OmniSIFT: Modality-Asymmetric Token Compression for Efficient Omni-modal Large Language Models

Anonymous Authors[1]

Abstract

Omni-modal Large Language Models (Omni-LLMs) have demonstrated strong capabilities in audio-video understanding tasks. However, their reliance on long multimodal token sequences leads to substantial computational overhead. Despite this challenge, token compression methods designed for Omni-LLMs remain limited. To bridge this gap, we propose **OmniSIFT** (**Omni**-modal **S**patio-temporal **I**nformed **F**ine-grained **T**oken compression), a modality-asymmetric token compression framework tailored for Omni-LLMs. Specifically, OmniSIFT adopts a two-stage compression strategy: (i) a spatio-temporal video pruning module that removes video redundancy arising from both intra-frame structure and inter-frame overlap, and (ii) a vision-guided audio selection module that filters audio tokens. The entire framework is optimized end-to-end via a differentiable straight-through estimator. Extensive experiments on five representative benchmarks verify the efficacy and robustness of OmniSIFT. Notably, for Qwen2.5-Omni-7B, OmniSIFT adds 4.85M parameters while still achieving lower latency than training-free baselines such as OmniZip. With only 25% of the original token context, OmniSIFT consistently outperforms all compression baselines and even surpasses the full-token model on several tasks.

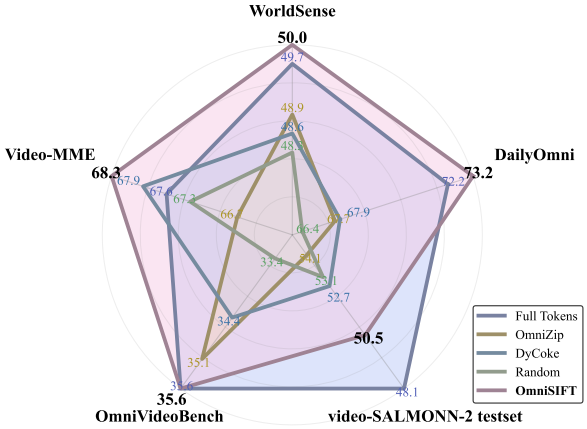

Figure 1. **Performance comparison across five audio–video benchmarks**. Results are obtained using Qwen2.5-Omni-7B with a 35% token retained ratio, comparing OmniSIFT against three baseline token compression methods and the full-token baseline.

1. Introduction

The rapid evolution of Omni-LLMs (Cheng et al., 2024; Xu et al., 2025b; Liu et al., 2025a) has significantly advanced holistic audio-video-language understanding (Hong et al., 2025; Zhou et al., 2025; Li et al., 2025). However, video signals are composed of densely sampled consecutive frames (Chen et al., 2024b; Jiang et al., 2025a), and audio streams must be encoded at high temporal resolution to capture acoustic dynamics (Ji et al., 2024). When these high-resolution streams are tokenized and interleaved for joint reasoning, the resulting sequence length grows rapidly. For example, a typical 20-second multimodal clip can yield more than 20K tokens (Xu et al., 2025a). Such long token sequences significantly increase computational cost, particularly for long video understanding (Fu et al., 2025).

Token compression (Chen et al., 2024a; Liu et al., 2025d;b; Ye et al., 2025) has emerged as a practical solution to mitigate the prohibitive computational cost caused by excessive token sequences. In the context of vision-centric MLLMs, a substantial body of work has explored effective strategies for pruning redundant visual tokens (Chen et al., 2024a; Tao et al., 2025a; Yao et al., 2025), demonstrating that significant efficiency gains can be achieved with minimal performance degradation. However, directly extending these approaches to audio–video understanding in Omni-LLMs is far from straightforward. As illustrated in Fig. 2, the modality-decoupled compression method directly transfers

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

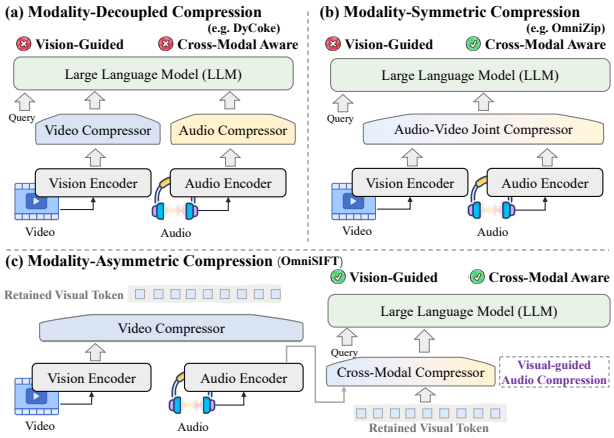

Figure 2. **Compression paradigm comparison for Omni-LLMs**. Token compression for Omni-LLMs can be categorized into three paradigms: (a) modality-decoupled compression (left top), which applies audio and video compression independently; (b) modality-symmetric compression (right top), which treats the two modalities equally informative; and (c) modality-asymmetric compression (bottom, ours), which first prunes visual redundancy and then performs visually guided audio compression.

vision-only techniques to both video and audio streams. While simple, this strategy completely ignores cross-modal semantic dependencies (Seo et al., 2023) and may discard tokens that are jointly informative.

A recent line of work adopts a modality-symmetric token compression paradigm. OmniZip (Tao et al., 2025b) follows this paradigm by first compressing audio tokens using attention scores from the audio encoder, and then guiding video token pruning with audio-derived saliency. Its reliance on attention-based saliency limits compatibility with efficient operators such as FlashAttention (Shah et al., 2024). In addition, treating the two modalities as equally informative collapses the compression process into selecting salient temporal positions, rather than capturing modality-specific semantic cues. EchoingPixels (Gong et al., 2025) also adopts a modality-symmetric design, performing global cross-modal contextualization over all audio and video tokens via four additional LLM decoder layers before compression. This compression method delays compression to a late stage and introduces substantial computational overhead.

In practice, humans process audio–video content asymmetrically (Koppen et al., 2008). Visual redundancy can typically be resolved using visual cues alone, whereas the saliency of audio signals depends on whether the visual scene provides a semantic anchor (Zhao et al., 2018; Arandjelovic & Zisserman, 2017), such as a visible speaker or a visually grounded event (Chowdhury et al., 2025). This perceptual asymmetry suggests that effective omni-modal token compression should be guided by visual semantics rather than treated symmetrically across modalities.

Taken together, these observations suggest ***three design principles*** for Omni-LLM token compression: (1) Modality-asymmetric, vision-guided compression; (2) Lightweight compression; (3) Compatibility with efficient operators.

Based on the above analysis, we present **OmniSIFT** (**Omni**-modal **S**patio-temporal **I**nformed **F**ine-grained **T**oken compression), a modality-asymmetric framework for visually guided token compression. As illustrated in Figure 2, OmniSIFT first prunes spatial and temporal redundancy in video to produce a compact set of visual anchors, and then uses these anchors to select the audio tokens that are most informative for the scene. This two-stage pipeline removes uninformative signals while preserving the key multimodal cues required for reasoning.

With only 4.85M additional parameters, OmniSIFT achieves lower latency than training-free baselines such as OmniZip on Qwen2.5-Omni-7B. Moreover, with only 25% of the original tokens retained, it consistently outperforms all compression baselines and even surpasses the full-token model on several settings, as illustrated in Figure 1.

Our main contributions are summarized as follows:

- Based on the asymmetric dependency between audio and video, we derive practical design principles for omni-modal token compression.
- We present OmniSIFT, a modality-asymmetric framework that first removes spatial and temporal redundancy in video tokens and then uses the resulting visual anchors to select informative audio tokens.
- Extensive experiments across five benchmarks show that OmniSIFT delivers strong performance–efficiency gains, achieving higher accuracy even with only 25% of the original tokens.

