OMNIRELAY: INFORMATION-FLOW-GUIDED TOKEN COMPRESSION FOR OMNI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (Omni-LLMs) achieve strong performance on audiovisual understanding tasks, but processing long audiovisual token sequences incurs substantial computational and memory costs. To reduce these costs, existing token compression methods exploit local redundancy at the *pre-LLM stage* and task relevance at the *within-LLM stage*, but make limited use of global modality dependencies. To determine when these dependencies should guide compression, we analyze attention patterns and perform attention knockout across layers. We observe a depth-varying relay: audiovisual information is integrated through intra-modal interactions within and across windows in early layers, transferred to text in middle layers, and finally relayed to the final input token for prediction. We therefore propose **OmniRelay**, a training-free pre&within-LLM framework that adapts compression guidance across layers. Its ***Depth-Adaptive Relay Compression (DARC)*** progressively shifts from intra-modal to text guidance before removing the remaining audiovisual tokens at late layers. Complementing ***DARC***, our proposed ***Saliency–Diversity and Coverage-Aware Compression (SDCC)*** reduces local redundancy before the LLM while preserving salient and diverse tokens with broad cross-window coverage. We evaluate OmniRelay on 3 Omni-LLMs and 5 audiovisual benchmarks. It retains 99.51% of full-token performance at 25% layer-averaged audiovisual token retention, with 79% fewer FLOPs and a 3.3× prefill speedup. Code is available at `https://anonymous.4open.science/r/OmniRelay-FF35/`.

1 INTRODUCTION

Omni-modal large language models (Omni-LLMs) (Hurst et al., 2024; Comanici et al., 2025; Xu et al., 2025a;b; Fu et al., 2026; Yao et al., 2024) process video, audio and text as a unified token sequence within a single LLM, achieving competitive performance on joint audiovisual understanding tasks (Yang et al., 2025b; Hong et al., 2026; Fu et al., 2025a; Zhou et al., 2025). However, high-resolution or long-duration audiovisual inputs typically produce a large number of tokens. The computational cost of standard self-attention grows quadratically with sequence length, while longer sequences also increase memory requirements (Tay et al., 2022; Dao et al., 2022). These costs reduce inference efficiency and hinder deployment in resource-constrained settings.

Token compression addresses these costs by reducing the number of audiovisual tokens processed by Omni-LLMs (Gong et al., 2025; Li & Huang, 2026; Su et al., 2026a; Deng et al., 2026; Zhong et al., 2026). Depending on when compression is applied, existing approaches include *pre-LLM* and *pre&within-LLM* methods. *Pre-LLM* methods such as OmniZip (Tao et al., 2026) and OmniSIFT (Ding et al., 2026) use saliency signals from the encoders and cross-modal guidance to selectively reduce audio and video tokens before they enter the LLM. However, aggressive compression at this stage can discard task-relevant evidence before the LLM can integrate it with the task context (Xin et al., 2026; Park et al., 2026). To reduce reliance on early pruning, *pre&within-LLM* methods defer part of the compression to LLM layers, where token selection can exploit contextualized task-relevance signals. For instance, OmniDrop (Park et al., 2026) combines pre-LLM intra-modal redundancy reduction with progressive text-guided pruning. SEATS (Xin et al., 2026) performs window-wise pre-LLM selection and adapts within-LLM budgets across windows and modalities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

