OmniSelect: Dynamic Modality-Aware Token Compression for Efficient Omni-modal Large Language Models

Morunliu Yang[1,3]*, **Ruotao Xu**[1]*, **Le Li**[1], **Yue Wang**[1], **Jianxin Zhang**[1], **Yihang Lou**[2], **Siwei Feng**[1], **Peifeng Li**[1] **Juntao Li**[1,3] †

[1]Soochow University [2]Peking University [3]Meta-Stone (Pinghu) Co., Ltd.

{mrlyangnlp, 20255227018}@stu.suda.edu.cn {ljt}@suda.edu.cn

Code: https://github.com/Yangmrl-nlp/OmniSelect

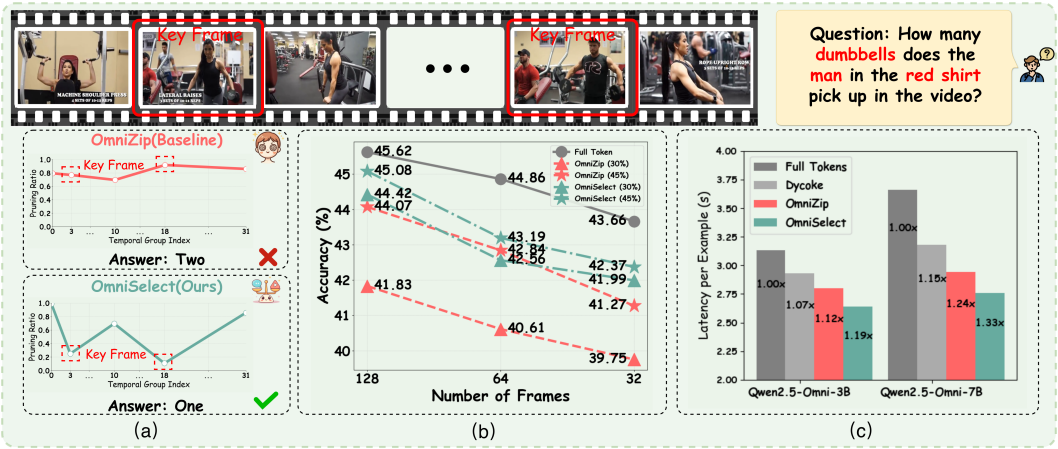


Figure 1: **(a)**: OmniSelect prunes fewer tokens in key frames and more in less important ones, while OmniZip prunes uniformly. **(b)**: OmniSelect retains 94% to 99% of the original full-token accuracy (Qwen2.5-Omni-3B, Worldsense, 128 Frames) and achieves competitive performance among existing training-free approaches.**(c)**: OmniSelect achieves an inference speedup of 1.19× to 1.33×.

Abstract

Omnimodal large language models (OmniLLMs) have recently gained increasing attention for unified audio-video understanding. However, processing long multimodal token sequences introduces substantial computational overhead, making efficient token compression crucial. Existing methods typically rely on fixed, modality-specific guidance, which fails to account for the varying importance of modalities across different queries. To address this limitation, we propose **OmniSelect**, a training-free, modality-adaptive token pruning framework that dynamically selects appropriate compression strategies for multimodal inputs. Specifically, we leverage a lightweight AudioCLIP model to estimate cross-modal relevance and categorize each input into three pruning regimes: Audio-Centric, Video-Centric, and Uniform pruning. Based on these relevance scores, OmniSelect further performs fine-grained token pruning within each temporal group, adaptively allocating pruning ratios to preserve informative tokens across modalities. By explicitly modeling modality preference and enabling dynamic strategy selection, OmniSelect

*Equal Contribution

†Corresponding author

‡Part of this work was completed when Morunliu Yang interned at Meta-Stone (Pinghu) Co., Ltd.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

effectively avoids the pitfalls of one-size-fits-all compression. Extensive experiments demonstrate that our method achieves efficient multimodal token reduction while maintaining strong performance, without requiring any additional training.

1 Introduction

Conventional Video Large Language Models (VLLMs) have achieved remarkable success in video question-answering and comprehension [2, 3, 16, 28], but they are often limited to processing visual cues, neglecting the rich auditory information present in videos. To address this, Omni-Modal Large Language Models (Omni-LLMs) [19, 20, 29, 30] were developed to integrate visual, auditory, and textual modalities within a unified autoregressive architecture. By capturing intricate cross-modal relationships and contextual nuances, this paradigm significantly enhances the model's capacity to perceive and interpret complex environments.

Since Omni-LLMs must process high-fidelity video and audio streams during inference, the resulting tokenization leads to an excessive number of tokens. This substantially increases the quadratic complexity of attention in Omni-LLMs, leading to significant computational and memory bottlenecks. To alleviate these issues, token compression has emerged as a promising technique. Substantial progress has already been made in single-modality compression involving images [32, 33], videos [18, 23], and audios [14, 15], which demonstrates its effectiveness in improving efficiency.

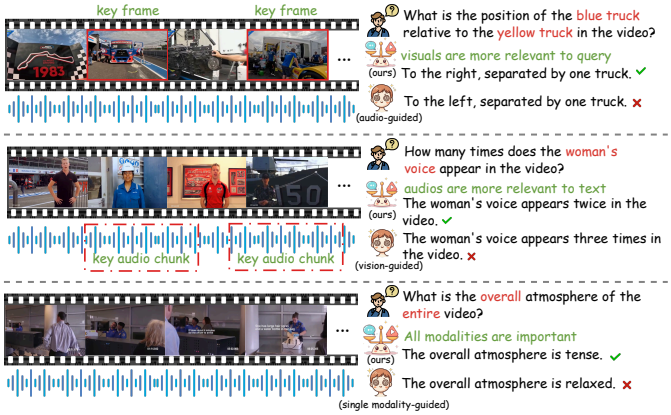


Figure 2: Illustration of modality importance variation across different questions.

Nevertheless, very few works have addressed token compression within the context of Omni-LLMs. OmniZip [25] introduces an innovative audio-visual token compression framework that utilizes audio anchors to dynamically guide video pruning through an interleaved spatiotemporal scheme, effectively accelerating OmniLLM inference in a training-free manner. OmniSIFT [5] introduces a modality-asymmetric token compression framework that employs a two-stage strategy of spatio-temporal video pruning and vision-guided audio selection, optimized end-to-end via a differentiable straight-through estimator.

While OmniZip [25] and OmniSIFT [5] utilize static guidance mechanisms based on audio or vision respectively, our analysis reveals that such one-size-fits-all strategies are inadequate. As shown in Figure 2, the relative importance of visual and auditory information varies significantly across different questions. While some questions rely heavily on vision, others are driven by audio or require a balanced integration of both modalities. Employing an inappropriate modality to guide token compression can lead to erroneous model responses. Therefore, we argue that the guidance strategy for token compression should be selected dynamically rather than relying on a fixed modality.

To facilitate dynamic selection of the guidance modality for token compression, it is essential to identify which modality is more relevant to the question. We employ AudioCLIP [8] to evaluate the cross-modal correlations between the question and various modalities. In the field of video keyframe sampling, existing works such as Q-Frame [34] and AKS [22] also utilize CLIP-based vision-language models to compute the relative importance between questions and visual information. Based on these principles, we present **OmniSelect**, a training-free framework that utilizes a two-stage strategy to dynamically compress visual and audio tokens. First, we utilize AudioCLIP [8] to calculate the similarity scores of visual-text and audio-text pairs, which defines the pruning ratio and categorizes the strategy into three distinct types: Uniform Pruning, Video-Centric Pruning, and

Audio-Centric Pruning. Second, redundant multimodal tokens are then pruned based on the attention score and cosine similarity matrix within each temporal group.

Experimental results demonstrate that OmniSelect delivers exceptional performance on audio-visual understanding tasks, and achieves competitive performance among existing training-free token compression methods. As shown in Figure 1, OmniSelect achieves an inference speedup of 1.19× to 1.33× while reducing GPU memory consumption by 2.61GB to 2.81GB. Despite these substantial resource savings, OmniSelect retains 94% to 99% of the original full-token accuracy.

Overall, our contributions are as follows:

- We analyze that existing methods typically rely on fixed, modality-specific guidance, which fails to account for the varying importance of modalities across different queries. Consequently, we propose the insight that modality-guided pruning should be selected dynamically.
- We propose **OmniSelect**, a training-free, modality-adaptive token compression framework that dynamically selects appropriate compression strategies for multimodal inputs and performs fine-grained token pruning within each temporal group, adaptively allocating pruning ratios to preserve informative tokens across modalities.
- Experimental results on audio-visual benchmarks show that OmniSelect accelerates inference speed and reduces GPU memory overhead while maintaining high performance and achieving competitive performance among training-free token compression methods.

2 Related Work

2.1 Omni-modal Large Language Models

Omni-Modal Large Language Models (Omni-LLMs) [10] represent an advanced evolution of traditional VideoLLMs [1, 2], incorporating audio alongside visual and textual data within a unified autoregressive architecture. This comprehensive approach enables the models to better understand complex environments by capturing intricate inter-modal dependencies and contextual nuances [19, 4, 31, 13, 20, 27]. Leading proprietary systems such as Qwen 3.5-Omni [26] and GPT-4o [17] have set high standards in audio-visual comprehension benchmarks [35, 9]. Simultaneously, the open-source community has created notable models such as Qwen2.5-Omni [29] and Qwen3-Omni [30]. These models use an end-to-end perception strategy that aligns specialized modality encoders with a central large language model (LLM) backbone through optimized projection layers.