2. Related Works

2.1. Omni-modal Large Language Models

Omni-LLMs (Jiang et al., 2025b) extend large language models to process heterogeneous modalities within a unified autoregressive framework. Unlike Video-LLMs (An et al., 2025; Bai et al., 2025), which focus on visual–text inputs, Omni-LLMs additionally incorporate audio signals (Cheng et al., 2024; Tang et al., 2025; Liu et al., 2025a). Proprietary systems such as GPT-4o (Hurst et al., 2024) and Gemini (Comanici et al., 2025) further demonstrate strong performance on audio–visual understanding tasks (Li et al., 2025; Hong et al., 2025). In the open-source community, models like Qwen2.5-Omni (Xu et al., 2025a) adopt a typical architecture that aligns modality-specific encoders with an LLM through learned projection layers.

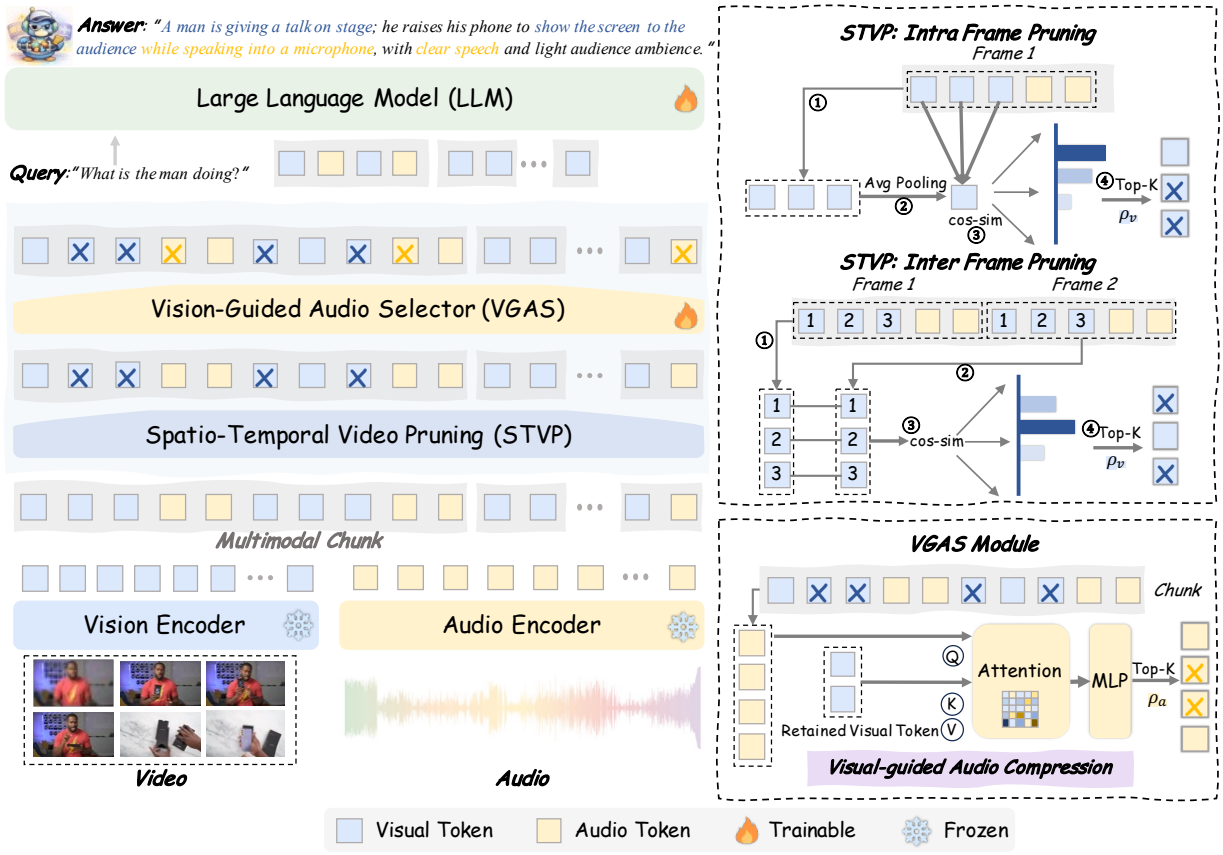

Figure 3. **Architecture of OmniSIFT, a modality-asymmetric compression framework**. The framework operates in two stages. In the first stage, STVP removes spatial and temporal redundancy in video tokens to obtain a compact set of visual anchors. In the second stage, VGAS selects audio tokens conditioned on these visual anchors. The resulting compressed multimodal sequence is then fed into the LLM backbone for downstream reasoning.

2.2. Token Compression in Multimodal Models

In the video domain, token compression methods such as VisionZip (Yang et al., 2025), VidCom^2 (Liu et al., 2025c), TimeChat-Online (Yao et al., 2025), and DyCoke (Tao et al., 2025a) estimate token importance through various saliency or similarity metrics. Recent work has begun to explore compression in the audio–video setting. OmniZip (Tao et al., 2025b) represents an early attempt, selecting salient audio tokens based on encoder attention and using them to guide video compression. EchoingPixels (Gong et al., 2025) takes a more tightly coupled approach, performing global audio–video contextualization before token compression.

3. Method

3.1. Preliminary

A typical Omni-LLM architecture (Xu et al., 2025a) includes modality-specific encoders Φ_v and Φ_a, cross-modal projectors, and a generative LLM backbone. Given a video clip $\mathcal{V}$ and synchronized audio $\mathcal{A}$, the encoders map each modality into token sequences compatible with the LLM backbone. Specifically,

$$\mathbf{Z}_v = \Phi_v(\mathcal{V}), \quad \mathbf{Z}_a = \Phi_a(\mathcal{A}) \tag{1}$$

where $\mathbf{Z}_v \in \mathbb{R}^{N_v \times D}$ and $\mathbf{Z}_a \in \mathbb{R}^{N_a \times D}$ are the resulting visual and audio token sequences, with N_v and N_a denoting the numbers of visual and audio tokens extracted by the encoders, and D denoting the LLM hidden dimension.

To maintain temporal alignment, Omni-LLMs group tokens from both modalities into aligned chunks. Let $\mathcal{C}_t$ denote the t-th chunk. We define the multimodal block as $\mathcal{C}_t = [\mathbf{Z}_v^{(t)}; \mathbf{Z}_a^{(t)}]$, where $\mathbf{Z}_v^{(t)} \in \mathbb{R}^{n_v \times D}$ and $\mathbf{Z}_a^{(t)} \in \mathbb{R}^{n_a \times D}$ are the visual and audio tokens in the same interval, with n_v and n_a denoting the visual and audio tokens per chunk, respectively. The final input sequence $\mathcal{S} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ is interleaved with textual instructions as the LLM's input.

Each visual sub-sequence $\mathbf{Z}_v^{(t)}$ corresponds to two consecutive frames. Let $n_p \triangleq n_v/2$ denote the number of visual tokens per frame. Let $\mathbf{F}_1^{(t)}, \mathbf{F}_2^{(t)} \in \mathbb{R}^{n_p \times D}$ be the token sequences of the two frames.

3.2. OmniSIFT

As illustrated in Figure 3, OmniSIFT operates in two stages: (1) a Spatio-Temporal Video Pruning (STVP) module that removes spatial and temporal redundancy from visual tokens within each chunk, and (2) a Vision-Guided Audio Selector (VGAS) module that selects audio tokens with the refined visual context. Each multimodal chunk $\mathcal{C}_t$ serves as the basic processing unit for OmniSIFT.