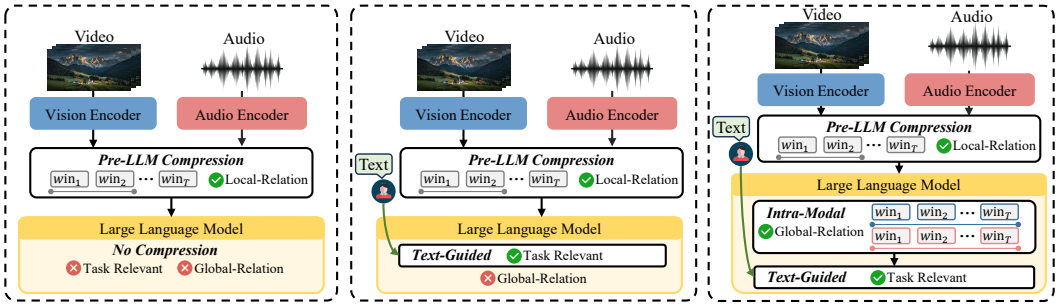


Figure 1: **Information utilization in different Omni-LLM token compression paradigms.** (a) Pre-LLM methods compress local redundancy before the LLM. (b) Existing pre&within-LLM methods further exploit text relevance within the LLM, but overlook global intra-modal dependencies. (c) OmniRelay (ours) exploits global intra-modal dependencies and text relevance at different layer depths following the information relay.

These *Pre&within-LLM* methods reduce local redundancy before the LLM. However, the main challenge remains within the LLM: as illustrated in Figure 1(b), existing methods primarily rely on text relevance for token selection, leaving global redundancy across the entire temporal sequence insufficiently resolved. Thus, within-LLM compression should account for both global modality dependencies and text relevance. This further raises a question: *When should these modality-related dependencies be considered, and when should text relevance instead be considered for compression within the LLM?* To investigate this question, we analyze attention patterns and perform layer-wise attention knockout in Section 2.2. Our analysis reveals a **depth-varying information relay**: in early layers, audiovisual information is processed primarily through intra-modal interactions within and across windows, with limited direct audio-video interaction. This information is progressively transferred to text tokens in middle layers and finally relayed to the final input token in late layers for prediction. Accordingly, intra-modal and text-guided compression should be applied at different layer depths to align with this information relay.

Motivated by these insights, we propose **OmniRelay**, a training-free pre&within-LLM token compression method for improving the efficiency of Omni-LLMs. The key design of OmniRelay is **D**epth-**A**daptive **R**elay **C**ompression **(DARC)** within the LLM. Following the information relay, DARC uses intra-modal attention in early layers to reduce global redundancy and then gradually shifts to text-guided compression, as shown in Figure 1(c). Once relevant modality information has been sufficiently transferred to text representations, the remaining audiovisual tokens are removed. To complement DARC, **S**aliency–**D**iversity and **C**overage-Aware **C**ompression **(SDCC)** is applied before the LLM to conservatively remove evident local redundancy using token saliency and diversity for token selection. It also leverages cross-window coverage for video budget allocation to preserve important evidence and global coverage for subsequent DARC. Together, OmniRelay progressively removes both local and global modality redundancy while preserving task-relevant information, thereby improving efficiency.

Experiments on three representative Omni-LLMs and five audiovisual understanding benchmarks demonstrate that OmniRelay achieves a state-of-the-art (SOTA) trade-off between performance and efficiency. Even at an average retention ratio of 25%, OmniRelay retains 99.51% of the full-token performance, reduces LLM FLOPs by 79% and achieves a 3.3× prefill speedup on Qwen2.5-Omni-7B (Xu et al., 2025a). The main contributions are summarized as follows:

- We identify a **depth-varying information relay in Omni-LLMs**: audiovisual information is primarily propagated through intra-modal interactions in early layers, progressively transferred to text in middle layers, and finally relayed to the final input token.
- We propose **OmniRelay, a training-free pre&within-LLM token compression method that aligns within-LLM compression with the information relay.** It progressively shifts from intra-modal to text-guided compression and removes remaining modality tokens finally within the LLM, complemented by a conservative pre-LLM compression.
- Extensive experiments across three Omni-LLMs and five audiovisual benchmarks demonstrate that OmniRelay achieves **a SOTA trade-off between performance and efficiency**.