2.2 Token Compression in Omni-LLMs

Processing high-fidelity video and audio streams generates an overwhelming number of multimodal tokens, hindering efficient deployment. The main challenge now is to maintain high-level reasoning performance while creating techniques to compress or simplify these integrated audio-visual inputs to reduce computational overhead [11]. While token compression methods for images [32, 33], videos [18, 23], and audios [14, 15] tasks have been widely studied, recent research has paid little attention to token compression in the omnimodal setting. As shown in Figure 3, Omnizip [25] identifies salient audio tokens and computes an audio retention score for each time group, and uses them to guide video token pruning. OmniSIFT [5] introduces a two-stage compression framework with spatio-temporal video pruning and vision-guided audio selection, optimized end-to-end via a differentiable straight-through estimator. Existing methods typically rely on fixed, modality-specific guidance, which fails to account for the varying importance of modalities across different queries. To address this limitation,

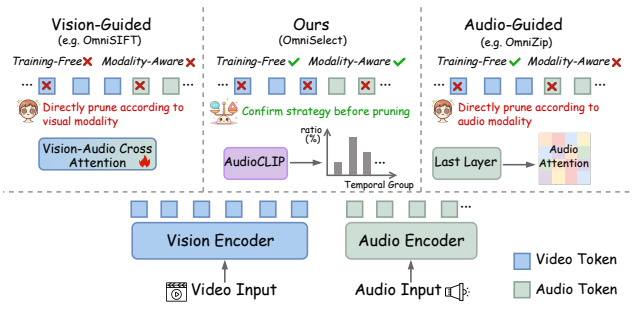


Figure 3: Illustration of different token compression strategies: vision-guided, audio-guided, and ours.

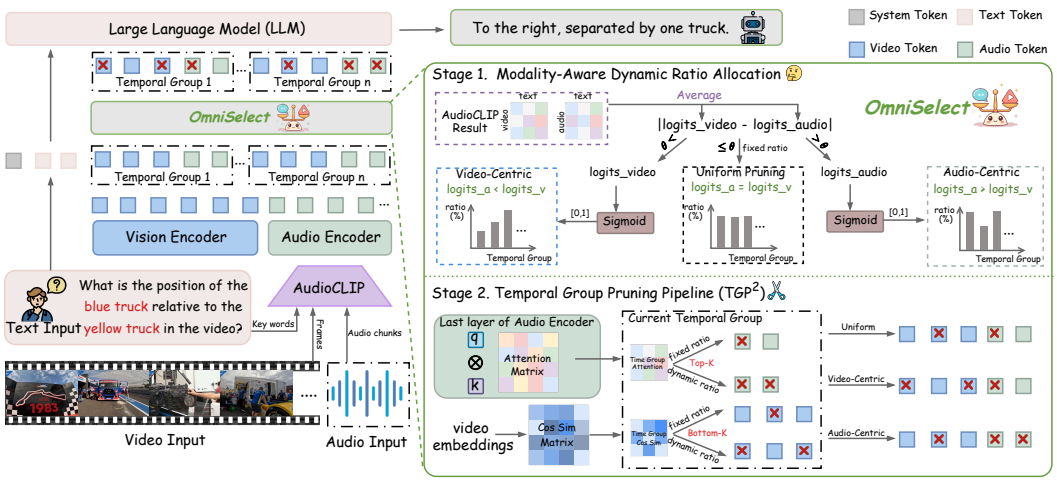


Figure 4: Overview pipeline of our method **OmniSelect**. The overall process is divided into two stages: (1) Modality-Aware Dynamic Ratio Allocation that allocates the pruning ratio for each temporal group while ensuring the total pruning ratio meets the expected value; (2) **T**emporal **G**roup **P**runing **P**ipeline (TGP^2) that prunes audio and video tokens based on attention score and cosine similarity score within each temporal group.

our approach is based on Modality-Aware techniques that dynamically determine the best pruning strategies.

3 Methodology

3.1 Our Method: OmniSelect

As illustrated in Figure 4, our method **OmniSelect** is fully training-free and utilizes a two-stage strategy to dynamically compress visual and audio tokens. First, we calculate the similarity scores of visual-text and audio-text pairs, map the resulting logits to the range $[0, 1]$, and use them to define the pruning ratio ρ_i for each temporal group i. This step also categorizes the pruning strategy into three distinct types: Uniform Pruning, Video-Centric Pruning, and Audio-Centric Pruning. Second, the method employs the **T**emporal **G**roup **P**runing **P**ipeline (**TGP**^{2}). After determining the pruning strategy, redundant multimodal tokens are pruned based on attention scores and cosine similarity matrix within each temporal group i. Notably, we set fixed pruning ratios ρ_v and ρ_a for visual and audio tokens, resulting in retained ratios of $1 - \rho_v$ and $1 - \rho_a$, respectively.

3.2 Modality-Aware Dynamic Ratio Allocation

Previous methods usually perform token pruning based on a single modality. However, in omni-modal reasoning, different modalities often contribute unequally depending on the query. Motivated by this observation, we aim to dynamically determine whether visual or audio information is more informative or whether both modalities should be treated equally. Preliminary analysis is provided in Appendix A. As discussed in this section, visual and audio tokens are grouped into temporal chunks before entering the LLM backbone. Following this design, we allocate pruning ratios at the chunk level, where each audio segment is temporally aligned with a sampled video frame.

To estimate modality importance, we compute cross-modal similarities between visual, audio, and textual representations using AudioCLIP [8], which maps all modalities into a shared embedding space. To reduce computational cost, video frames are temporally downsampled. For text, we retain only nouns and adjectives through lightweight linguistic filtering, since these words mainly capture

semantic objects and attributes. The similarity scores are computed as:

$$
\begin{aligned}
s_i^{(v)} &= \frac{\mathbf{v}_i \cdot \mathbf{t}}{\|\mathbf{v}_i\| \, \|\mathbf{t}\|}, \quad s_i^{(a)} = \frac{\mathbf{a}_i \cdot \mathbf{t}}{\|\mathbf{a}_i\| \, \|\mathbf{t}\|}, \\
\bar{s}^{(v)} &= \frac{1}{F} \sum_{i=1}^{F} s_i^{(v)}, \quad \bar{s}^{(a)} = \frac{1}{F} \sum_{i=1}^{F} s_i^{(a)}, \quad i = 1, 2, \ldots, F,
\end{aligned} \tag{1}
$$

where $\{\mathbf{v}_i\}_{i=1}^{F}$, $\{\mathbf{a}_i\}_{i=1}^{F}$, and $\mathbf{t}$ denote the visual embeddings of the F video frames, audio embeddings of the F audio chunks, and textual embeddings, respectively. The averaged scores $\bar{s}^{(v)}$ and $\bar{s}^{(a)}$ reflect the overall relevance of the visual and audio modalities to the query, respectively.

Based on the score difference and a threshold θ, we divide the pruning process into three cases. If $|\bar{s}^{(v)} - \bar{s}^{(a)}| \leq \theta$, both modalities are considered equally important and we apply **Uniform Pruning**. Otherwise, pruning becomes modality-aware. When $\bar{s}^{(v)} > \bar{s}^{(a)}$, we adopt **Video-Centric Pruning**; otherwise, we use **Audio-Centric Pruning**. For the less informative modality, we still employ Uniform Pruning to avoid excessive information loss.

To satisfy the global pruning budget while preserving score-based importance, we further design an adaptive allocation strategy. Let $\mathbf{s} = \{\hat{s_1}, \hat{s_2}, \ldots, \hat{s_G}\}$ denote the similarity scores and $\mathbf{n} = \{n_1, n_2, \ldots, n_G\}$ the token counts of the G temporal groups. The expected number of pruned tokens is defined as $N_{exp} = \rho_{v/a} \cdot \sum_{i=1}^{G} n_i$, where $\rho_{v/a}$ is the target pruning ratio for audio tokens or video tokens.

We first normalize the scores using their mean μ_s and standard deviation σ_s. When $\sigma_s > \epsilon$, the base pruning probability is obtained through a sigmoid mapping:

$$
\begin{aligned}
p_i &= \sigma\left(-\frac{\hat{s}_i - \mu_s}{\tau \cdot \sigma_s}\right), \\
\sigma(x) &= \frac{1}{1 + \exp(-x)},
\end{aligned} \tag{2}
$$

where τ controls the sharpness of the pruning distribution. If $\sigma_s \leq \epsilon$, all groups share the same pruning probability $p_i = \eta$. Next, we rescale the probabilities to match the global pruning budget and iteratively refine the allocation:

$$
\begin{aligned}
\rho_i^{(0)} &= \text{clip}\left(p_i \cdot \frac{N_{exp}}{\sum_{j=1}^{G} p_j n_j}, 0, 1\right), \\
\rho_i^{(k+1)} &= \text{clip}\left(\rho_i^{(k)} + \frac{N_{exp} - \sum_{j=1}^{G} \rho_j^{(k)} n_j}{\sum_{j \in \mathcal{A}} n_j}, 0, 1\right), \quad \forall i \in \mathcal{A},
\end{aligned} \tag{3}
$$

where $\mathcal{A} = \{i \mid 0 < \rho_i^{(k)} < 1\}$. The refinement continues until the pruning budget is satisfied or the maximum iteration number is reached. Finally, the pruning ratio for the i-th temporal group in modality $m \in \{v, a\}$ is defined as:

$$
\rho_{i,m} = \begin{cases}
\rho_{v/a}, & |\bar{s}^{(v)} - \bar{s}^{(a)}| \leq \theta, \\
\sigma\left(-\frac{\hat{s}_{i,v} - \mu_{s_v}}{\tau \cdot \sigma_{s_v}}\right), & |\bar{s}^{(v)} - \bar{s}^{(a)}| > \theta \quad \text{and} \quad \bar{s}^{(v)} > \bar{s}^{(a)}, \\
\sigma\left(-\frac{\hat{s}_{i,a} - \mu_{s_a}}{\tau \cdot \sigma_{s_a}}\right), & |\bar{s}^{(a)} - \bar{s}^{(v)}| > \theta \quad \text{and} \quad \bar{s}^{(v)} < \bar{s}^{(a)},
\end{cases} \tag{4}
$$

where τ is a temperature hyperparameter controlling the concentration of pruning ratios. The retained token number for each temporal group is then computed as:

$$
K_{i,m} = \max\left(1, \lfloor (1 - \rho_{i,m}\%) \cdot n_i \rfloor\right). \tag{5}
$$

Overall, this strategy dynamically allocates computational resources toward the modality that is more semantically aligned with the query while maintaining a global pruning constraint.