We denote by $\rho_v, \rho_a \in (0, 1]$ the visual and audio compression ratios, representing the proportions of tokens removed in the video and audio stages. The corresponding retention ratios used for token selection are $\alpha_v = 1-\rho_v, \alpha_a = 1-\rho_a$.

3.3. Spatio-Temporal Video Pruning

Visual tokens in Omni-modal LLMs exhibit substantial redundancy, arising from spatial redundancy within each frame and temporal overlap across consecutive frames. The problem we aim to solve is: *how can we retain spatially distinctive regions and temporally changing areas, while discarding redundant patches under a fixed visual compression ratio ρ_v?*

We introduce a Spatio-Temporal Video Pruning (STVP) module that operates at the chunk level. We adopt a two-stage pruning strategy: (1) compute spatial saliency scores on $\mathbf{F}_1^{(t)}$ and temporal saliency scores on $\mathbf{F}_2^{(t)}$, and (2) select the top-ranked tokens according to the retention ratio α_v.

Spatial Saliency Estimation. The first frame captures the static visual layout of the scene. To identify spatially distinctive patches, we compute a frame-level representation via mean pooling.

$$\bar{\mathbf{v}}_1^{(t)} = \frac{1}{n_p} \sum_{i=1}^{n_p} \mathbf{v}_{1,i}^{(t)}, \tag{2}$$

which summarizes the global visual context. Spatial saliency for token $\mathbf{v}_{1,i}^{(t)}$ is then defined as its cosine deviation from this representation:

$$s_{1,i}^{(t)} = 1 - \frac{\mathbf{v}_{1,i}^{(t)} \cdot \bar{\mathbf{v}}_1^{(t)}}{\|\mathbf{v}_{1,i}^{(t)}\| \, \|\bar{\mathbf{v}}_1^{(t)}\|}. \tag{3}$$

Tokens with higher scores correspond to patches that deviate more from the global frame context and more informative.

Temporal Saliency Estimation. The second frame reflects temporal evolution, such as object motion or newly content. Using positional encodings, each token $v_{2,i}^t \in F_2^t$ can be matched to its corresponding patch token $v_{1,i}^t$ in the first frame, enabling the computation of temporal saliency:

$$s_{2,i}^{(t)} = 1 - \frac{\mathbf{v}_{2,i}^{(t)} \cdot \mathbf{v}_{1,i}^{(t)}}{\|\mathbf{v}_{2,i}^{(t)}\| \, \|\mathbf{v}_{1,i}^{(t)}\|}. \tag{4}$$

A higher temporal saliency score indicates a stronger deviation over time, capturing motion dynamics or appearance changes that contribute new information.

Saliency-guided Token Selection. Given the spatial and temporal saliency scores, STVP retains the most informative patches under the visual retention ratio α_v. Let $\hat{n}_p = \alpha_v n_p$ denote the number of tokens to keep per frame. We select the top-scoring tokens from each frame:

$$\begin{aligned} \hat{\mathbf{F}}_1^{(t)} &= \mathrm{TopK}(\mathbf{F}_1^{(t)}, \mathbf{s}_1^{(t)}, \hat{n}_p), \\ \hat{\mathbf{F}}_2^{(t)} &= \mathrm{TopK}(\mathbf{F}_2^{(t)}, \mathbf{s}_2^{(t)}, \hat{n}_p). \end{aligned} \tag{5}$$

The pruned visual sequence is $\hat{\mathbf{Z}}_v^{(t)} = [\hat{\mathbf{F}}_1^{(t)}; \hat{\mathbf{F}}_2^{(t)}]$.

3.4. Vision-Guided Audio Selector

Audio streams are sampled at high temporal resolutions and therefore contain substantial redundancy. The problem to solve is: *under a fixed audio compression ratio ρ_a, how can we identify audio tokens carrying scene-relevant information while safely discarding those that are uninformative?*

We rely on the intrinsic modality-asymmetric nature of audio–video data: whether a sound is vital can only be judged when paired with the corresponding visual cues. Motivated by this, we design the Vision-Guided Audio Selector (VGAS), which leverages compressed video tokens to guide audio token selection.

Formally, for the t-th chunk $\mathcal{C}_t$, VGAS takes as input

$$\mathbf{Z}_a^{(t)} \in \mathbb{R}^{n_a \times D}, \quad \hat{\mathbf{Z}}_v^{(t)} \in \mathbb{R}^{\hat{n}_v \times D}, \tag{6}$$

where $\mathbf{Z}_a^{(t)}$ denotes the whole audio token sequence and $\hat{\mathbf{Z}}_v^{(t)}$ is the pruned video token sequence produced by STVP. VGAS then applies a cross-modal attention mechanism to compute context-aware audio representations.

Vision-Guided Semantic Interaction. VGAS employs a lightweight cross-attention mechanism, where audio tokens act as queries $\mathbf{Q}_a$ and pruned video tokens act as keys $\mathbf{K}_v$ and values $\mathbf{V}_v$. Concretely, we compute

$$\mathbf{H}_a^{(t)} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_a \mathbf{K}_v^\top}{\sqrt{d}}\right) \mathbf{V}_v, \tag{7}$$

where d is the attention head dimension. This yields context-aware audio representations $\mathbf{H}_a^{(t)} \in \mathbb{R}^{n_a \times D}$, in which each audio token is enriched with visual information that emphasizes acoustics aligned with the observed scene.

Saliency Scoring and Token Selection. We project the context-aware audio representations $\mathbf{H}_a^{(t)}$ through a two-layer MLP followed by a sigmoid activation to obtain a scalar saliency score for each token:

$$s_{a,j}^{(t)} = \sigma(\mathrm{MLP}(\mathbf{h}_{a,j}^{(t)})), \tag{8}$$

Table 1. **Performance comparison Results**. Results are evaluated on Qwen2.5-Omni-7B and Qwen2.5-Omni-3B across multiple benchmarks, using retained ratios of 35% and 25%. The **best** result among token compression methods for each metric is bolded.

Method	Retained Ratio (%)	WorldSense (↑)	OmniVideo Bench (↑)	VideoMME (↑)				video-SALMONN-2 testset (↓)		
				Short	Medium	Long	Avg.	Miss	Hal	Total
Qwen2.5-Omni-7B										
Full Tokens	100	49.7	35.6	78.9	66.9	57.1	67.6	29.1	19.0	48.1
OmniZip	35	48.9	35.1	77.1	67.0	56.0	66.7	34.1	20.0	54.1
Random	35	48.3	33.4	77.2	**68.1**	56.6	67.3	33.2	19.9	53.1
DyCoke	35	48.6	34.4	78.7	68.0	56.9	67.9	32.6	20.1	52.7
OmniSIFT	35	**50.0**	**35.6**	**79.0**	67.9	**58.0**	**68.3**	**30.7**	**19.8**	**50.5**
OmniZip	25	48.1	34.1	76.4	66.1	55.3	66.0	35.8	21.4	57.2
Random	25	47.1	32.6	77.0	66.1	55.1	66.1	36.2	20.7	56.9
DyCoke	25	48.1	34.1	76.4	66.2	55.0	65.9	35.3	**20.0**	56.3
OmniSIFT	25	**49.9**	**35.4**	**78.6**	**67.8**	**58.3**	**68.2**	**30.9**	20.3	**51.2**
Qwen2.5-Omni-3B										
Full Tokens	100	45.8	33.5	76.1	63.4	52.9	64.2	32.8	20.8	53.6
OmniZip	35	44.1	**33.7**	74.7	**63.8**	53.1	63.5	36.9	22.2	59.1
Random	35	45.5	33.4	74.3	61.6	52.1	62.7	37.0	21.7	58.7
DyCoke	35	45.3	32.8	73.7	62.7	**53.7**	63.3	36.9	**21.6**	58.5
OmniSIFT	35	**45.7**	**33.7**	**76.1**	62.2	52.8	**63.7**	**35.2**	21.8	**56.9**
OmniZip	25	43.8	32.4	72.7	61.9	**52.3**	62.3	39.5	22.6	62.1
Random	25	43.3	33.0	74.0	61.9	50.9	62.3	39.3	22.6	62.0
DyCoke	25	44.1	33.0	73.3	**62.3**	51.9	62.5	40.2	**21.7**	61.9
OmniSIFT	25	**45.8**	**33.1**	**75.0**	62.0	52.1	**63.0**	**36.4**	21.9	**58.3**

forming a score sequence $\mathbf{s}_a^{(t)} = \{s_{a,j}^{(t)}\}_{j=1}^{n_a}$. Given the audio retention ratio α_a, we keep the top $\hat{n}_a = \alpha_a n_a$ tokens with the highest scores using a TopK operator, yielding the pruned audio sequence $\hat{\mathbf{Z}}_a^{(t)}$.

