs ratio as the metric to ensure a fair comparison. For FastV, we set the attention-computation layer to layer 5. For video input, to better match the time-window granularity—and given that VideoMME videos are relatively long—we cap the maximum number of frames at 768. For other datasets, we cap inputs at 128 frames. For each time window, it has 50 audio tokens and 288 video tokens. For hyperparameter settings, we set $\rho_{max} = 0.75$, $\rho_{min} = 0.35$, $k = 5$ and $\mathcal{G} = 15$ for AVUT and $\mathcal{G} = 3$ for others. For 45% and 35% retained ratio, we set $\rho_a = 0.3, \rho_v = 0.6$ and $\rho_a = 0.4, \rho_v = 0.7$ respectively, except for ShortVid-Bench. For all experiments, we leverage FlashAttention to reduce memory usage.

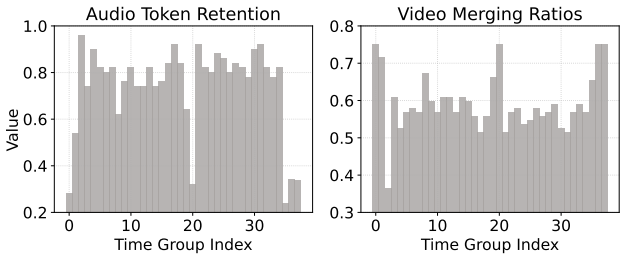

Figure 5. **Visualization of dynamic pruning ratios.** The figure illustrates how audio token retention guides the allocation of video token pruning. Specifically, for time windows with low audio retention, we allocate a higher video pruning ratio, while maintaining a constant total pruning rate.

4.2. Main Results

We evaluate our approach on recent mainstream models Qwen2.5-Omni at two parameter scales (7B and 3B). For the VideoMME, we use the LMMs-Eval [24, 71] for evaluation, and for other benchmarks, we follow the unified testing code for all experimental settings. We evaluated performance and inference cost at two distinct token retention rates. To facilitate a comprehensive evaluation, the results in Tab. 1 are normalized and presented as percentages, where the baseline model's accuracy is set to 100%. Notably, unlike conventional purely video understanding tasks, audio-video understanding tasks present greater *challenges* and exhibit *increased sensitivity* to token pruning.

Comparison with State-of-the-Art Methods. As shown in Tab. 1, the results indicate that OmniZip maintains optimal performance with the fewest tokens across diverse test benchmarks. Even with a 60% reduction in computational FLOPs, the model retains an average accuracy of 99.1%. In contrast, the random pruning leads to significant performance degradation. FastV similarly fails to achieve effective results, a limitation attributable to the uneven attention distribution between video and audio tokens and the consequent disruption of temporal windows. DyCoke is designed to reduce redundancy in the temporal dimension while preserving the time window structure. However, as it is designed for single-modal video and neglects spatial redundancy, its omnimodal performance is suboptimal. At lower retention rates, OmniZip maintains its leading performance. Besides, as shown in Tab. 2 for the WorldSense Benchmark, OmniZip at a 35% token retention rate outperforms other methods operating at a 50% retention rate.

Furthermore, our experiments across different model scales reveal that models with fewer parameters are more amenable to compression, corroborating prior studies [39, 48]. We also note that the missing FastV results for the 7B model are attributable to its incompatibility with Flash Attention, which requires the explicit calculation of the attention matrix and subsequently causes an out-of-memory (OOM) error. We circumvent this problem in our method design.

Method	GPU Mem. ↓	Prefilling Time ↓	Acc. ↑	Latency per Example ↓
Qwen2.5-Omni-7B				
Full Tokens	35G	291ms (1.00×)	46.8	4.52s (1.00×)
FastV	OOM			
DyCoke (V&A)	31G	184ms (1.58×)	44.6	3.64s (1.24×)
Ours (45%)	28G	116ms (2.51×)	45.9	3.40s (1.33×)
Ours (35%)	**25G**	**85ms (3.42×)**	45.3	**3.18s (1.42×)**
Qwen2.5-Omni-3B				
Full Tokens	25G	258ms (1.00×)	46.4	3.61s (1.00×)
FastV	45G	222ms (1.16×)	44.4	3.45s (1.05×)
DyCoke (V&A)	20G	171ms (1.51×)	44.0	3.12s (1.16×)
Ours (45%)	17G	104ms (2.48×)	45.2	2.86s (1.26×)
Ours (35%)	**16G**	**79ms (3.27×)**	44.3	**2.75s (1.31×)**

Table 3. **Actual inference efficiency comparison on WorldSense**. Experiments with the 7B and 3B models are conducted on a single A6000 GPU. FastV computes the full attention matrix in memory, a process that results in Out-of-Memory (OOM) errors attributable to the large number of tokens. Our method can achieve the best model performance, the lowest memory consumption, and the greatest inference acceleration.

