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

Supplementary Material

A. Dynamic Pruning Rate Allocation Algorithm

This section expands upon the audio-guided video token compression algorithm described in Sec. 3.3. Algorithm 1 defines the calculation for the dynamic pruning rate and illustrates that while this rate is adaptive, the overall pruning rate remains constant.

Algorithm 1 Audio-guided Video Token Pruning

Parameter: $\rho_{\min}$, $\rho_{\max}$, ρ_v
Input: Audio-retention ratio $S_a = [S_a(1), ..., S_a(N)]$
Output: DP rates $\rho'_v = [\rho'_v(1), ..., \rho'_v(N)]$
$N \leftarrow \text{length}(S_a)$
$\rho'_{v\_initial} \leftarrow []$
{Step 1: Compute initial pruning ratios (Equation (5))}
for $i \leftarrow 1$ **to** N **do**
    $\rho'_v(i) \leftarrow \rho_{\max} - (\rho_{\max} - \rho_{\min}) \cdot S_a(i)$
    $\rho'_{v\_initial}$.append($\rho'_v(i)$)
{Step 2: Normalize to meet the global budget}
$T_{budget} \leftarrow \rho_v \times N$
$T_{initial} \leftarrow \sum(\rho'_{v\_initial})$
$\rho'_v \leftarrow \text{NormalizeRatios}(\rho'_{v\_initial}, T_{initial}, T_{budget})$
return ρ'_v
end function

B. Discussion

B.1. Adaptivity of OmniZip

The design of OmniZip is motivated by an analysis of audio-visual tokens and the dominant paradigm of their time-window-based arrangement in OmniLLMs. Notably, current mainstream models are generally based on this time-window paradigm [17, 47, 60, 61, 64, 67]. This approach divides the continuous audio-visual stream into discrete time segments, fuses or concatenates the tokens from each modality within their respective segments, and finally inputs the combined sequence into a large language model. This architectural commonality facilitates the adaptation of OmniZip to other existing models.

We also acknowledge that the field of OmniLLMs is still nascent, which raises the reasonable question of whether OmniZip would lose efficacy if some models no longer rely on explicit time-window concatenation. We argue that the core principle of OmniZip exploits the inherent temporal locality of audio-visual data streams. Within any short time segment, there is a high degree of correlation and synchronization between audio and video, accompanied by significant redundancy. Therefore, OmniZip remains a viable strategy, as its core mechanism—guiding token pruning by analyzing multi-modal tokens within a local temporal window—is fundamentally feasible and effective.

B.2. Hardness of Omnimodal Token Compression

While prior work in visual token compression has achieved high reduction rates (e.g., 70-85%), this is because a single modality is inherently simpler to compress. However, for OmniLLMs, the variable contribution of audio and video across different tasks, and the fact that audio information, as a high-dimensional feature, is less intuitively compressible than visual data, complicates this process. Additionally, recent models increasingly incorporate token efficiency as a core design principle, making further gains from simple pruning more difficult to achieve. Therefore, token pruning audio-video tokens is significantly more challenging. Nevertheless, achieving comprehensive video understanding necessitates the joint processing of both audio and visual information, making an effective token compression strategy all the more critical. In summary, as the first audio-visual token compression method, OmniZip sets a new benchmark for future technological advancements.

C. Computing Cost Evaluation

We examine the total FLOPs introduced by *audio tokens* and *video tokens* of the prefilling stage and the decoding stage. In OmniLLMs, a transformer layer comprising a multi-head attention (MHA) module and a feed-forward network (FFN) module is considered. Here, n denotes the token count, d the hidden state dimension, and m the FFN intermediate dimension. In the prefilling phase, the total FLOPs can be approximated as $4nd^2+2n^2d+2ndm$. In the decoding phase, taking into account the significant contribution introduced by the KV cache the computational consumption for $\mathcal{R}$ total iterations (*i.e.*, predicting $\mathcal{R}$ tokens) is $\mathcal{R}\left(4d^2 + 2dm\right) + 2\sum_{i=1}^{\mathcal{R}} d \times (n + i)$. We unify $\mathcal{R} = 100$ for calculation in the experiments. Thus, for an LLM with T total transformer layers, the total FLOPs can be expressed as follows,

$$\begin{aligned}\text{FLOPs} = T(4nd^2 + 2n^2d + 2ndm) \\ + T\mathcal{R}\left((4d^2 + 2dm) + 2\left(dn + \frac{d(\mathcal{R}+1)}{2}\right)\right).\end{aligned} \tag{9}$$

Method	Settings			Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.
	Retained Ratio	ρ_a	ρ_v									
Qwen2.5-Omni-7B												
Full Tokens	100%	-	-	52.4	50.1	48.5	44.6	43.8	41.6	41.6	47.3	46.8
Random	55%	0.45	0.45	47.1	47.0	44.4	41.2	40.0	40.1	40.1	46.3	43.6
FastV	50%	0.5	0.5	48.8	47.4	44.2	44.1	41.2	38.3	40.0	46.6	44.3
DyCoke (V&A)	50%	0.5	0.5	48.4	49.9	46.7	41.4	39.9	40.8	40.2	46.5	44.6
OmniZip (Ours)	50%	0.5	0.5	50.4	49.5	47.7	42.5	41.6	41.2	42.8	47.8	46.1
DyCoke (V&A)	45%	0.55	0.55	47.1	49.5	44.5	41.2	40.8	40.7	40.5	46.6	44.1
OmniZip (Ours)	45%	0.55	0.55	50.0	49.8	47.6	42.7	40.1	40.7	41.2	47.8	45.5
OmniZip (Ours)	45%	0.3	0.6	50.1	51.1	47.6	43.9	40.1	40.8	41.9	46.7	45.9

Table 6. **Comparison of different methods on the WorldSense benchmark**. 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. ρ_a and ρ_v are the pruning ratios of audio tokens and video tokens, respectively.