End-to-End Optimization. To enable gradient-based training with discrete TopK selection, we optimize VGAS with a straight-through estimator that uses a binary mask in the forward pass and passes gradients through the scores in the backward pass. During the forward pass, a binary mask $m_j \in \{0, 1\}$ is generated for each audio token based on its score $s_{a,j}^{(t)}$ via Top-k, and the selected tokens are passed to the LLM backbone. In the backward pass, we apply an identity surrogate for the TopK mask and approximate $\partial m_j / \partial s_{a,j}^{(t)} \approx 1$ to support end-to-end optimization.

4. Experiment

4.1. Experimental Setting

Model and Data. Following OmniZip (Tao et al., 2025b), we evaluate OmniSIFT on the Qwen2.5-Omni series (Xu et al., 2025a). For VGAS alignment, we fine-tune on the AVoCaDO SFT dataset (Chen et al., 2025), which provides 107K synchronized audio–visual captioning pairs.

Benchmarks. We evaluate OmniSIFT on four audio–visual QA benchmarks: VideoMME (with audio) (Fu et al., 2025), DailyOmni (Zhou et al., 2025), WorldSense (Hong et al., 2025), and OmniVideoBench (Li et al., 2025), as well as the video-SALMONN-2 captioning testset (Tang et al., 2025).

Baselines. We choose three baselines: (i) **OmniZip** (Tao et al., 2025b), the first method designed for Omni-LLMs; (ii) **DyCoke** (Tao et al., 2025a), a video-centric approach whose TTM module we adapt to prune video and audio tokens independently; (iii) **Random Pruning**, which drops video and audio tokens uniformly at random.

Implementation Details. The VGAS module uses a lightweight multi-head cross-attention layer with 8 heads and a 512-dimensional hidden size. We fine-tune only the LLM decoder and the VGAS module using a learning rate of 1×10^{-5} and a total batch size of 128. For fair comparison, we first fine-tune the Qwen2.5-Omni backbone under the same setting and then apply compression baselines to this model. Additional details are provided in Appendix B.

4.2. Main Results

4.3. Efficiency Analysis

State-of-the-Art Compression Performance. As shown in Table 1 and Table 2, we evaluate OmniSIFT on five audio-visual benchmarks using Qwen2.5-Omni-7B and 3B under

Table 2. **Performance comparison results on DailyOmni.** Results are evaluated on Qwen2.5-Omni-7B and Qwen2.5-Omni-3B across multiple benchmarks, using retained ratios of 35% and 25%. The **best** result among token compression methods is bolded.

Method	Retained Ratio (%)	Event Sequence	AV Event Alignment	Inference	Reasoning	Context Understanding	Comparative	Average
Qwen2.5-Omni-7B								
Full Tokens	100	66.7	70.6	79.2	76.6	69.9	77.1	72.2
OmniZip	35	63.7	63.0	77.3	76.6	59.1	74.8	67.7
Random	35	58.5	61.8	77.9	73.7	63.2	74.0	66.3
DyCoke	35	61.4	63.9	77.9	75.4	63.7	74.8	67.9
OmniSiFT	35	**66.7**	**70.2**	**83.1**	**78.9**	**69.9**	**79.4**	**73.2**
OmniZip	25	61.8	59.7	75.3	75.4	60.6	74.0	66.2
Random	25	61.1	56.7	78.6	71.4	60.1	73.3	65.2
DyCoke	25	57.2	56.7	80.0	74.3	61.1	71.0	64.7
OmniSiFT	25	**66.7**	**68.9**	**82.5**	**77.7**	**71.0**	**76.3**	**72.5**
Qwen2.5-Omni-3B								
Full Tokens	100	60.1	62.2	78.6	74.9	62.2	74.8	67.0
OmniZip	35	**60.5**	56.7	76.6	72.0	59.6	**72.5**	64.7
Random	35	55.9	54.2	76.0	74.3	60.1	68.7	62.9
DyCoke	35	53.9	52.5	**79.2**	**76.0**	60.1	**72.5**	63.2
OmniSiFT	35	57.8	**58.8**	77.3	73.7	**64.8**	69.5	**65.3**
OmniZip	25	57.8	55.0	75.3	70.3	58.5	70.0	64.2
Random	25	53.3	54.2	**76.6**	72.0	55.4	67.9	61.2
DyCoke	25	52.6	54.6	74.7	**74.3**	58.0	69.5	61.7
OmniSiFT	25	**58.5**	**59.7**	75.3	73.7	**60.6**	**70.2**	**64.7**

Table 3. **Efficiency comparison results**. Results are evaluated on Qwen2.5-Omni-7B and Qwen2.5-Omni-3B using the WorldSense benchmark, reporting peak GPU memory usage, inference latency, and accuracy. The **best** result among token compression methods for each metric is in bold, the second best result is underlined.

Method	Retained Ratio (%)	GPU Mem (GB)↓	Total Time (s)↓	Prefill Lat. (s)↓	E2E Lat. (s)↓	Acc (%)↑
Qwen2.5-Omni-7B						
Full Tokens	100	27.59	15097.1	4.76	4.94	49.7
OmniZip	35	22.92	8886.4	2.80	2.89	48.9
DyCoke	35	23.09	**8718.3**	**2.75**	**2.85**	47.3
OmniSIFT	35	**22.91**	8756.0	2.76	2.86	**50.0**
Qwen2.5-Omni-3B						
Full Tokens	100	18.91	11399.4	3.59	3.79	45.8
OmniZip	35	**14.75**	7750.4	2.44	2.59	44.1
DyCoke	35	14.92	**7578.8**	**2.39**	**2.53**	43.9
OmniSIFT	35	14.79	7596.3	**2.39**	**2.53**	**45.7**

35% and 25% token retention ratios. Across all settings, OmniSIFT consistently achieves the highest accuracy among compression methods. Notably, OmniSIFT matches or even surpasses the full-token model on multiple benchmarks; for example, with only 35% tokens retained on Qwen2.5-Omni-7B, it achieves 50.0 on WorldSense compared to 49.7 from the full-token model. We attribute this phenomenon to OmniSIFT's ability to remove redundant audio–visual tokens that may introduce noise, while preserving the key audio-visual cues required for reasoning.

Fine-Grained Category Results. Table 2 reports fine-grained DailyOmni results for both Qwen2.5-Omni-7B and Qwen2.5-Omni-3B under the two retained-ratio settings. In the most demanding categories, accuracy drops substantially for existing token compression methods. For example, with Qwen2.5-Omni-7B at the 25% retained ratio, OmniZip obtains only 61.8 on *Event Sequence* and 59.7 on *AV Event Alignment*, reflecting difficulties in modeling temporal dynamics and cross-modal consistency under aggressive compression. In contrast, OmniSIFT achieves 66.7 and 68.9 on these two categories, respectively, maintaining strong performance even at low token budgets.