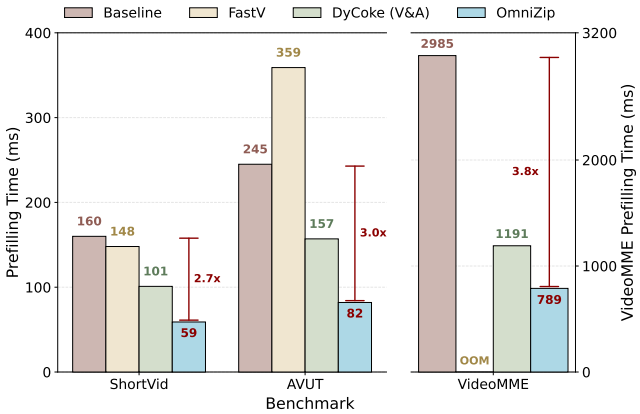

Figure 6. **Achieving superior inference speedup.** We visualize the inference speedup achieved by OmniZip during the prefilling stage on the 7B model. As video sequence length increases, the speedup effect becomes more pronounced. OmniZip achieves a 2.7–3.8× inference speedup while robustly maintaining model accuracy.

Sensitivity Analyses on ρ_a and ρ_v. As illustrated in the left and middle plots of Fig. 4, excessive pruning of either audio or video significantly degrades model performance. This suggests that due to the varying redundancy and attention given to audio and video tokens, identifying an optimal pruning ratio is crucial for maximizing compression effectiveness, a conclusion supported by the data. Thus, we suggest that the pruning rate can be dynamically adapted based on the specificity of the task, such as its relative dependence on video or audio information. Moreover, our results suggest that the audio token pruning rate should be lower than the video token pruning rate. Finally, the right plot indicates that OmniZip outperforms other methods across all pruning ratios. As the pruning rate increases, our accuracy declines more gradually, highlighting the robustness of our method.

Visualization of Dynamic Pruning. Fig. 5 visualizes the dynamic allocation of pruning rates in OmniZip, illustrating

ID	Select Method			AVUT	WorldSense	ShortVid-Bench
	Video	Audio	GS	Avg.	wo	Avg. Score
Baseline	-	-	-	64.5	46.8	70.5
1	ISTC	Random	✗	60.0	45.1	69.0
2	DyCoke	Ours	✗	62.1	45.0	69.2
3	VisionZip	Ours	✓	61.4	44.2	68.0
4	Random	Random	✓	60.4	43.3	68.1
OmniZip	ISTC	Ours	✗	**63.0**	**45.9**	**69.9**

Table 4. **Ablation study of the token selection method.** We compare our token selection method against baseline strategies on 7B model. Furthermore, to substantiate the design rationale of OmniZip, we compare it against VisionZip [65], a method that performs global video token selection (GS).

Settings			AVUT	WorldSense	ShortVid-Bench
Re. Ratio	DP	AC	Avg.	Avg.	Avg.
100%	-	-	64.5	46.8	70.5
45%	✓	✓	63.0	45.9	69.9
45%	✗	✓	62.0 (-1.0)	45.0 (-0.9)	69.3 (-0.6)
45%	✗	✗	61.7 (-1.3)	44.8 (-1.1)	69.0 (-0.9)

Table 5. **Ablation study of DP & AC Technology.** To validate the efficacy of our method, we conduct an ablation study evaluating the impact of our two key components on final model accuracy: audio-guided dynamic video pruning (DP) and audio anchor consolidation (AC) on Qwen2.5-Omni-7B.

that across different time windows, the pruning rate of video tokens changes dynamically in conjunction with that of audio tokens. Our method employs a dynamic pruning rate while simultaneously maintaining a constant overall pruning rate, which facilitates a fair comparison against other methods. Collectively, this finding demonstrates the efficacy of our proposed approach; it also underscores the necessity of developing specialized research for the OmniLLMs.

4.3. Efficiency Analyses

We evaluated the inference speed and memory consumption across four benchmarks. As shown in Tab. 3, we conducted more detailed analyses on the WorldSense benchmark. The results indicated that our method significantly accelerated inference speed compared to the full token model. On the 3B model, our method achieves 3.27× speedup in the prefilling stage. This advantage became more pronounced for larger models (7B), yielding a 1.42× speedup in overall inference and a 3.42× speedup in prefilling. Moreover, our method significantly reduces the memory cost during inference. While maintaining an accuracy of approximately 97%, the method reduces memory consumption by *10G*, which is crucial for the practical deployment of OmniLLMs.