Table 1: **Main results on OmniVideoBench, VideoMME, and DailyOmni.** * denotes performance exceeding Full Tokens. The best and second-best results are **bolded** and **underlined** for each column, respectively.

Method	Retained Ratio (%)	OmniVideo Bench (↑)	VideoMME (↑)				DailyOmni (↑)						
			Short	Medium	Long	Avg.	Con.	Event	AV Event	Com.	Inf.	Rea.	Avg.
Qwen2.5-Omni-3B													
Frame Budgets 128 (VideoMME 512):													
Full Tokens	100%	32.9	74.8	64.1	52.8	63.89	55.44	56.21	53.78	70.99	79.22	74.29	62.82
Random	55%	31.0	68.8	61.6	52.4	61.00	47.67	44.44	45.38	64.12	71.43	63.43	53.55
DyCoke (V&A)	50%	31.3	71.1	62.8	52.4	62.11	48.19	49.35	47.48	**65.65**	72.73	65.14	55.89
OmniZip	45%	32.5	**74.0**	64.6	52.7	63.74	47.15	47.71	46.64	64.12	**76.62**	**69.71**	56.14
OmniZip	30%	31.4	70.6	61.2	52.7	61.48	46.63	46.08	42.86	59.54	68.18	68.00	53.05
OmniSelect (Ours)	45%	32.7	73.4	**65.0***	**53.3***	**63.93***	**50.78**	**49.67**	**52.10**	**65.65**	74.03	69.14	**58.06**
OmniSelect (Ours)	30%	**33.3***	69.7	61.9	53.2	61.59	46.63	45.42	48.32	62.60	72.08	65.14	54.39
Qwen2.5-Omni-7B													
Frame Budgets 128 (VideoMME 512):													
Full Tokens	100%	34.6	77.1	66.8	55.4	66.44	56.99	60.13	50.84	71.76	78.57	78.86	64.16
Random	55%	32.1	73.9	66.6	55.3	65.07	52.33	49.67	41.48	70.99	72.73	71.43	56.89
DyCoke (V&A)	50%	32.4	73.9	66.4	55.1	65.15	50.26	52.94	44.96	70.23	77.92	**76.57**	59.48
OmniZip	45%	**33.3**	**75.4**	67.2	55.4	66.03	**53.89**	51.96	46.22	70.23	**77.27**	**76.57**	59.98
OmniZip	30%	32.8	74.3	66.1	55.4	65.29	47.15	47.39	43.70	63.36	76.62	73.71	55.97
OmniSelect (Ours)	45%	33.1	74.9	**67.7***	**56.4***	**66.33**	53.37	**53.92**	**46.64**	**72.52***	76.62	**76.57**	**60.65**
OmniSelect (Ours)	30%	32.7	73.8	66.3	56.3	65.48	48.19	50.00	46.22	67.18	74.68	73.14	57.39

3.3 Temporal Group Pruning Pipeline (TGP^2)

After allocating the pruning ratios for each temporal group, we prune the visual and audio tokens within each temporal group to retain more salient tokens before model inference. To achieve this, we propose the **T**emporal **G**roup **P**runing **P**ipeline (**TGP**^{2}). For the audio token pruning strategy, we compute the attention matrix from the last layer of the audio encoder $\mathcal{E}_a$ as follows:

$$\mathbf{Q}_a = \mathbf{W}_{\mathbf{Q}_a}(\mathcal{T}_a), \quad \mathbf{K}_a = \mathbf{W}_{\mathbf{K}_a}(\mathcal{T}_a),$$
$$\mathbf{A}^{(a)} = \text{Softmax}\left(\frac{\mathbf{Q}_a\mathbf{K}_a^\top}{\sqrt{d_k}}\right) \in \mathbb{R}^{n_{i,a}\times n_{i,a}}, \tag{6}$$

where $n_{i,a}$ represents the total number of audio tokens in the temporal group i. After performing the same pooling operation as in the encoder to align with actual audio token indices, the averaged attention score is denoted as $\mathbf{A}^{(a)}_{avg}$. Finally, we select the salient audio tokens by:

$$\mathcal{I}_{i,a} = \text{TopK}\left(\mathbf{A}^{(a)}_{avg},\ k = \lfloor(1-\rho_{i,a}\%)\cdot(\lceil\frac{n_{i,a}}{2}\rceil)\rfloor\right). \tag{7}$$

For the vision token pruning strategy, we compute the cosine similarity matrix of the vision embeddings $\mathbf{V}_i$ in the i-th temporal group as $\mathbf{S} = \bar{\mathbf{V}}_i\bar{\mathbf{V}}_i^\top$, where $\bar{\mathbf{V}}_i$ denotes the ℓ_2-normalized embeddings. We then compute the average similarity score for each token by averaging across the last dimension. The indices of salient tokens $\mathcal{I}_{i,v}$ are selected via:

$$\mathcal{I}_{i,v} = \text{BottomK}\left(\frac{1}{N_v}\sum_{j=1}^{N_v}\mathbf{S}_{v,j},\quad k = \lfloor(1-\rho_{i,v}\%)\cdot n_{i,v}\rfloor\right). \tag{8}$$

We use video embeddings instead of attention weights because they are more computationally efficient and provide a more reliable measure of token redundancy without the attention sink bias. More detailed analyses about why to choose the Bottom-K strategy are provided in the Appendix D.2.

4 Experiments

4.1 Experiments Setup

Models and Baselines. We conduct experiments on Qwen2.5-Omni models with 3B and 7B scales. We compare our method OmniSelect with three baselines. (i) **OmniZip** [25] is a training-free framework that dynamically compresses multimodal tokens by leveraging salient audio cues to guide video token pruning. (ii) **Random Pruning** randomly removes audio and video tokens under the same pruning ratio. We include this baseline to verify the effectiveness of our structured pruning

Table 2: **WorldSense results.** * denotes performance exceeding Full Tokens. The best and second-best results are **bolded** and **underlined** for each column, respectively.

Method	Retained Ratio	Tech & Science	Games	Daily Life	Film & TV	Music	Sports	Culture & Politics	Performance	Avg. (↑)
Qwen2.5-Omni-3B										
Frame Budgets 32:										
Full Tokens	100%	49.39	37.77	44.07	42.48	45.32	39.77	47.57	38.20	43.66
OmniZip	45%	**47.35**	37.77	41.64	39.58	43.10	37.67	42.39	36.33	41.27
OmniZip	30%	46.12	**39.48***	39.51	36.15	44.58	35.35	38.83	34.83	39.75
OmniSelect (Ours)	45%	46.94	38.20	**42.86**	**40.90**	44.33	**39.30**	44.98	37.45	**42.37**
OmniSelect (Ours)	30%	46.94	36.48	41.19	40.63	**47.04***	36.74	**45.95**	**37.83**	41.99
Frame Budgets 64:										
Full Tokens	100%	51.02	39.48	43.62	42.22	45.32	41.63	50.16	43.45	44.86
OmniZip	45%	**48.78**	38.63	43.16	40.11	**45.32**	38.84	44.66	**39.33**	42.84
OmniZip	30%	46.33	39.06	40.12	37.47	44.33	36.05	42.07	37.08	40.61
OmniSelect (Ours)	45%	**48.78**	38.20	**43.62***	**41.69**	44.58	**39.53**	**46.93**	37.83	**43.19**
OmniSelect (Ours)	30%	**48.78**	**41.63***	41.64	39.84	44.33	38.37	46.28	37.83	42.56
Frame Budgets 128:										
Full Tokens	100%	52.65	39.06	43.47	44.06	46.55	42.79	52.43	41.20	45.62
Random	55%	46.53	36.05	40.27	38.79	45.81	39.53	46.60	36.70	41.68
DyCoke (V&A)	50%	49.18	40.34	42.10	39.84	43.60	41.86	45.95	38.95	43.06
OmniZip	45%	50.20	39.91	43.16	41.16	46.55	40.23	48.54	40.07	44.07
OmniZip	30%	47.76	**41.20***	40.88	38.52	45.57	37.44	44.01	37.45	41.83
OmniSelect (Ours)	45%	**50.61**	40.34	43.77	**43.27**	46.31	**42.56**	**49.51**	**41.95***	**45.08**
OmniSelect (Ours)	30%	49.80	**41.20***	**43.92***	43.01	**46.80***	39.53	47.57	41.20	44.42
Qwen2.5-Omni-7B										
Frame Budgets 128:										
Full Tokens	100%	49.39	42.49	46.96	44.59	48.03	42.56	53.72	45.69	46.82
Random	55%	47.76	36.48	42.86	40.90	48.28	38.14	49.19	38.95	43.25
DyCoke (V&A)	50%	46.73	40.77	43.62	39.58	47.29	41.86	48.54	41.57	43.95
OmniZip	45%	47.35	41.20	45.14	42.74	**49.01***	42.33	49.51	43.07	45.27
OmniZip	30%	45.92	39.91	43.92	**43.27**	48.03	40.23	46.93	40.07	43.85
OmniSelect (Ours)	45%	**47.96**	**43.78***	**46.20**	43.01	48.03	**43.72***	**50.81**	43.07	**46.00**
OmniSelect (Ours)	30%	46.33	39.91	44.38	42.22	45.32	42.79	46.93	**43.45**	44.17

strategy. (iii) **DyCoke** [24] is a training-free dynamic token pruning method for video LLMs. In our experiments, we apply its first-stage Token Temporal Merging (TTM) module to compress both video and audio tokens. We do not include OmniSIFT [25] in the primary comparison because it requires training with a straight-through estimator, whereas our comparison focuses on training-free token compression methods.