Robustness Across Compression Ratios. Figure 4 compares OmniSIFT with other token compression methods under varied visual and audio compression ratios. For instance, when the audio compression ratio increases from 0.3 to 0.9, accuracy for OmniZip plummets from above 48.9 to around 44.0. In contrast, OmniSIFT maintains performance above 49.3 across the entire range, showing only minor fluctuations even at high compression levels.

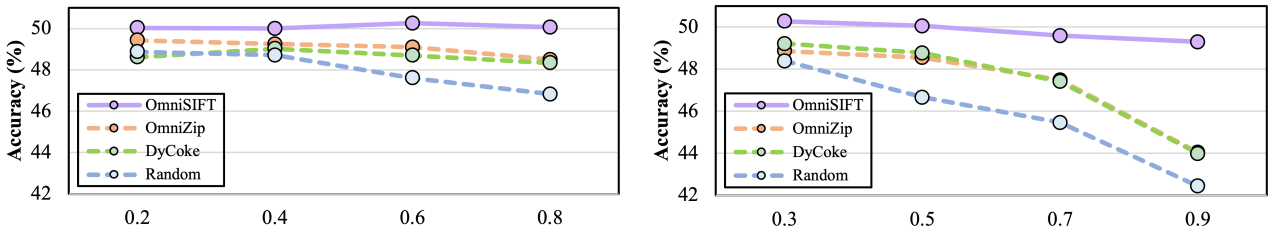

Figure 4. **Ablation results for video and audio compression ratios**, evaluated on the Qwen2.5-Omni-7B model using the WorldSense benchmark. **Left**: Varying the video compression ratio ρ_v with audio compression ratio $\rho_a = 0.5$; **Right**: Varying the audio compression ratio ρ_a with video compression ratio $\rho_v = 0.8$.

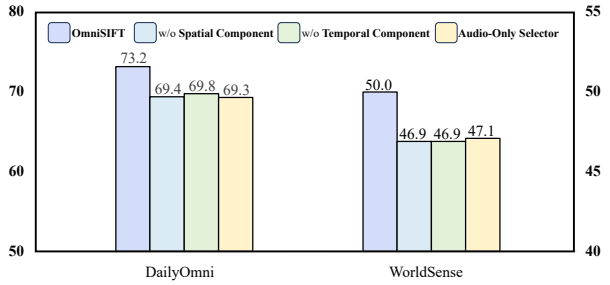

Figure 5. **Ablation results for OmniSIFT's architecture**. **w/o Spatial Component**: all visual tokens are selected using temporal saliency only, removing the spatial compression component. **w/o Temporal Component**: all visual tokens are selected based on spatial saliency only, removing the temporal compression component. **Audio-Only Selector**: audio tokens are selected solely based on intra-audio self-attention without any visual guidance.

Overall, these results demonstrate that OmniSIFT achieves the best balance between compression and performance, maintaining reliable audio-visual understanding even when retaining only a small fraction of the original tokens.

Table 3 summarizes the efficiency comparison on the WorldSense benchmark using both Qwen2.5-Omni-7B and Qwen2.5-Omni-3B, reporting inference latency and peak GPU memory for OmniSIFT and other token compression methods. Across both model scales, OmniSIFT delivers consistently better efficiency than the full-token model. On the 7B model, it reduces total inference time by more than 40% while lowering average peak memory usage by more than 4.6 GB, with similar improvements observed on the 3B variant. Although OmniSIFT introduces a lightweight cross-modal module, its end-to-end latency and peak GPU memory remain comparable to training-free approaches such as OmniZip and DyCoke.

4.4. Ablation Study

We investigate two dimensions of OmniSIFT through ablation studies: the contribution of its video and audio compression modules, and the effect of the asymmetric token compression paradigm. All ablation experiments are conducted on the Qwen2.5-Omni-7B model for consistency.

Structural Ablation: STVP and VGAS. Figure 5 presents the structural ablation results of the STVP and VGAS modules in OmniSIFT. For the STVP module, we evaluate the contributions of its spatial and temporal components. Removing either part leads to clear accuracy drops on both DailyOmni and WorldSense, indicating that spatial and temporal redundancy capture complementary aspects of video information and both are required for effective visual token reduction. For the VGAS module, we compare an Audio-Only Selector whose audio selection relies solely on intra-audio dependencies. Specifically, we replace the cross-attention module in VGAS with an audio self-attention module within each chunk, resulting in accuracy drops of 3.9% and 2.9% on DailyOmni and WorldSense, respectively. This indicates that audio relevance is highly context-dependent and cannot be reliably assessed without visual cues. Additional ablation results for the VGAS module are detailed in Appendix D.4.

Token Compression Paradigm Ablation. Table 4 compares our modality-asymmetric paradigm, which performs vision-guided audio compression, with an OmniZip-style modality-symmetric paradigm that performs audio-guided video compression. Both paradigms are evaluated under three retained-ratio settings on DailyOmni and WorldSense. To instantiate the modality-symmetric baseline, we fine-tune the same Qwen2.5-Omni-7B backbone with the OmniZip pruning rule. Across all retained ratios, OmniSIFT performs better consistently than this symmetric variant, and the performance gap widens as the retained ratio decreases. These results demonstrate that compared to modality-symmetric pruning strategies, OmniSIFT consistently preserves more salient audio-visual tokens by explicitly modeling cross-modal dependencies between visual and audio tokens. Consequently, OmniSIFT exhibits outstanding stability across diverse audio-visual compression schemes.

4.5. Case Study

Figure 6 presents a case study comparing OmniSIFT with OmniZip on OmniVideoBench, illustrating a key limitation of modality-symmetric compression methods: assume au-