Furthermore, Fig. 6 summarizes the inference speedup on other benchmarks. Our method significantly reduces the prefilling stage time. In contrast, FastV is incompatible with Flash Attention due to its requirement for explicit attention matrix computation, which incurs extra overhead and consequently slows inference. Furthermore, due to inherent dataset characteristics (i.e., ShortVid comprises shorter videos while VideoMME features longer ones), the speedup on VideoMME is correspondingly more pronounced. Compared to the baseline, OmniZip achieves a 2.7–3.8× inference speedup, the highest among all methods.

4.4. Ablation Study

Ablation Study of DP and AC Technology. Tab. 5 presents an ablation study on the two core components of the OmniZip framework: dynamic video pruning (DP) and audio anchor consolidation (AC). As shown in the table, removing the dynamic pruning allocation for video tokens significantly decreases model accuracy. Further eliminating the audio anchor consolidation strategy leads to an additional performance degradation. This result validates the efficacy and design rationale of OmniZip.

Ablation Study about Token Selection Method. Tab. 4 presents a comparative analysis of different token selection strategies. First, Tab. 4 (ID:2) demonstrates the superior performance of ISTC over DyCoke for video tokens. Furthermore, VisionZip [65], a global token selection (GS) strategy, is included in the comparison. The results indicate that the GS strategy is suboptimal for the omnimodal setting. The GS strategy extracts focused video and audio tokens independently, ignoring semantic alignment and disrupting the temporal structure, making it difficult to maintain model accuracy. Notably, the additional computation required by VisionZip to compute the visual attention matrix frequently causes OOM, a limitation that OmniZip avoids. Besides, a comparison against random selection underscores the effectiveness of our method. Therefore, our method represents a specialized design that accounts for the characteristics of multimodal information, offering clear advantages over prior single-modal token compression methods.

5. Conclusion

This paper presents *OmniZip*, a novel *training-free* method to dynamically reduce the audio-video tokens based on audio-guidance for faster omnimodal large language models (OmniLLMs). Specifically, the framework first identifies salient audio tokens and calculates an audio retention rate for each time window, which is then used to dynamically guide the pruning of video tokens in conjunction with a corresponding spatio-temporal compression module. To the best of our knowledge, this is the first token pruning method tailored to OmniLLMs that jointly optimizes the compression of multimodal audio-video tokens. Extensive benchmark and analysis results on a wide range of audio-video understanding tasks with two OmniLLMs (3B, 7B parameters) demonstrate that our method consistently surpasses prior single-modal methods. Our method achieves up to a 10G memory reduction and a 2.7–3.8× prefill speedup, while maintaining nearly identical performance.

Acknowledgement

This work was supported by the Ant Group Research Intern Program, Young Scientists Fund of the National Natural Science Foundation of China (NSFC) (No. 62506305), Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (No. 2024R01007), Key Research and Development Program of Zhejiang Province (No. 2025C01026), Scientific Research Project of Westlake University (No. WU2025WF003), Chinese Association for Artificial Intelligence (CAAI) & Ant Group Research Fund - AGI Track (No. 2025CAAI-ANT-13). It is also supported by the research funds of the National Talent Program and Hangzhou Municipal Talent Program.

References

[1] Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025. 3, 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2

[3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3, 2

[4] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, 2024. 2, 3, 5, 6

[5] Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*, 2025. 2, 3

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 1, 2

[7] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024. 3, 5, 2

[8] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5, 2

[9] 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. *arXiv preprint arXiv:2503.00540*, 2025. 3

[10] Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016. 5

[11] Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025. 2

[12] Sicheng Feng, Kaiwen Tuo, Song Wang, Lingdong Kong, Jianke Zhu, and Huan Wang. Rewardmap: Tackling sparse rewards in fine-grained visual reasoning via multi-stage reinforcement learning. *arXiv preprint arXiv:2510.02240*, 2025. 2

[13] Sicheng Feng, Song Wang, Shuyi Ouyang, Lingdong Kong, Zikai Song, Jianke Zhu, Huan Wang, and Xinchao Wang. Can mllms guide me home? a benchmark study on fine-grained visual reasoning from transit maps. *arXiv preprint arXiv:2505.18675*, 2025. 2

[14] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *ICML*, 2023. 3

[15] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 1, 2

[16] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025. 6

[17] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. 1

[18] Yuying Ge, Yixiao Ge, Chen Li, Teng Wang, Junfu Pu, Yizhuo Li, Lu Qiu, Jin Ma, Lisheng Duan, Xinyu Zuo, et al. Arc-hunyuan-video-7b: Structured video comprehension of real-world shorts. *arXiv preprint arXiv:2507.20939*, 2025. 1, 2, 6