Benchmarks. We evaluate OmniSelect and the baselines on four benchmarks: WorldSense [9], DailyOmni [35], VideoMME (with audio) [6], and OmniVideoBench [12]. WorldSense contains 3,172 QA pairs and focuses on real-world omnimodal understanding, requiring joint reasoning over audio, visual, and textual signals across diverse scenarios. DailyOmni consists of 1,197 samples and emphasizes audio-visual reasoning in daily-life scenarios with strong temporal dependencies. OmniVideoBench includes 1,000 QA pairs and is designed to evaluate synergistic audio-visual reasoning with step-by-step annotations, highlighting modality complementarity and logical consistency. VideoMME comprises approximately 2,700 samples and provides a comprehensive evaluation of video understanding, covering perception, temporal reasoning, and multimodal integration.

Implementation Details. When applying OmniSelect, we set the threshold θ following Section 4.4. Specifically, we use $\theta = 0$ when $|\bar{s}^{(v)} - \bar{s}^{(a)}| \leq 2$, and $\theta = 5$ when $|\bar{s}^{(v)} - \bar{s}^{(a)}| > 2$. This setting is used across all benchmarks in Tables 1 and 2, providing stable performance across datasets. Note that a single threshold is not optimal for all cases due to varying modality dominance, and per-instance best results are reported in Appendix D.3. We evaluate robustness under different frame budgets on WorldSense and DailyOmni, while setting 512 frames for VideoMME and 128 frames for OmniVideoBench. FlashAttention is used in all experiments to improve inference efficiency and reduce memory overhead. Additional implementation details are provided in Appendix D.2.

4.2 Main Results

Competitive Performance among Training-Free Compression Methods. As shown in Table 1 and Table 2, when retaining 45% of tokens, OmniSelect surpasses all training-free compression baselines on most datasets. It preserves 97.7% of the full-token performance on OmniVideoBench across all compression settings and achieves 99.95% performance retention on VideoMME at 45% token retention. Compared with existing methods, OmniZip may underperform when visual information

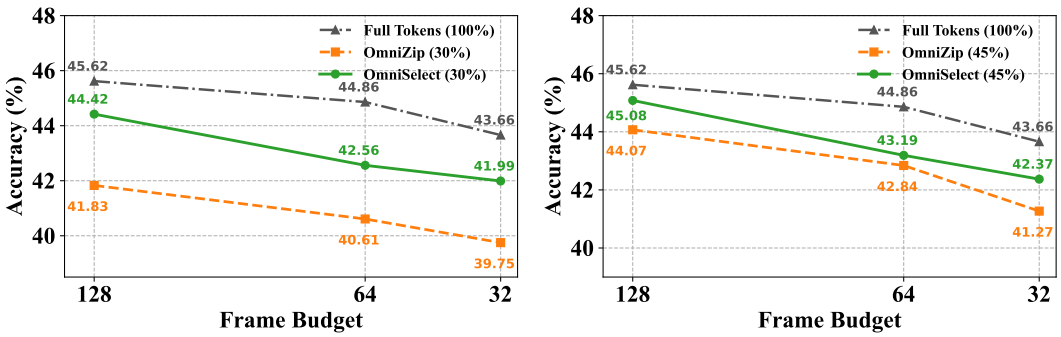
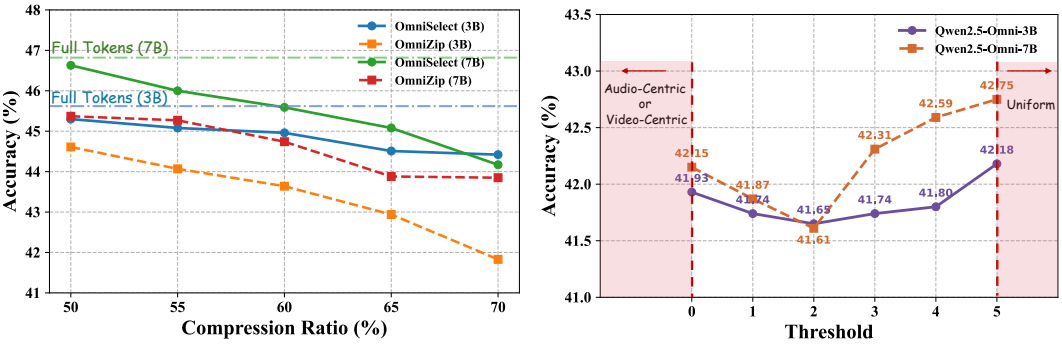


Figure 5: Qwen-2.5-Omni-3B performance under varying frame budgets at 30% and 45% pruning.



(a) Performance under different compression ratios.

(b) Performance under different thresholds.

Figure 6: Performance under different compression ratios and different thresholds.

is more critical due to its audio-guided pruning strategy, while DyCoke can mistakenly compress salient tokens because of only temporal merging. In contrast, OmniSelect occasionally even surpasses full-token inference, suggesting that removing redundant tokens can improve prediction quality. Moreover, the advantage becomes more pronounced on long videos, where accurate selection of key frames and salient audio segments is particularly important.

Robustness across Frame Budgets and Compression Ratios. Existing omni-modal token compression methods usually adopt fixed frame budgets. To evaluate the robustness of OmniSelect, we conduct experiments on the WorldSense benchmark under varying frame budgets. As shown in Figure 5, at both 30% and 45% compression ratios, OmniSelect consistently outperforms OmniZip across 32, 64, and 128 frame budgets. This indicates that OmniSelect preserves informative tokens more effectively under limited input. We further evaluate performance under different compression ratios on WorldSense. As shown in Figure 6a, OmniSelect consistently surpasses OmniZip across all tested ratios. Interestingly, we observe that token pruning yields larger gains on the 3B model compared to the 7B model, indicating a more favorable efficiency–performance trade-off in smaller backbones. This observation is consistent with the phenomenon reported in OmniZip [25].

4.3 Efficiency Analyses

To further validate the effectiveness of OmniSelect, we evaluate the average peak GPU memory usage, average prefilling time, and average inference latency on Qwen2.5-Omni-3B and 7B models. Specifically, the latency measurements in Table3 use the same starting point for all methods after common raw multimodal data loading and preprocessing, while including all method-specific computation. Measured separately on WorldSense, AudioCLIP feature extraction and modality scoring require an average of 0.191 seconds per sample. After including this overhead, OmniSelect still reduces total latency. For OmniSelect, the reported latency already includes AudioCLIP feature extraction, similarity computation, strategy selection, and token pruning. As shown in Table 3, OmniSelect achieves a GPU memory reduction of 2.58G–2.77G and provides a speedup of 1.19×–1.33× during

inference when compressing 70% of the audio and video tokens. It outperforms both OmniZip and DyCoke while maintaining the highest accuracy among all methods. Regarding prefilling efficiency, OmniSelect substantially reduces the computational overhead during prompt prefilling, achieving 3.53×-3.62× speedup under 70% token compression. This gain mainly comes from early removal of redundant multimodal tokens, which reduces unnecessary self-attention computation in later transformer layers.

Moreover, OmniSelect maintains strong robustness under aggressive compression, preserving 94.3%–97.4% of the original model accuracy on the WorldSense benchmark, while OmniZip and DyCoke exhibit larger performance degradation. These results demonstrate that OmniSelect achieves a favorable trade-off between inference efficiency and reasoning quality.

Table 3: **Efficiency Analyses on WorldSense benchmark**. The best result among the compression methods is **bolded**.