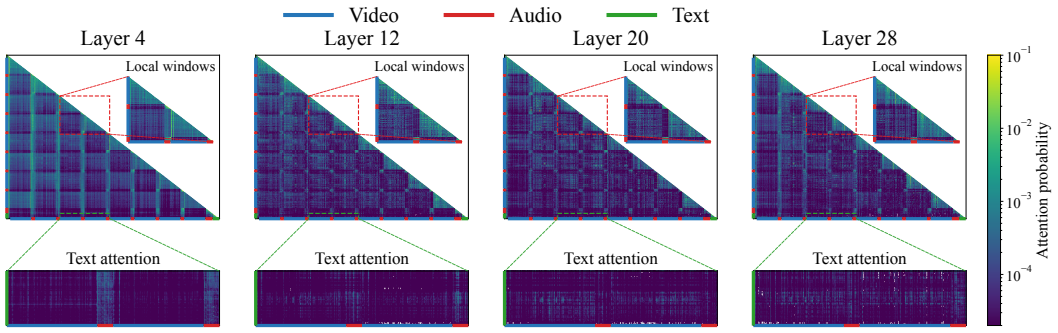


Figure 2: **Attention patterns across layers.** Enlarged views highlight interactions within modality windows and between adjacent windows, and text attention to modality tokens.

2 Preliminaries & Motivation

2.1 Preliminaries

Tokenization of Omni-LLMs. Omni-LLMs (Xu et al., 2025a;b; Yao et al., 2024) encode video and audio inputs into token representations and project them into the LLM embedding space. Let $\mathbf{V}$ and $\mathbf{A}$ denote all video and audio tokens, with n^v and n^a tokens, respectively. For w temporal windows, we partition them as $\mathbf{V} = [\mathbf{V}_1; \dots; \mathbf{V}_w]$ and $\mathbf{A} = [\mathbf{A}_1; \dots; \mathbf{A}_w]$, where $\mathbf{V}_t \in \mathbb{R}^{n_t^v \times d}$ and $\mathbf{A}_t \in \mathbb{R}^{n_t^a \times d}$ denote the video and audio tokens in the t-th window, $n^v = \sum_{t=1}^{w} n_t^v$, $n^a = \sum_{t=1}^{w} n_t^a$, and d is the hidden dimension. Given the system prompt embeddings $\mathbf{H}_{\text{sys}}$ and text embeddings $\mathbf{T}$, the temporally interleaved sequence is constructed as

$$\mathbf{H}^{(1)} = [\mathbf{H}_{\text{sys}}; \mathbf{V}_1; \mathbf{A}_1; \cdots; \mathbf{V}_w; \mathbf{A}_w; \mathbf{T}] \in \mathbb{R}^{N \times d}, \tag{1}$$

where $\mathbf{H}^{(1)}$ denotes the first-layer hidden states and N denotes the total sequence length.

Causal attention. Let H denote the number of query heads. At layer l, for each query head $h \in \{1, \dots, H\}$, the hidden states $\mathbf{H}^{(l)}$ are projected into queries and keys as $\mathbf{Q}^{(l,h)} = \mathbf{H}^{(l)}\mathbf{W}_Q^{(l,h)}$ and $\mathbf{K}^{(l,h)} = \mathbf{H}^{(l)}\mathbf{W}_K^{(l,h)}$, $\mathbf{W}_Q^{(l,h)}, \mathbf{W}_K^{(l,h)} \in \mathbb{R}^{d \times d_h}$, and d_h is the head dimension. The causal attention weights are computed as

$$\mathbf{P}^{(l,h)} = \text{softmax}\left(\frac{\mathbf{Q}^{(l,h)}(\mathbf{K}^{(l,h)})^\top}{\sqrt{d_h}} + \mathbf{M}\right), \tag{2}$$

where $\mathbf{M}$ is the causal mask, with $M_{i,j} = 0$ for $j \le i$ and $M_{i,j} = -\infty$ otherwise. $P_{i,j}^{(l,h)}$ denotes the attention weight from query token i to key token j at layer l and head h.

Attention Knockout. Attention knockout (Geva et al., 2023) is used to assess how information transferred through selected attention paths contributes to predictions. At the layer l, we replace $\mathbf{M}$ in Eq. (2) with $\mathbf{M}_l$, setting $M_{l,ij} = -\infty$ when query token i belongs to the target set and key token j belongs to the source set. This blocks information flow from the source tokens to the target tokens. Specifically, we use $X \nrightarrow Y$ to denote blocking information flow from token set X to token set Y and $X \nleftrightarrow Y$ to denote blocking both directions. We quantify the intervention effect as $\Delta p = (p_{\text{KO}} - p_{\text{base}})/p_{\text{base}} \times 100\%$, where p_{base} and p_{KO} denote the probabilities assigned to the same baseline-predicted answer token before and after the intervention.