D. Related Work

D.1. Video Large Language Models

Video large language models (VideoLLMs) extend traditional LLMs and visual-language models [12, 13, 33], integrating video and language understanding into a unified framework [2, 6, 25, 27, 30, 33, 49, 54, 69, 70]. By jointly processing text and video inputs, VideoLLMs can perform complex cross-modal reasoning tasks, such as visual question answering and video captioning. They typically utilize pre-trained visual encoders and leverage powerful language backbones to align heterogeneous representations in a shared semantic space. Recent advancements, such as Qwen3-VL [51], InternVL3.5 [55], and Kimi-VL [50], have significantly advanced video-text understanding capabilities. However, as video inherently contains both visual and audio information, audio-video understanding is a key future research direction.

D.2. 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 enabled 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 token input derived from audio-video tokens is essential.

D.3. 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, 11, 22, 38, 46, 59, 63, 65, 68], video [5, 21, 35, 39, 41, 42, 48, 68], 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.

E. More Experimental Results

This section presents supplementary experimental results and ablation studies.

Tab. 6 presents comparison results under various pruning rates, primarily to further demonstrate that our method significantly outperforms other methods. Furthermore, OmniZip is designed to prune audio tokens more aggressively than video tokens (a heuristic derived from our analysis), but the data also demonstrates that our method's superior results are *not solely dependent on this specific ratio*. For example, at a 50% overall compression rate with a balanced 1:1 pruning ratio (ρ_a=0.5, ρ_v=0.5), OmniZip still achieves significantly

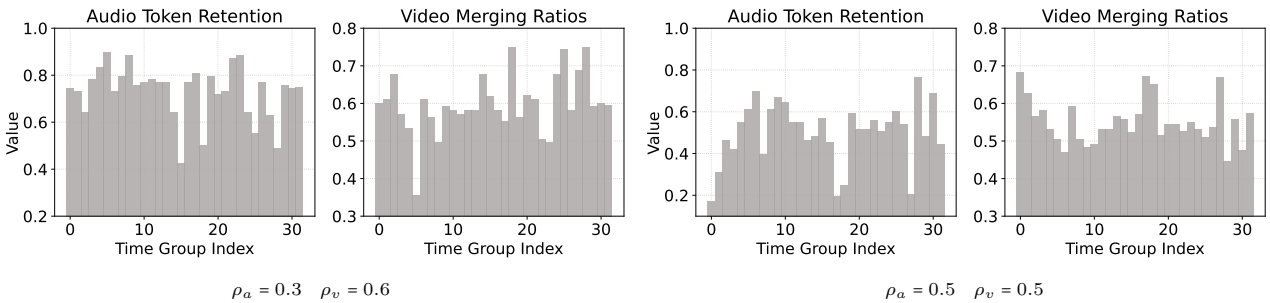

Figure 7. **More visualization of dynamic pruning ratios.** The figure illustrates how audio token retention guides the allocation of video token pruning. Specifically, for time windows with low audio retention, we allocate a higher video pruning ratio while maintaining a constant total pruning rate.

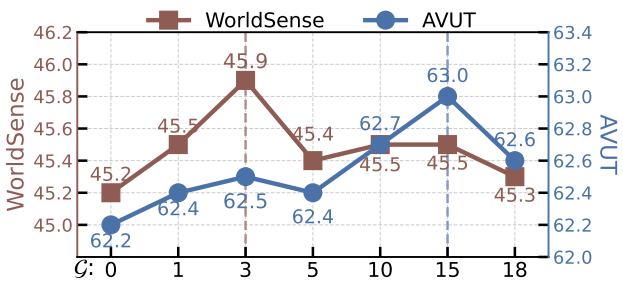

Figure 8. **Ablation study on $\mathcal{G}$.** The accuracy of our method in a 45% retained ratio is analyzed with the value of $\mathcal{G}$, which is defined as the number of tokens merged by each audio token anchor. All experiments illustrated in the figure were carried out on the Qwen2.5-Omni-7B model.

better performance than other methods.

In addition, for the dynamic pruning ratio allocation, we provide more visualization results as shown in Fig. 7.

Ablation Study on $\mathcal{G}$. As shown in Fig. 8, we evaluate the effect of $\mathcal{G}$. Primarily, the application of our audio token merging method yields substantial performance gains. On the AVUT [66], which is *audio-centric*, allocating a higher $\mathcal{G}$ proves to be appropriate. Conversely, in other benchmarks where audio is more balanced with video or serves as a supplementary modality, $\mathcal{G} = 3$ achieves the best results, while larger values introduce noise and slightly degrade performance. This finding indicates that $\mathcal{G}$ can be dynamically tuned based on the task's reliance on audio information.

F. Limitations and Future Work

While this work is the first to demonstrate the acceleration of OmniLLMs via audio-visual token compression, it is important to acknowledge its current limitations. Firstly, the relative informational requirements of audio and video vary significantly across different tasks and contexts. Consequently, determining the optimal compression balance between audio and video tokens remains a significant challenge. Secondly, this method is designed primarily for offline inference and does not natively support online or arbitrary-length streaming audio-visual input [5, 9, 37, 62]. Developing a streaming video inference framework that effectively incorporates audio will be a primary focus of our future work. Finally, the substantial parameter count of larger models continues to impede their practical deployment. Consequently, investigating how to combine token compression with other advanced efficiency techniques, such as model quantization [31, 36, 53, 57] and pruning [14, 45, 56, 73], represents a promising research direction.