[19] Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Zipvl: Efficient large vision-language models with dynamic token sparsification. *arXiv preprint arXiv:2410.08584*, 2024. 3, 5, 2

[20] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omni-modal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025. 1, 6

[21] Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *ACL*, 2025. 2, 3

[22] Xin Jin, Siyuan Li, Siyong Jian, Kai Yu, and Huan Wang. Mergemix: A unified augmentation paradigm for visual and multi-modal understanding. *arXiv preprint arXiv:2510.23479*, 2025. 2

[23] Taehan Lee and Hyukjun Lee. Token pruning in audio transformers: Optimizing performance and decoding patch importance. *arXiv preprint arXiv:2504.01690*, 2025. 3, 2

[24] Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Accelerating the development of large multimodal models, 2024. 7

[25] 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. *TMLR*, 2025. 1, 2

[26] Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025. 1

[27] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1, 2

[28] Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. Accelerating transducers through adjacent token merging. In *Interspeech*, 2023. 3, 2

[29] Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2024. 1, 2

[30] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, 2024. 1, 2

[31] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *NeurIPS*, 2024. 3

[32] Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai Li, Yiran Chen, et al. Speechprune: Context-aware token pruning for speech information retrieval. In *ICME*, 2025. 3, 2

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 2

[34] Jinming Liu, Junyan Lin, Yuntao Wei, Kele Shao, Keda Tao, Jianguo Huang, Xudong Yang, Zhibo Chen, Huan Wang, and Xin Jin. Revisiting mllm token technology through the lens of classical visual coding. *arXiv preprint arXiv:2508.13460*, 2025. 2

[35] Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *EMNLP*, 2025. 2

[36] Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*, 2024. 3

[37] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *NeurIPS*, 2024. 3

[38] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *ICCV*, 2025. 2, 3

[39] Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025. 2, 3, 4, 5, 7

[40] Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *arXiv preprint arXiv:2507.20198*, 2025. 3, 2

[41] Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. Fastvid: Dynamic density pruning for fast video large language models. *arXiv preprint arXiv:2503.11187*, 2025. 2, 3, 4

[42] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In *ICML*, 2025. 2, 3

[43] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. In *CVPR*, 2025. 1, 2

[44] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024. 2, 3

[45] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023. 3

[46] Xudong Tan, Peng Ye, Chongjun Tu, Jianjian Cao, Yaoxin Yang, Lin Zhang, Dongzhan Zhou, and Tao Chen. Tokencarve: Information-preserving visual token compression in multimodal large language models. *arXiv preprint arXiv:2503.10501*, 2025. 2, 3

[47] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025. 1, 2

[48] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *CVPR*, 2025. 2, 3, 4, 5, 6, 7

[49] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 1, 2

[50] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 2

[51] Qwen Team. Qwen3 technical report, 2025. 2

[52] Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, et al. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. *arXiv preprint arXiv:2510.13747*, 2025. 1, 2

[53] Mart Van Baalen, Andrey Kuzmin, Ivan Koryakovskiy, Markus Nagel, Peter Couperus, Cedric Bastoul, Eric Mahurin, Tijmen Blankevoort, and Paul Whatmough. Gptvq: The blessing of dimensionality for llm quantization. *arXiv preprint arXiv:2402.15319*, 2024. 3

[54] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2

[55] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2

[56] Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023. 3

[57] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *ICML*, 2023. 3

[58] Zhifei Xie and Changqiao Wu. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*, 2024. 2

[59] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *CVPR*, 2025. 2, 3

[60] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 1, 2, 6

[61] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 1, 2, 3

[62] Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. Streamingvlm: Real-time understanding for infinite video streams. *arXiv preprint arXiv:2510.09608*, 2025. 3

[63] Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *CVPR*, 2025. 2, 3

[64] Qize Yang, Shimin Yao, Weixuan Chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *arXiv preprint arXiv:2506.21277*, 2025. 1, 2, 3

[65] 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. In *CVPR*, 2025. 2, 3, 5, 8

[66] Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *EMNLP*, 2025. 6, 3

[67] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025. 1, 3, 2

[68] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *AAAI*, 2025. 2, 3

[69] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1, 2

[70] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP*, 2023. 1, 2

[71] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 7

[72] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 1, 2

[73] Junhan Zhu, Hesong Wang, Mingluo Su, Zefang Wang, and Huan Wang. Obs-diff: Accurate pruning for diffusion models in one-shot. *arXiv preprint arXiv:2510.06751*, 2025. 3