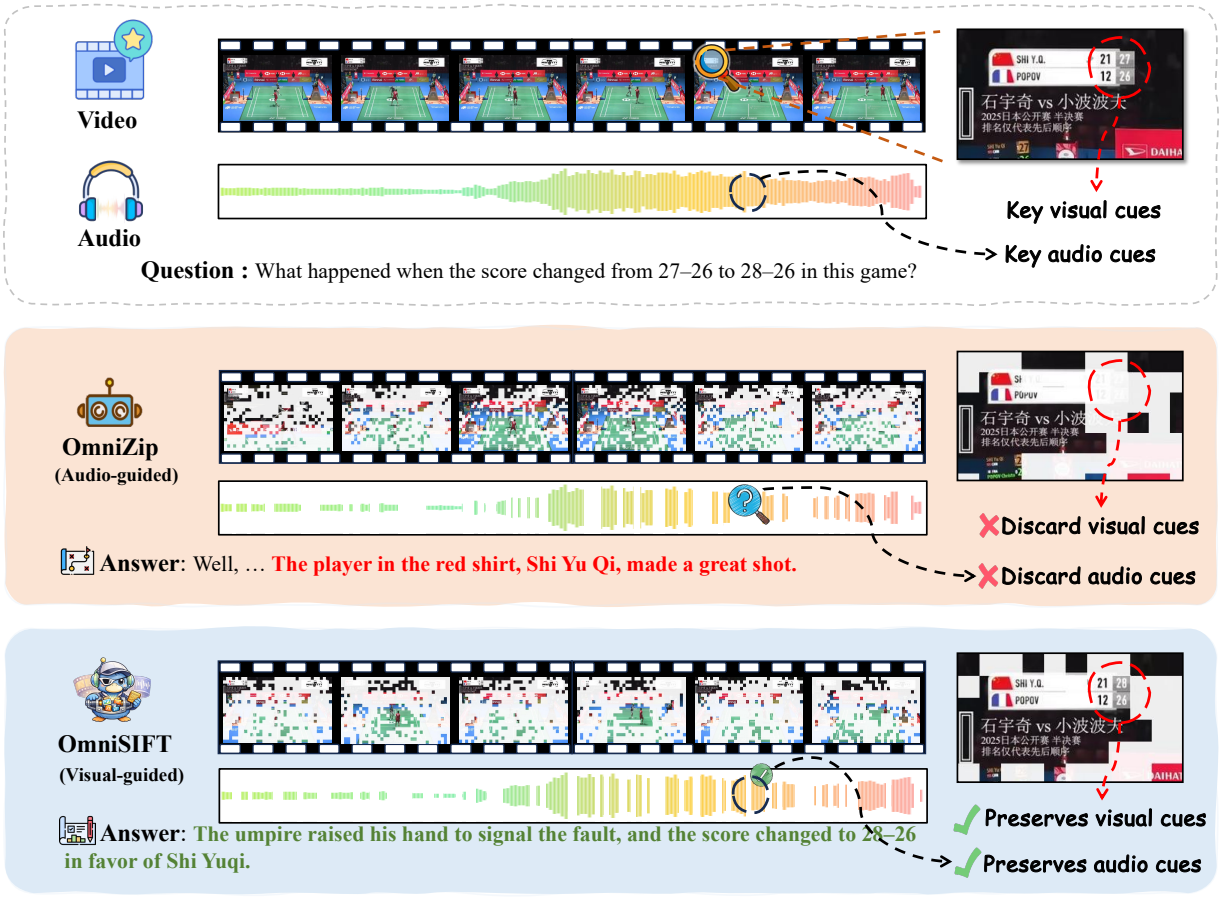

Figure 6. **Visualization of token compression methods for Omni-LLMs.** White blocks denote discarded video and audio tokens. The vertical amplitude of the waveform reflects the audio information density. OmniZip removes key visual information and audio cues, leading to an incorrect interpretation of the score change. In contrast, OmniSIFT preserves both the salient visual transitions and the informative audio segments required for correct event reasoning.

Table 4. **Ablation results for compression paradigm**. We compare our video-guided audio compression with an OmniZip-style trained compression method. All experiments use the Qwen2.5-Omni-7B backbone and evaluate three retained ratios on Daily-Omni and WorldSense. The **best** results are bolded.

Method	Retained Ratio (%)	Daily-Omni	World-Sense
OmniZip-Trained	35	70.5	49.7
OmniSIFT	35	**73.2**	**50.0**
OmniZip-Trained	30	69.3	49.3
OmniSIFT	30	**72.8**	**50.0**
OmniZip-Trained	25	68.8	48.7
OmniSIFT	25	**72.5**	**49.9**

dio and video signals at the same time carry comparable importance. In this example, when the score changes, the audio signal receives a low saliency score and allocates a small compression budget to video; as a result, the scoreboard patches are pruned, yielding an incorrect answer. In contrast, OmniSIFT adopts a modality-asymmetric compression that preserves the salient video patches and contextually informative audio cues necessary for correct reasoning.

5. Conclusion

In this work, we introduce OmniSIFT, a modality-asymmetric token compression framework for Omni-LLMs. Inspired by the asymmetric nature of human audio–video perception, OmniSIFT first decouples spatial and temporal redundancy in video tokens to obtain compact visual cues, and then uses these cues to guide audio token selection. Experiments on five audio–visual benchmarks show that OmniSIFT consistently outperforms existing compression baselines and, in several settings, even exceeds the performance of full-token models. It also delivers substantial gains in inference speed and memory usage. Overall, OmniSIFT provides an effective and efficient approach for reducing token counts in Omni-LLMs while preserving the key audio–visual information required for downstream tasks.

Impact Statement

OmniSIFT improves the efficiency of Omni-modal LLMs by reducing redundant tokens while preserving or enhancing performance, enabling wider deployment in resource-constrained or real-time settings. By encouraging semantically meaningful, cross-modal representations, it benefits applications such as audio-visual QA and video captioning.

References

An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Arandjelovic, R. and Zisserman, A. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pp. 609–617, 2017.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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

Chen, X., Ding, Y., Lin, W., Hua, J., Yao, L., Shi, Y., Li, B., Zhang, Y., Liu, Q., Wan, P., et al. Avocado: An audio-visual video captioner driven by temporal orchestration. *arXiv preprint arXiv:2510.10395*, 2025.

Chen, Y., Xue, F., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024b.

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

Chowdhury, S., Nag, S., Dasgupta, S., Wang, Y., Elhoseiny, M., Gao, R., and Manocha, D. Avtrustbench: Assessing and enhancing reliability and robustness in audio-visual llms. 2025.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

Gong, C., Wang, D., Wei, Z., Guo, Y., Zhu, H., and Chen, J. Echoingpixels: Cross-modal adaptive token reduction for efficient audio-visual llms. *arXiv preprint arXiv:2512.10324*, 2025.

Hong, J., Yan, S., Cai, J., Jiang, X., Hu, Y., and Xie, W. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Ji, S., Jiang, Z., Wang, W., Chen, Y., Fang, M., Zuo, J., Yang, Q., Cheng, X., Wang, Z., Li, R., et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Jiang, J., Li, X., Liu, Z., Li, M., Chen, G., Li, Z., Huang, D.-A., Liu, G., Yu, Z., Keutzer, K., et al. Storm: Token-efficient long video understanding for multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5830–5841, 2025a.

Jiang, S., Liang, J., Wang, J., Dong, X., Chang, H., Yu, W., Du, J., Liu, M., and Qin, B. From specific-mllms to omni-mllms: a survey on mllms aligned with multi-modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8617–8652, 2025b.

Koppen, C., Alsius, A., and Spence, C. Semantic congruency and the colavita visual dominance effect. *Experimental brain research*, 184(4):533–546, 2008.

Li, C., Chen, Y., Ji, Y., Xu, J., Cui, Z., Li, S., Zhang, Y., Tang, J., Song, Z., Zhang, D., et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025.

Liu, K., Li, J., Sun, Y., Wu, S., Gao, J., Zhang, D., Zhang, W., Jin, S., Yu, S., Zhan, G., et al. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. *arXiv preprint arXiv:2512.22905*, 2025a.

Liu, X., Gui, X., Zhang, Y., and Zhang, L. Mixing importance with diversity: Joint optimization for kv cache compression in large vision-language models. *arXiv preprint arXiv:2510.20707*, 2025b.

Liu, X., Wang, Y., Ma, J., and Zhang, L. Video compression commander: Plug-and-play inference acceleration for video large language models. *arXiv preprint arXiv:2505.14454*, 2025c.

Liu, X., Wang, Z., Chen, J., Han, Y., Wang, Y., Yuan, J., Song, J., Zhang, L., Huang, S., and Chen, H. Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. *arXiv preprint arXiv:2501.05179*, 2025d.

Seo, P. H., Nagrani, A., and Schmid, C. Avformer: Injecting vision into frozen speech models for zero-shot av-asr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22922–22931, 2023.

Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., and Dao, T. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

Tang, C., Li, Y., Yang, Y., Zhuang, J., Sun, G., Li, W., Ma, Z., and Zhang, C. video-salmonn 2: Captioning-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*, 2025.

Tao, K., Qin, C., You, H., Sui, Y., and Wang, H. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18992–19001, 2025a.

Tao, K., Shao, K., Yu, B., Wang, W., liu, J., and Wang, H. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. *arXiv preprint arXiv:2511.14582*, 2025b.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., and Lin, J. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., and Jia, J. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19792–19802, 2025.

Yao, L., Li, Y., Wei, Y., Li, L., Ren, S., Liu, Y., Ouyang, K., Wang, L., Li, S., Li, S., et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025.