Method	Retained Ratio (%)	GPU Mem. (G) ↓	Prefilling Time (ms) ↓	Average Latency (s) ↓	Acc. (%) ↑
Qwen2.5-Omni-3B					
Frame Budgets 128:					
Full Tokens	100%	15.08	594 (1.00×)	3.13 (1.00×)	45.62
DyCoke (V&A)	50%	13.16	282 (2.11×)	2.93 (1.07×)	43.06
OmniZip	30%	12.49	**163** (3.64×)	2.80 (1.12×)	41.83
OmniSelect (Ours)	30%	**12.47**	164 (3.62×)	**2.64** (1.19×)	**44.42**
Qwen2.5-Omni-7B					
Frame Budgets 128:					
Full Tokens	100%	25.08	1119 (1.00×)	3.66 (1.00×)	46.82
DyCoke (V&A)	50%	23.00	536 (2.09×)	3.18 (1.15×)	43.95
OmniZip	30%	22.31	319 (3.51×)	2.94 (1.24×)	43.85
OmniSelect (Ours)	30%	**22.28**	**317** (3.53×)	**2.76** (1.33×)	**44.17**

4.4 Ablation Study

Ablation Study of Threshold θ. The threshold θ controls the transition between modality-centric pruning and Uniform pruning. To study its effect, we evaluate different fixed θ values on the WorldSense benchmark when compressing 70% tokens. As shown in Figure 6b, the extreme cases correspond to pure modality-centric pruning when $\theta \leq 0$ and pure Uniform pruning when $\theta \geq 5$. Both Qwen2.5-Omni-3B and 7B exhibit a clear U-shaped performance trend as θ varies. Specifically, modality-centric pruning at $\theta = 0$ achieves 41.93% and 42.15% accuracy for the 3B and 7B models, while Uniform pruning at $\theta = 5$ improves performance to 42.18% and 42.75%, respectively. In contrast, intermediate thresholds lead to noticeable degradation, with the 7B model dropping to 41.61% around $\theta = 2$, which suggests that a fixed intermediate threshold cannot reliably balance modality importance across diverse samples. This indicates that a single fixed threshold is suboptimal and motivates our piecewise design in Section 4.1, where θ is set to 0 for small modality gaps and 5 for large gaps to better balance modality importance across different samples.

The Significance of Pruning Strategy Classification. To validate the necessity of nuanced token selection, we evaluate the performance of OmniSelect against three baseline strategies: Audio-Centric, Video-Centric, and Uniform pruning. Experiments are conducted on the DaliyOmni benchmark across two models under two aggressive compression constraints (retaining 30% and 45% of audio and video tokens).

As illustrated in Figure 7, single-dimensional strategies exhibit clear limitations. For instance, at a 70% pruning rate, Audio-Centric pruning often discards vital visual context, while Uniform pruning fails to prioritize modality-specific information density, resulting in suboptimal accuracies of 54.14% and 55.47% respectively. Even when the retention budget increases to 45%, these fixed heuristics struggle to match the adaptive allocation of OmniSelect. Notably, OmniSelect achieves a peak performance of 60.65% on the 7B model, consistently outperforming the best-performing baseline. These results demonstrate the importance of our proposed taxonomy. Relying on any

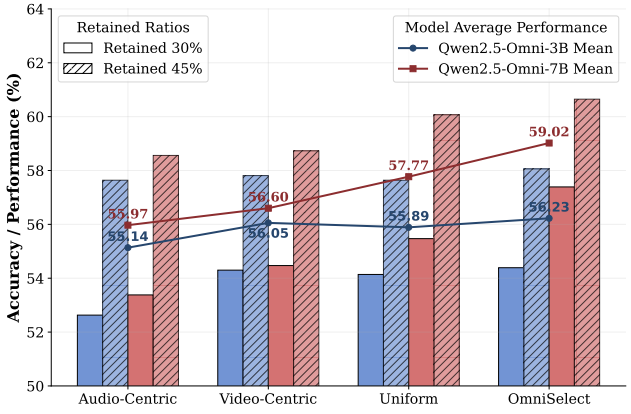


Figure 7: Comparison of different token compression strategies under 30% and 45% retained ratios on Qwen-2.5-Omni-3B and Qwen-2.5-Omni-7B. OmniSelect consistently achieves the best performance across all settings.

single strategy alone is insufficient to robustly handle diverse scenarios; instead, integrating all three categories is necessary to achieve consistently superior performance.

Asymmetric Pruning Analysis. We study the effect of modality-specific pruning operators on WorldSense at a 30% retention ratio. As shown in Table 4, our asymmetric design, which combines similarity-based video pruning (V_{sim}) with attention-based audio pruning (A_{attn}), achieves the highest overall accuracy of 44.42%. It outperforms both the Similarity Only variant ($V_{\text{sim}} + A_{\text{sim}}$, 43.25%) and the Attention Only variant ($V_{\text{attn}} + A_{\text{attn}}$, 40.70%). Furthermore, keeping audio pruning fixed and replacing V_{sim} with V_{attn} reduces the accuracy of the Games category by 6.87 points, from 41.20% to 34.33%. This result suggests that similarity-based video pruning is more effective for removing temporally redundant visual tokens. Conversely, keeping video pruning fixed and replacing A_{attn} with A_{sim} reduces the accuracy of Music category by 2.96 points, from 46.80% to 43.84%. This indicates that attention-based audio pruning better preserves salient acoustic information in audio-sensitive tasks.

Table 4: **Asymmetric Pruning Analysis on WorldSense.** Accuracy (%) at a 30% retention ratio. The **bold** number in each row indicates the best result.

Category	Ours $V_{\text{sim}} + A_{\text{attn}}$	Similarity Only $V_{\text{sim}} + A_{\text{sim}}$	Attention Only $V_{\text{attn}} + A_{\text{attn}}$
Overall	**44.42**	43.25	40.70
Music	**46.80**	43.84	44.33
Games	**41.20**	39.48	34.33
Tech./Science	**49.80**	48.37	47.14
Daily Life	**43.92**	41.95	40.58
Film/TV	**43.01**	41.69	39.58
Sports	**39.53**	39.07	35.81
Culture/Pol.	47.57	**48.22**	41.75
Performance	41.20	**42.70**	37.45

We further performed a latency spot check on 20 randomly selected samples. The attention-based audio pruning variant was faster on 13 samples, corresponding to 65% of the cases, with an average latency reduction of 11.2 ms and a maximum reduction of 134.4 ms. This result is consistent with the lower computation overhead of reusing attention-derived saliency, while similarity-based audio pruning requires additional pairwise feature similarity computation. Overall, the results support the use of modality-specific pruning operators and provide empirical evidence for our asymmetric design.

5 Conclusion

This paper presents OmniSelect, a training-free, modality-adaptive token compression framework for Omni-LLMs. Specifically, the framework dynamically selects appropriate compression strategies for multimodal inputs and performs fine-grained token pruning within each temporal group, adaptively allocating pruning ratios to preserve informative tokens across modalities. Extensive experimental results on audio-visual understanding tasks with two OmniLLMs (3B and 7B parameters) demonstrate that OmniSelect achieves competitive performance among existing training-free token compression approaches. OmniSelect achieves an inference speedup of 1.19× to 1.33× while reducing GPU memory consumption by 2.58GB to 2.77GB. Despite these substantial resource savings, OmniSelect retains 94% to 99% of the original full-token accuracy.

6 Acknowledgments

We want to thank all the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (NSFC No. 62622608 and 62576232).

References

[1] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025.

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

[3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024.

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

[5] Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, Yuanxing Zhang, Jiaheng Liu, Qiang Liu, Pengfei Wan, and Liang Wang. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models, 2026.

[6] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.

[7] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190. IEEE, 2023.

[8] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[9] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2026.

[10] Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. From specific-mllms to omni-mllms: A survey on mllms aligned with multi-modalities, 2025.

[11] Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models–from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025.

[12] 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.

[13] Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-omni-1.5 technical report, 2025.

[14] Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. Accelerating transducers through adjacent token merging, 2023.

[15] Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai "Helen" Li, and Yiran Chen. Speechprune: Context-aware token pruning for speech information retrieval, 2025.

[16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[17] OpenAI. Gpt-4o system card, 2024.

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

[19] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding, 2023.

[20] Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Caption-enhanced audio-visual large language models, 2025.

[21] Changli Tang, Qinfan Xiao, Ke Mei, Tianyi Wang, Fengyun Rao, and Chao Zhang. Wave: learning unified & versatile audio-visual embeddings with multimodal llm. In *International Conference on Learning Representations*, volume 2026, pages 61596–61612, 2026.

[22] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding, 2025.

[23] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models, 2025.

[24] Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18992–19001, 2025.

[25] Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Jian liu, and Huan Wang. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models, 2025.

[26] Qwen Team. Qwen3.5-omni: Scaling up, toward native omni-modal agi, March 2026.

[27] Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, Hanming Deng, Yuxuan Song, Dinghao Zhou, Guiping Zhong, Ken Zheng, Shiyin Kang, and Lewei Lu. Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue, 2025.

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

[29] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.

[30] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025.

[31] 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, 2025.

[32] 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, 2026.

[33] 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, 2024.

[34] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms, 2025.

[35] Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities, 2026.

A Preliminary

An OmniLLM [10] typically consists of a vision encoder $\mathcal{E}_v$, an audio encoder $\mathcal{E}_a$, and an LLM backbone. In the input pipeline, video, audio, and textual prompts are treated as raw data. Specifically, video sequences are converted into frames V via uniform temporal sampling, while audio signals A are processed using a fixed sampling rate. The inputs of the two modalities are mapped into token sequences as follows:

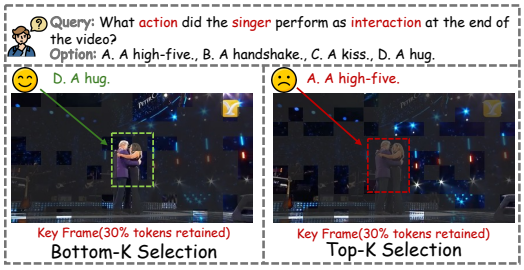
$$\mathcal{T}_v = \mathcal{P}_v(\mathcal{E}_v(V)), \quad \mathcal{T}_a = \mathcal{P}_a(\mathcal{E}_a(A)), \tag{9}$$

where $\mathcal{T}_v \in \mathbb{R}^{N_v \times D}$ and $\mathcal{T}_a \in \mathbb{R}^{N_a \times D}$ represent the token embedding sequences of visual and audio modalities. Here, N_v and N_a denote the number of visual and audio tokens, and D is the hidden dimension of the LLM backbone. The projectors $\mathcal{P}_v$ and $\mathcal{P}_a$ align the encoded features from their respective latent spaces into the joint embedding space of the LLM backbone.