2.2 Motivation

To determine when to use modal-related or text guidance for within-LLM compression, we examine attention interactions across layers, as attention establishes token dependencies through information transfer from attended tokens (Elhage et al., 2021).

2.2.1 Attention Patterns

We first inspect attention patterns to obtain a qualitative overview of how token dependencies are organized and evolve across layers, thereby identifying the key pathways for subsequent knockout

Table 1: **Accuracy of Qwen2.5-Omni-7B on AVUT and WorldSense under attention knockout applied to all layers.** We report accuracy (%) and changes from baseline in percentage points. A, V and T denote audio, video and text tokens, respectively.

Dataset	Baseline	$A \nleftrightarrow V$	$V \nrightarrow V$	$A \nrightarrow A$	$A \nrightarrow T$	$V \nrightarrow T$
AVUT	64.76	64.76 (0.00)	55.59 (−9.17)	52.94(−11.82)	55.02 (−9.74)	50.35(−14.41)
WorldSense	46.63	46.66 (+0.03)	41.14 (−5.49)	39.12 (−7.51)	39.60 (−7.03)	41.02 (−5.61)

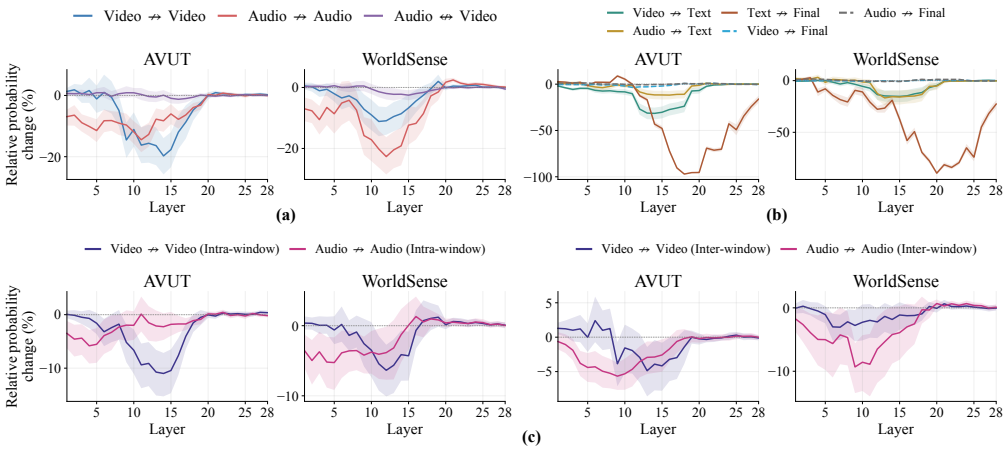


Figure 3: **Layer-wise attention knockout results for Qwen2.5-Omni-7B on AVUT and WorldSense.** (a) Intra-modal and direct audio–video interactions. (b) Modality-to-text, text-to-final and modality-to-final interactions. (c) Intra-window and inter-window intra-modal interactions.

analysis. As illustrated in Figure 2, we observe three patterns: (1) Intra-modal attention is consistently stronger than cross-modal attention, suggesting a dominant role for within-modal processing. (2) Intra-modal attention occurs both within and across windows indicating the contextual modeling capacity of LLM to capture global dependencies. (3) Early text queries focus on audio tokens also favored by other queries, while their attention allocation changes substantially in later layers.

2.2.2 Attention Knockout

We then use attention knockout (Kim et al., 2026; Suharitdamrong et al., 2026) to investigate how audiovisual information is transferred across layers to the final input token, which predicts the first output token. We conduct attention knockout on AVUT (Yang et al., 2025b) and WorldSense (Hong et al., 2026), as their questions involve sufficient audiovisual interactions and span diverse domains.