Ye, W., Wu, Q., Lin, W., and Zhou, Y. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 22128–22136, 2025.

Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., and Torralba, A. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 570–586, 2018.

Zhou, Z., Wang, R., and Wu, Z. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

A. Expanded Benchmark Details

To comprehensively evaluate OmniSIFT across diverse audio-visual understanding regimes, we select five representative benchmarks that jointly cover: (i) long-horizon temporal understanding, (ii) audio–visual fusion and cross-modal alignment, (iii) fine-grained multi-dimensional comprehension, and (iv) generative captioning. Our selection follows two principles: *capability coverage*—including competencies most relevant to practical omni-modal assistants (e.g., temporal integration, causal reasoning, and modality interaction), and *protocol availability*—adopting benchmarks with public, standardized evaluation protocols to ensure reproducible comparisons. Table 5 summarizes the benchmarks, dataset scales, and metrics used in this work.

Table 5. Evaluation benchmarks used in this work.

Benchmark	#Videos	#QA	#Caps	Metric
DailyOmni	684	1,197	–	Acc. (overall & by QA type)
Video-MME	900	2,700	–	Acc.
WorldSense	1,662	3,172	–	Acc.
OmniVideoBench	628	1,000	–	Acc.
video-SALMONN-2	483	–	483	GPT-judge (Comp., Hall.)

"–" indicates not applicable (QA-only or caption-only evaluation).

B. Expanded Implementation Details

Frame Rating and Pixels Settings During both training and evaluation, video inputs are sampled at a rate of 2 frames per second, with a total frame count capped at 256 frames. The maximum resolution per frame is limited to 320×28×28 pixels.

Selecting Visual and Audio Compression Ratios Table 6 shows the specific rho_v and rho_a values we select for different retained ratios during evaluation. Specifically, for a total retained ratio of 35%, we choose audio and video compression ratios based on the compression ratios outlined in OmniZip. Considering the varying actual compression rates across different compression methods, we dynamically adjust these ratios to ensure that the proportion of audio and video tokens actually retained across different methods remained roughly equivalent. At the 25% retained ratio, we determine the optimal audio-video token compression ratios through multiple experiments, again striving to maintain similar audio-video tokens ratios across different compression methods.

Evaluation Prompt Figure 7 shows the prompt we use when evaluating QA benchmarks like Videomme, DailyOmni, WorldSense and OmniVideoBench. Figure 8 shows the prompt we use when evaluating the video-SALMONN-2 tsetset and the prompt used to evaluate model captions with GPT-4.1 is shown in Figure 9.

Table 6. ρ_v (video) and ρ_a (audio) for different retrained ratios.

Methods	35%		25%	
	ρ_a	ρ_v	ρ_a	ρ_v
OmniZip	0.4	0.7	0.6	0.98
DyCoke	0.4	0.9	0.6	0.99
Random	0.4	0.67	0.5	0.77
OmniSIFT	0.4	0.67	0.5	0.77

C. Computing Cost Evaluation

The computational overhead of OmniSIFT consists of two parts: the **TransformerAudioSelector** parameters and the **similarity-based pruning** operations. We analyze these costs relative to the Qwen2.5-Omni-7B backbone ($d_{model} = 3584$).

C.1. Parameter Efficiency

The TransformerAudioSelector is designed to be extremely lightweight. Given the dimensions $d_{model} = 3584$ and $hidden\_dim = 512$, the parameter count is calculated as follows:

- **Projections**: Two linear layers (v_proj, a_proj) map features from 3584 to 512 dimensions, totaling $2 \times (3584 \times 512) \approx 3.67$M parameters.
- **Cross-Attention**: A single-layer Multi-head Attention (MHA) with $embed\_dim = 512$ accounts for approximately $4 \times (512^2) \approx 1.05$M parameters.
- **Score Head**: A small MLP ($512 \to 256 \to 1$) accounts for ≈ 0.13M parameters.

The total parameter count is approximately **4.85M**, which is less than **0.1%** of the total parameters in a typical 7B-class LLM backbone.

C.2. Computational Complexity (FLOPs)

The complexity of our compression unit is significantly lower than the self-attention mechanism in the LLM:

1. **Spatial and Temporal Pruning**: These operations rely on cosine similarity and Top-K selection. For a frame with N patches, the complexity is $O(N \cdot d_{model})$, which is linear with respect to the sequence length.
2. **Audio Selection**: We perform cross-attention where the audio tokens (L_a) act as queries and compressed video tokens (L_v) act as keys/values. The complexity is $O(L_a \cdot L_v \cdot hidden\_dim)$. Since this is performed within localized chunks and uses a reduced $hidden\_dim$ (512 vs 3584), the FLOPs are marginal.

Prompts for evaluation

Select the best answer to the following multiple-choice question based on the video. Respond with only the letter (A, B, C, or D) of the correct option.
What visual elements were displayed immediately after Dr. Rajani's 'BOTOX WITHOUT THE BOTOX' video concluded?
A. Still product bottle → Price text overlay
B. Facial treatment demonstration → Presenter holding product while explaining
C. Presenter's torso shot → Secondary screen activation
D. Bookshelf backdrop → Close-up of string lights
The best answer is:

Figure 7. Prompts for evaluation in QA benchmarks.

Prompts for evaluation

Please provide a thorough description of all the content in the video, including every detail. As you describe, ensure that you also cover as much information from the audio as possible, and be mindful of the synchronization between the audio and video as you do so.

Figure 8. Prompts for evaluation in video-SALMONN-2 testset.

Table 7. Efficiency comparison in theoretical FLOPs in evaluation on the WorldSense with Qwen2.5-Omni-7B.

Method	Retained Ratio	Selector FLOPs (T)	LLM FLOPs(T)	Total FLOPs (T)
Full Tokens	100%	/	555.74	555.74
OmniSIFT	35%	0.06	292.10	292.16
OmniSIFT	25%	0.04	250.79	250.83

C.3. Theoretical Speedup and Latency

The primary bottleneck in Omni models is the quadratic complexity $O(L^2)$ of the LLM's self-attention, where L is the total number of multi-modal tokens. By reducing the video tokens by r_v and audio tokens by r_a, OmniSIFT reduces the total sequence length from L to L'.

The total time cost T_{total} can be expressed as:

$$T_{total} = T_{selector}(L) + T_{LLM}(L') \tag{9}$$

where $T_{selector}$ is the overhead of our method and T_{LLM} is the backbone inference time. Since $T_{selector} \ll T_{LLM}$ and $L' < L$, the reduction in T_{LLM} (which scales quadratically with L) far outweighs the linear overhead of $T_{selector}$. In our experiments, OmniSIFT achieves a significant reduction in Time-to-First-Token (TTFT) and total inference latency while maintaining high performance.

D. More Experimental Results

D.1. Visualization of Attention Sparsity

To investigate the internal mechanism of multi-modal understanding and justify the rationale for token compression, we visualize the attention score distributions of the Qwen2.5-Omni-7B backbone. We specifically analyze Layer 15 (middle layer) and Layer 27 (deep layer) during the inference process of an audio-video task.

As shown in Figure 10, we observe a high degree of **attention sparsity** across both layers:

- **Concentrated Information**: The majority of the attention scores are near zero (represented by the dark regions in the log-scale heatmap). Only a small subset of tokens—primarily certain pivotal video patches and audio segments—receive significant attention weights.
- **Modality Redundancy**: A large proportion of video (yellow) and audio (green) tokens contribute minimally to the final output generation. This suggests that the original dense representation contains substantial redundant information that does not influence the model's decision-making.
- **Consistency across Layers**: This sparsity pattern persists from the middle layers to the deeper layers. It indicates that the LLM naturally learns to filter out irrelevant spatial-temporal details to focus on high-level semantics.