Moreover, to achieve temporal alignment between visual and auditory modalities, OmniLLMs often employ a token-level concatenation strategy within synchronized time windows. Specifically, the original token sequences are partitioned into K temporal segments, where the visual and audio tokens within the same time interval are concatenated to form joint multimodal embeddings. This process can be formally defined as:

$$\mathcal{T}_m^{(i)} = [\mathcal{T}_v^{(i)} \parallel \mathcal{T}_a^{(i)}], \quad i = 1, 2, \ldots, K \tag{10}$$

where $\mathcal{T}_v^{(i)}$ and $\mathcal{T}_a^{(i)}$ denote the i-th segments of visual and audio tokens respectively, and $\parallel$ represents the concatenation operation. Consequently, the resulting sequence $\hat{\mathcal{T}} = \{\mathcal{T}_m^{(1)}, \mathcal{T}_m^{(2)}, \ldots, \mathcal{T}_m^{(K)}\}$ encapsulates synchronized bimodal information within each unified token. This enables the LLM backbone to model cross-modal dependencies in a temporally coherent latent space. These chunked multimodal tokens are subsequently concatenated with the textual embeddings $\mathcal{T}_t$ to form the unified input sequence.



(a) Bottom-K Selection vs. Top-K Selection.

Prompt Template

These are the frames of a video and the corresponding audio. Please answer the following multiple-choice question based on the video and audio content.Choose the correct option and respond with **only the letter** (A, B, C, ...) of your choice.

Question: {question} Options: {options_str} Answer:

(b) Prompt Template.

Figure 8: **Left:** The comparison of Bottom-K Selection strategy and Top-K Selection strategy. **Right:** Prompt Template for multiple-choice QA evaluations.

B Why Bottom-K Strategy Work for Video Token Pruning?

We adopt video embeddings instead of attention weights for two main reasons. First, computing full attention maps over a large number of vision tokens incurs prohibitive computational overhead, whereas embedding-based similarity is significantly more efficient and scalable. Second, feature-level similarity offers a more direct and stable measure of token redundancy, effectively avoiding the attention sink bias and noise commonly observed in pre-trained attention heads.

Furthermore, pruning is performed within each temporal group to meet the pre-allocated token budget defined in Section 3.2. Thanks to the dynamic proportion allocation mechanism, which already mitigates a substantial portion of redundancy across the temporal dimension by assigning higher token budgets to more informative segments, the remaining redundancy becomes predominantly localized within each temporal group. As illustrated in Figure 8a, this makes the Bottom-K strategy particularly effective: informative visual cues are typically sparsely distributed over time, while most tokens inside a local temporal window remain highly redundant.

Algorithm 1 OmniSelect: Dynamic Modality-Aware Token Compression

Input: Video tokens T_v, audio tokens T_a, query q, pruning ratio $\rho_{v/a}$, number of Frames F

Output: Pruned multimodal tokens $\hat{T}$

Divide (T_v, T_a) into G aligned temporal groups
Compute AudioCLIP embeddings: $v_i \leftarrow \text{AudioCLIP}(T_v^i)$, $a_i \leftarrow \text{AudioCLIP}(T_a^i)$, $t \leftarrow \text{AudioCLIP}(q)$
for $i = 1$ to F **do**
    $s_i^{(v)} \leftarrow \cos(v_i, t)$, $s_i^{(a)} \leftarrow \cos(a_i, t)$
end for
$\bar{s}^{(v)} \leftarrow \frac{1}{F}\sum_i s_i^{(v)}$, $\bar{s}^{(a)} \leftarrow \frac{1}{F}\sum_i s_i^{(a)}$
if $|\bar{s}^{(v)} - \bar{s}^{(a)}| \leq \theta$ **then**
    Strategy $\leftarrow$ Uniform
else if $\bar{s}^{(v)} > \bar{s}^{(a)}$ **then**
    Strategy $\leftarrow$ Video-Centric
else
    Strategy $\leftarrow$ Audio-Centric
end if
for $i = 1$ to G **do**
    $\rho_i \leftarrow \sigma\left(-\frac{\hat{s}_i - \mu_s}{\tau\sigma_s}\right)$
    $K_{i,a/v} \leftarrow \lfloor(1-\rho_i)\cdot n_i\rfloor$ or $\lfloor(1-\rho_{v/a})\cdot n_i\rfloor$
    $A \leftarrow \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)$ ▷ global attention computed once
    $A_i \leftarrow A[\mathcal{G}_i]$ ▷ slice attention for group i
    $I_a^i \leftarrow \text{TopK}(A_i, K_{i,a})$
    $S_i \leftarrow \bar{V}_i \bar{V}_i^\top$
    $I_v^i \leftarrow \text{BottomK}(S_i, K_{i,v})$
    $\hat{T}_i \leftarrow T_v^i[I_v^i] \cup T_a^i[I_a^i]$
end for
$\hat{T} \leftarrow \{\hat{T}_1, \hat{T}_2, \ldots, \hat{T}_G\}$
return $\hat{T}$

Table 5: **WorldSense Results under Any-Correct Evaluation with Strategy Diversity.** * denotes performance exceeding Full Tokens. The best results are **bolded** for each column.

Method	Retained Ratio	Tech & Science	Games	Daily Life	Film & TV	Music	Sports	Culture & Politics	Performance	Avg. (↑)
Qwen2.5-Omni-3B										
Frame Budgets 128:										
Full Tokens	100%	**52.65**	39.06	43.47	44.06	46.55	42.79	52.43	41.20	45.62
OmniSelect (Ours)	45%	51.63	**43.78***	45.14*	**44.33***	48.03*	**43.72***	**53.07***	41.57*	**46.60***
OmniSelect (Ours)	30%	51.43	42.92*	**45.59***	43.27	**48.52***	41.63	50.81	**42.70***	46.12*
Qwen2.5-Omni-7B										
Frame Budgets 128:										
Full Tokens	100%	**49.39**	42.49	**46.96**	44.59	48.03	42.56	**53.72**	**45.69**	46.82
OmniSelect (Ours)	45%	48.78	**43.35***	46.81	**45.38***	**49.26***	44.19*	51.78	**45.69**	**47.04***
OmniSelect (Ours)	30%	47.96	42.06	46.50	44.85*	46.80	**46.05***	49.84	44.57	46.34

By selecting tokens with the lowest similarity scores, Bottom-K preferentially removes redundant or highly correlated content while preserving visually distinctive tokens that are more likely to align with textual semantics. As shown in the figure, when answering "What action did the singer perform as interaction at the end of the video?", the Bottom-K selection accurately captures the critical hugging scene (correct answer: D. A hug), whereas Top-K focuses on a less relevant region and leads to the wrong prediction (A. A high-five). This approach achieves more balanced temporal coverage and effectively mitigates information collapse within dense regions. Overall, the Bottom-K strategy provides a simple yet powerful mechanism to retain representative and semantically diverse visual tokens, which is crucial for maintaining strong reasoning performance under aggressive compression.

Table 6: **DailyOmni Results under Any-Correct Evaluation with Strategy Diversity.** * denotes performance exceeding Full Tokens. The best results are **bolded** for each column.

Method	Retained Ratio	Con.	Event	AV Event	Com.	Inf.	Rea.	Avg. (↑)
Qwen2.5-Omni-3B								
Frame Budgets 128:								
Full Tokens	100%	55.44	**56.21**	53.78	**70.99**	79.22	**74.29**	**62.82**
OmniSelect (Ours)	45%	54.92	54.25	57.14*	67.94	**79.87***	72.00	62.32
OmniSelect (Ours)	30%	**55.96***	52.94*	**57.56***	67.18	75.32	68.57	61.07
Qwen2.5-Omni-7B								
Frame Budgets 128:								
Full Tokens	100%	**56.99**	**60.13**	**50.84**	71.76	78.57	**78.86**	**64.16**
OmniSelect (Ours)	45%	54.92	54.90	**50.84**	**74.05***	77.27	76.57	62.24
OmniSelect (Ours)	30%	52.33	53.59	49.16	69.47	**79.22***	74.29	60.57

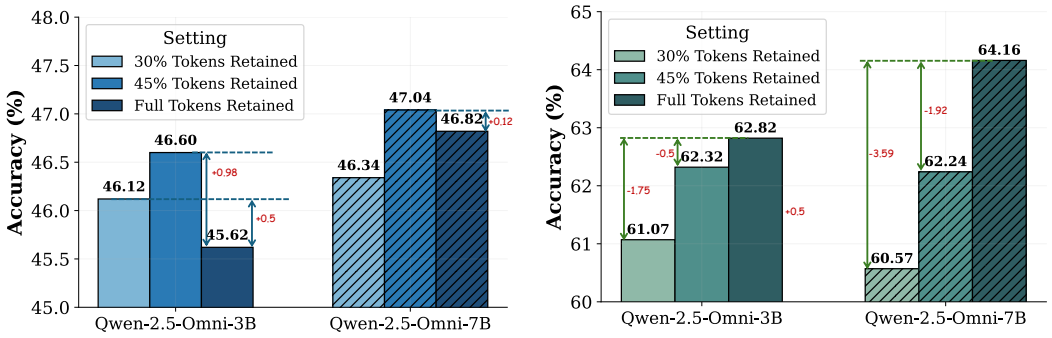


Figure 9: Qwen-2.5-Omni-3B and Qwen-2.5-Omni-7B performance on WorldSense (**Left**) and DailyOmni (**Right**) benchmarks at 30% and 45% pruning ratios.

C Dynamic Modality-Aware Token Compression Algorithm

To clearly present our overall pipeline, we summarize the full dynamic modality-aware token compression process in Algorithm 1. The algorithm first aligns video and audio tokens into G temporal groups to establish fine-grained cross-modal correspondence. It then computes AudioCLIP-based embeddings for each modality and the query, and estimates group-wise semantic similarities to determine the relative relevance of video and audio signals. Based on the aggregated similarity, a dynamic strategy is selected (video-centric, audio-centric, or uniform) to adaptively balance modality importance. Finally, under the chosen strategy, we perform group-wise token pruning with adaptive retention ratios and attention-guided selection, and merge the retained audio and video tokens to form the final compressed multimodal representation $\hat{T}$.

D More Experimental Details

D.1 More Experimental Results

We also evaluate the Qwen-2.5-Omni-7B model on the WorldSense benchmark under frame budgets of 32 and 64. As shown in Table 7, under the same compression settings, OmniSelect consistently outperforms OmniZip across different frame budgets. Furthermore, as illustrated in Figure 10, although the performance gains brought by token pruning on Qwen-2.5-Omni-7B are less pronounced than those observed on the 3B model, OmniSelect still consistently achieves better performance than OmniZip under the same compression ratio across all three frame budget settings. These results further demonstrate the robustness and effectiveness of our modality-aware dynamic pruning strategy on larger OmniLLMs.

Table 7: **WorldSense results when the frame budgets are set to 32 and 64.** * denotes performance exceeding Full Tokens. The best and second-best results are **bolded** and **underlined** for each column, respectively.

Method	Retained Ratio	Tech & Science	Games	Daily Life	Film & TV	Music	Sports	Culture & Politics	Performance	Avg. (↑)
Qwen2.5-Omni-7B										
Frame Budgets 32:										
Full Tokens	100%	47.55	41.63	44.68	42.74	46.31	41.40	49.51	42.32	44.70
OmniZip	45%	**45.92**	39.91	**43.77**	37.99	**47.04**	**41.40**	45.31	40.07	43.06
OmniZip	30%	45.10	**40.34**	40.43	38.52	46.80	38.60	42.07	38.58	41.49
OmniSelect (Ours)	45%	44.90	39.91	**43.62**	**39.05**	46.31	39.77	**47.90**	**41.95**	**43.10**
OmniSelect (Ours)	30%	45.10	36.91	41.34	38.52	45.32	41.16	44.66	38.95	41.87
Frame Budgets 64:										
Full Tokens	100%	48.78	41.20	45.39	44.85	47.04	41.16	51.46	44.19	45.71
OmniZip	45%	**47.14**	39.48	**44.53**	40.63	47.54*	40.93	46.93	43.07	44.10
OmniZip	30%	45.10	37.34	41.49	40.90	46.80	40.47	44.34	40.45	42.40
OmniSelect (Ours)	45%	45.51	**41.63***	43.77	**42.22**	**48.28***	**41.86***	**47.90**	**44.19**	**44.45**
OmniSelect (Ours)	30%	45.51	38.20	43.31	39.84	46.80	41.16	44.34	40.45	42.88

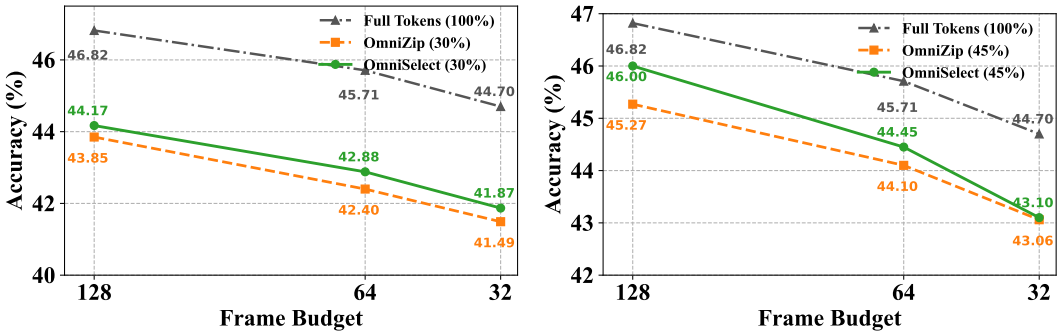


Figure 10: Qwen-2.5-Omni-7B performance under varying frame budgets at 30% and 45% pruning.

D.2 More Implementation Details

Prompt for QA Evaluation. For all QA-based benchmarks, we use a unified prompt template to ensure consistent evaluation across different datasets. Specifically, each input is formatted as shown in Figure 8. This simple and instruction-consistent template is used for all QA evaluations without task-specific modifications.

GPUs. All experiments are conducted on NVIDIA A100 GPUs with 40GB memory. For most benchmarks, a single GPU is sufficient to run inference. For VideoMME, due to its longer video duration and higher computational load, we use two GPUs to ensure efficient processing, while all other tasks are evaluated on a single GPU.

Input Configuration and Preprocessing. For inference, video inputs are uniformly sampled at a rate of 2 frames per second (FPS), with the total frame count restricted to a maximum of 32 / 64 / 128 / 512. Following the setting of baselines for fair comparison, the spatial resolution for each individual frame is configured at a maximum of $128 \times 28 \times 28$ pixels.

D.3 Any-Correct Evaluation under Strategy Diversity

To further analyze the robustness of our modality-aware strategy selection, we introduce a relaxed evaluation setting that considers all three candidate pruning strategies, including Uniform, Video-Centric, and Audio-Centric Pruning. In this setting, a sample is regarded as correctly solved if any one of the three strategies produces a correct prediction, regardless of the selected strategy.

We observe that the performance gap under the threshold-based strategy partition (θ) mainly stems from the limited discriminative capability of AudioCLIP in certain challenging cases, which may lead to suboptimal modality preference estimation. Since inference can only be performed once in practical settings, the strategy must be determined in a single forward pass, making the threshold-based decision inherently sensitive to such estimation errors. As shown in Table 5 and Table 6,

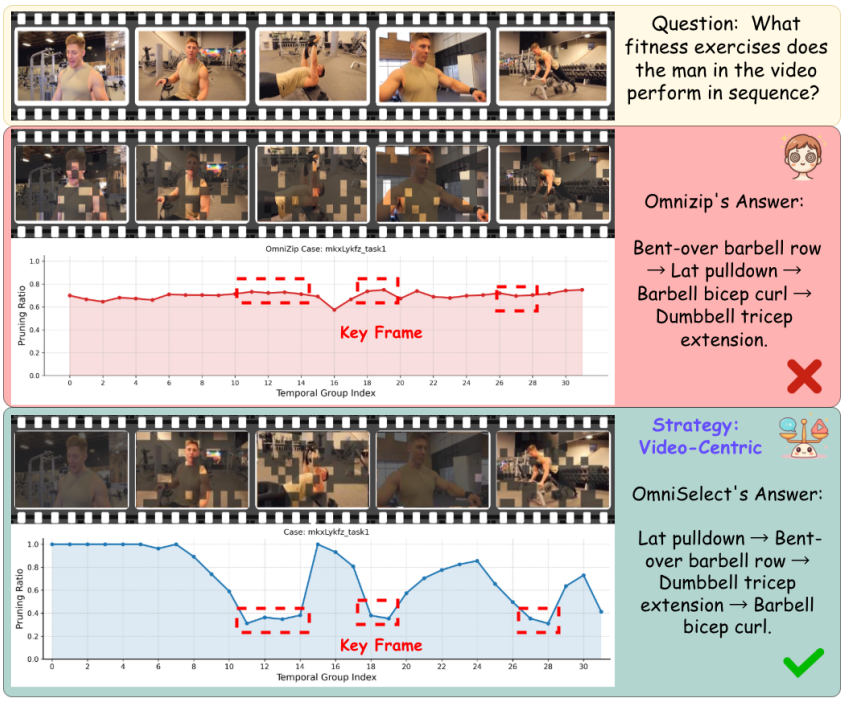


Figure 11: A Video-Centric pruning case, including pruning results, answers, and per-temporal-group pruning ratios for our method and the baseline.

this relaxed setting achieves performance very close to the Full Tokens setting, with only marginal degradation and, in some cases, even higher performance. For instance, as shown in Figure 9, when the retained ratio is set to 45%, OmniSelect improves the accuracy of Qwen-2.5-Omni-3B and Qwen-2.5-Omni-7B by 0.98% and 0.12%, respectively. Meanwhile, on the DailyOmni benchmark, it consistently preserves 94.4%–99.2% of the Full Tokens' performance, substantially outperforming threshold-based control methods. This indicates that the optimal strategy often exists within the proposed strategy space, and the main limitation lies in strategy selection rather than representation capacity.

Overall, these results further demonstrate that our dynamic modality-aware strategy is both reasonable and effective in balancing compression and performance, while also highlighting a potential direction for future improvement in more robust strategy prediction under weak modality discriminability.