Prompts used to evaluate captions in video-SALMONN-2

Task: A good video description should capture the detailed events in the video. The task is to judge whether a given description is good or not. The model is provided with a list of base events and a candidate description, and must determine which base events are covered by the description.

Instruction: Besides correctly described events, the description may contain missed events, incorrect events, or hallucinated events.

- **Missed Event:** An event in the base set whose main action, participants, and context are absent from the description.
- **Incorrect Event:** An event in the base set that is mentioned but described with significant factual errors.
- **Hallucination Event:** An event mentioned in the description but not included in the base set and not a plausible inference from it.

The model must also enumerate these events. Incorrect and hallucination events must not overlap.

Input Format:

- There are {`event_num`} base events given as a Python list: `["xxx", ...]`.
- A video description to be evaluated is provided.

Output Format (strict): {"Missed": x, "Incorrect": x, "Hallucination": x, "Missed Event": [...], "Incorrect Event": [...], "Hallucination Event": [...]}

Events in the Video: {`events_in_video`}

Video Description to be Rated: {`cap_to_be_rated`}

Given the base events and the candidate description, count missed, incorrect, and hallucinated events and list them out.

Figure 9. Prompt for caption evaluation on video-SALMONN-2 test set.

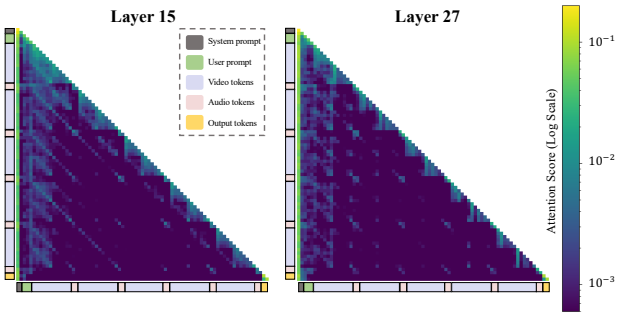

Figure 10. Attention score distribution maps for layers 15 and 27 of the LLM decoder in the Qwen2.5-Omni-7B model.

These empirical observations provide a strong motivation for **OmniSIFT**. By identifying and removing these low-contribution tokens through our selector and similarity-based pruning, we can significantly reduce the computational load without compromising the model's representative power.

D.2. Efficiency Gains across Video Lengths

To further demonstrate the necessity and effectiveness of token compression in long-video understanding, we analyze the inference efficiency of OmniSIFT on the WorldSense benchmark across varying video durations (from 0s to 120s). We compare our method against the baseline (Full Tokens) in terms of end-to-end (E2E) latency and peak GPU memory usage.

As illustrated in Figure 11, we observe several key trends:

- **Latency Reduction**: As video length increases, the E2E latency of the baseline model grows significantly due to the quadratic complexity of self-attention. OmniSIFT consistently achieves a latency reduction ratio of over 60% for videos longer than 60 seconds. This confirms that our compression strategy effectively mitigates the computational bottleneck for long-form content.

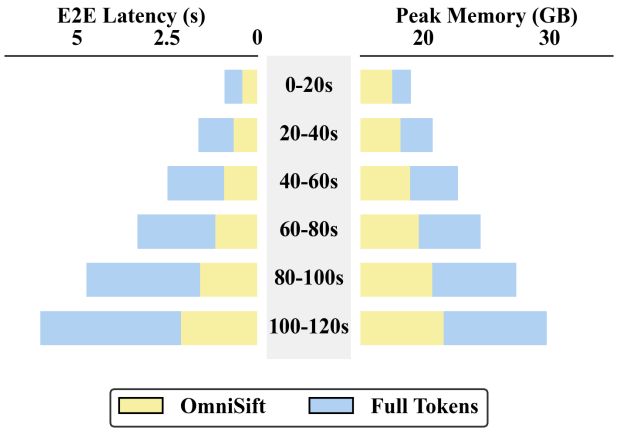

Figure 11. Peak Memory and End-to-End Latency per example for OmniSift and full tokens on Qwen2.5-Omni-7B when reasoning over videos of varying durations in Worldsense.

- **Memory Scalability**: The peak memory consumption of the baseline model increases rapidly with video length, posing a challenge for deployment on consumer-grade hardware. In contrast, OmniSIFT maintains a much slower growth rate in memory usage. Notably, the *Peak Memory Reduction Ratio* increases as the video becomes longer, reaching approximately 28% at 120s.
- **Criticality for Long Videos**: The widening gap between the "Full Tokens" and "OmniSIFT" curves highlights that token compression is not merely an optimization but a necessity for scaling Omni models to handle extended temporal contexts.

Through this analysis, we demonstrate that OmniSIFT provides a more sustainable scaling curve, enabling high-quality audio-visual understanding under restricted computational budgets.

D.3. Ablation on Selector Depth

To verify whether the complexity of the **TransformerAudioSelector** impacts the pruning quality, we conduct a comparative study between our default single-layer ($N = 1$) design and a deeper 3-layer ($N = 3$) variant. The goal is to determine if increasing the depth of cross-modal interaction leads to better selection of representative audio tokens.

As summarized in Table 8, the performance across multiple benchmarks remains largely stable when the selector depth is increased. For instance, the 3-layer variant yields only marginal gains (e.g., less than 0.1% on VideoMME), suggesting that the cross-modal correlation required for token selection is effectively captured even with a shallow architecture.

This result yields two key insights:

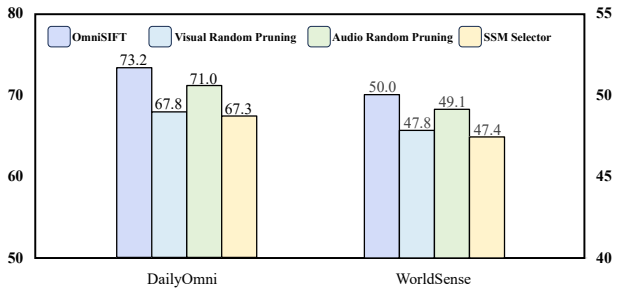

Figure 12. Results of extended ablation experiments conducted on the archi-tecture of OmniSift. **Visual Random Pruning**: randomly pruning visual tokens; **SSM Selector**: Using a State Space Model as a selector for audio tokens; Audio-Only Selector: randomly pruning audio tokens.

- **Information Sufficiency**: A single cross-attention layer is sufficient to identify the redundancy in audio tokens relative to the visual context. The bottleneck for token compression in Omni models lies more in the selection strategy itself than in the depth of the selector module.
- **Efficiency-Performance Trade-off**: While the 3-layer variant maintains comparable performance, it introduces additional computational latency. Given our objective of accelerating long-video inference, the single-layer configuration is the optimal choice as it minimizes the Time-to-First-Token (TTFT) while preserving high accuracy.

Table 8. Performance comparison of different selector depths at a 35% retained ratio on selected benchmarks. Performance is saturated at $N = 1$, making it the optimal choice for long video compression.

Configuration	Video-MME (%)	Daily-Omni (%)	World-Sense (%)	GPU Mem (GB) ↓
1-Layer (Ours)	**68.3**	**73.2**	**50.0**	**22.62**
3-Layer Variant	67.2	72.3	49.0	22.67

D.4. Extended Ablation Results

In addition to the experiments in Section 4.4, we conduct further ablation studies on the architecture of OmniSIFT. These results demonstrate that both the STVP and VGAS modules are essential for accurately capturing audio-visual dependencies and pruning redundant tokens. Furthermore, we observe that randomly pruning audio tokens yields better results than randomly pruning visual tokens, confirming the modality-asymmetric nature between audio and visual inputs. This indicates that incorrect visual inputs are more challenging for the model to correct than erroneous audio token inputs.