E Relevance Estimator Analysis

We evaluate the effect of the relevance estimator by replacing AudioCLIP [8] with ImageBind [7] under the same experimental setting: WorldSense, Qwen2.5-Omni-7B, 128 input frames, and a 30% token pruning ratio.

Table 8: **Comparison of relevance estimators.** Results are measured on WorldSense using Qwen2.5-Omni-7B, 128 input frames, and a 30% token pruning ratio. Best results are **bolded** in each column.

Estimator	Parameters	Accuracy (%) ↑	Latency (s) ↓	GPU Mem. (GB) ↓
AudioCLIP	∼200M	**44.17**	**0.191**	**22.28**
ImageBind	∼1B	43.88	2.701	26.45

As shown in Table 8, AudioCLIP achieves 44.17% accuracy, slightly higher than ImageBind at 43.88%. AudioCLIP also has a substantially smaller model size (∼200M versus ∼1B parameters), lower measured latency (0.191 s versus 2.701 s), and lower GPU memory consumption (22.28 GB

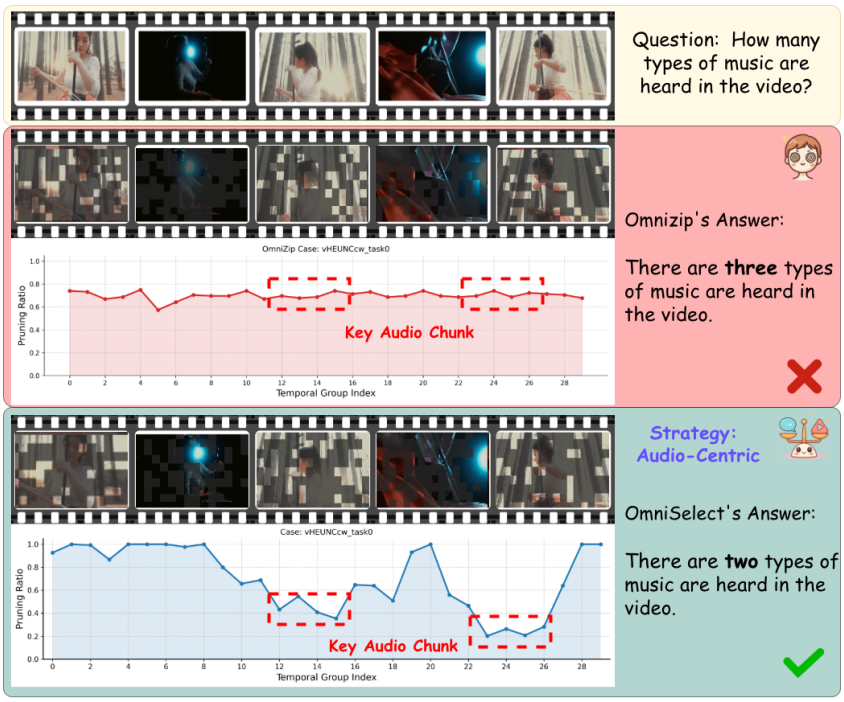


Figure 12: A Audio-Centric pruning case, including pruning results, answers, and per-temporal-group pruning ratios for our method and the baseline.

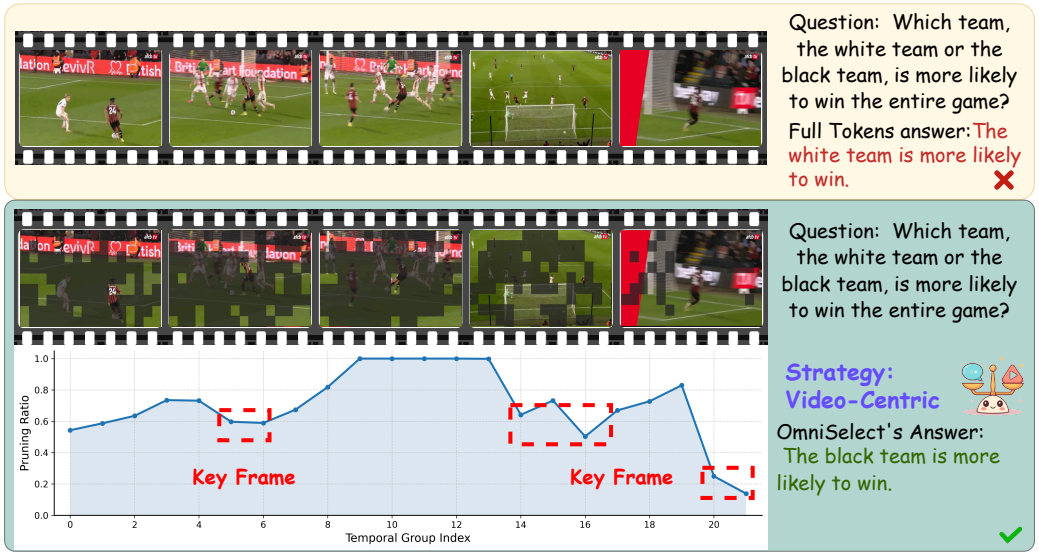


Figure 13: A Video-Centric pruning case that OmniSelect's answer corrects but Full Tokens does not, including pruning results, answers, and per-temporal-group pruning ratios for our method.

versus 26.45 GB). These results indicate that AudioCLIP offers a more favorable accuracy-efficiency trade-off for lightweight, training-free dynamic token selection. We do not include WAVE [21] in this comparison because its 7B parameters substantially exceed the scale of the lightweight relevance estimators considered here. Our goal is to use a training-free estimator with low computational overhead for omni-modal token selection; WAVE, as a large-scale multimodal foundation model, falls outside this comparison scope.

F Case Study

As shown in Figure 11, OmniSelect (our method) adopts a video-centric pruning strategy, aggressively removing tokens from frames that contain little relevant equipment or action information, while preserving more tokens for key fitness-related frames. In contrast, OmniZip (baseline) applies a nearly uniform pruning ratio across all frames regardless of their semantic importance. Consequently, OmniSelect successfully predicts the correct answer, whereas OmniZip fails.

In Figure 12, OmniSelect employs an audio-centric pruning strategy, uniformly pruning visual tokens while dynamically adjusting the pruning ratio for audio tokens, retaining more tokens in acoustically informative segments. By contrast, OmniZip uniformly prunes both visual and audio tokens, which leads to incorrect identification of the types and number of musical elements, potentially introducing duplication or omission errors. As a result, OmniSelect generates the correct answer, whereas OmniZip fails.

We further analyze cases where the Full Tokens setting produces the correct answer, while OmniSelect still maintains correct predictions after pruning 70% of the tokens. As illustrated in Figure 13, for the question *"Which team, the white team or the black team, is more likely to win the entire game?"*, OmniSelect preserves substantially more tokens around critical frames that are highly relevant to the game dynamics, while aggressively masking redundant tokens in less informative regions. This selective retention effectively reduces irrelevant contextual interference during reasoning, enabling the model to maintain accurate predictions even under heavy compression. This observation also helps explain why, under the Any-Correct Evaluation setting in Table 5, our method can even outperform the Full Tokens setting overall.

G Limitations and Future Work

Although OmniSelect demonstrates that dynamically categorizing pruning strategies based on modality importance is effective for omni-modal token compression, our method still has several limitations. First, the current strategy selection mechanism relies on manually designed threshold-based rules, which may not fully capture the complex and continuously changing relationships between audio and visual modalities across different queries. As observed in our experiments, a single fixed threshold is often insufficient to optimally balance modality importance under diverse scenarios, leading to suboptimal pruning decisions in challenging cases. Second, our framework depends on AudioCLIP to estimate cross-modal relevance. While lightweight and efficient, AudioCLIP cannot always accurately distinguish fine-grained modality importance, especially when audio and visual semantics are highly entangled or weakly aligned. Future work could explore stronger yet lightweight modality-aware models specifically optimized for multimodal importance estimation and dynamic strategy prediction. Our evaluation is currently conducted within the Qwen2.5-Omni model family, using its 3B and 7B variants, and our primary comparison focuses on training-free compression methods. We include OmniZip as a matched baseline, but do not compare with OmniSIFT because it requires training with a straight-through estimator and its implementation was unavailable when our experiments were conducted. Thus, generalization to other omni-modal backbones and learned compression methods remains underexplored. Finally, although OmniSelect achieves competitive performance under aggressive compression ratios, there remains substantial room for further exploration in omni-modal token compression. Future research should investigate how to remove even more redundant multimodal tokens while preserving most of the reasoning capability and performance of OmniLLMs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full set of assumptions and a complete (and correct) proof for each theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have reviewed the paper and confirmed that it discloses sufficient information to reproduce the main experimental results supporting the core claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have checked the paper, and it provides open access to the data and code with sufficient instructions to reproduce the main results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We have reviewed the paper and it specifies all the necessary training and test details (data splits, hyperparameters, optimizer, etc.) to understand the results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have reviewed the paper and it reports error bars suitably or provides appropriate information about statistical significance of the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have reviewed the paper and it provides sufficient information on the computer resources (compute workers, memory, execution time) needed to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the paper and confirm that the research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: OmniSelect improves the computational efficiency of existing OmniLLMs through token compression, without introducing new capabilities or application scenarios. It may reduce the computing resources needed to deploy these models, making them more accessible. While it inherits potential risks of the underlying models, such as biased or inaccurate outputs, we do not anticipate additional societal risks specific to the proposed compression method.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have reviewed the paper and confirm that the creators or original owners of all assets (code, data, models) are properly credited with licenses and terms of use explicitly mentioned and respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLMs were used to develop this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.