Observation (I): Different information pathways dominate at different depths. Figure 3 presents the layer-wise attention knockout results. Figure 3(a) shows that blocking intra-modal attention substantially reduces the baseline-predicted answer probability in early and middle layers, whereas blocking direct audio–video attention has little effect, consistent with the observed attention patterns. Figure 3(b) shows that modality-to-text interactions are most consequential in middle layers, followed by text-to-final interactions in later layers, while direct modality-to-final interactions have little effect. This corroborates the findings of Suharitdamrong et al. (2026). The accuracy results in Table 1 further confirm that direct audio–video knockout has little impact, whereas intra-modal and modality-to-text knockout cause substantial performance degradation.

Insight (I). Compression guidance should adapt across depth to align with the information relay. These results reveal a depth-varying information relay: key audiovisual information is processed within each modality in early layers, transferred to text tokens in middle layers and finally conveyed to the final input token in late layers. Compression should therefore shift from intra-modal guidance to text guidance across depth. Once the relevant information has been transferred to text, the remaining modality tokens can be removed.

Observation (II): Both intra- and inter-window interactions contribute to prediction. We further separate intra-modal interactions into intra- and inter-window pathways. As shown in Figure 3(c), blocking either pathway substantially impairs prediction, consistent with the attention patterns observed above. This confirms the global dependencies extend beyond each window.

Insight (II). Since pre-LLM compression primarily uses encoder features within individual windows, it may overlook tokens that are important in the global context. Within-LLM compression should therefore exploit intra-modal attention across windows to capture global dependencies and reduce the redundancy left by local pre-LLM compression.

3 The Proposed Method: OmniRelay

Motivated by the observations in Section 2.2, we propose **OmniRelay**, a training-free pre&within-LLM token compression method. As illustrated in Figure 4, before the LLM, **S**aliency–**D**iversity and **C**overage-Aware **C**ompression (**SDCC**) conservatively removes evident local redundancy while preserving salient and diverse evidence across windows. Then **D**epth-**A**daptive **R**elay **C**ompression (**DARC**) further reduces global redundancy through intra-modal attention in early layers, progressively shifts to text guidance and removes remaining modality tokens eventually.

3.1 Saliency–Diversity and Coverage-Aware Compression

SDCC combines intra-window saliency–diversity selection with inter-window coverage-aware video budget allocation, while retaining audio tokens at a fixed ratio.

Intra-Window Saliency–Diversity Selection. For modality $m \in \{v, a\}$, let $\{\mathbf{x}_{t,i}\}_{i=1}^{n_t^m}$ denote the token features in window t. We derive normalized saliency ranks $a_{t,i}$ from mean received encoder attention and set diversity weights $w_{t,i} = \epsilon_c + (1-\epsilon_c)a_{t,i}$, where ϵ_c is the saliency floor. The token distance combines feature dissimilarity and positional separation:

$$d_m(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) = 1 - \cos(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}) + \lambda_m \|\mathbf{p}_{t,i} - \mathbf{p}_{t,j}\|_2, \tag{3}$$

where $\mathbf{p}_{t,i}$ is the normalized 2D spatial position for video or 1D temporal position for audio within window t, and λ_m controls positional weighting.

Given previously selected indices $\mathcal{S}_t(k-1)$, we alternate saliency and diversity selection:

$$i_{t,k} = \arg \max_{i \notin \mathcal{S}_t(k-1)} \begin{cases} a_{t,i}, & k \text{ odd}, \\ w_{t,i} \min\limits_{j \in \mathcal{S}_t(k-1)} d_m(\mathbf{x}_{t,i}, \mathbf{x}_{t,j}), & k \text{ even}. \end{cases} \tag{4}$$

Inter-Window Coverage-Aware Budget Allocation. We group similar consecutive video windows, selecting a medoid window to represent each group. For window t with k selected candidates, the coverage loss is

$$L_t(k) = \frac{\sum_{i=1}^{n_t^v} w_{t,i} \min_{j \in \mathcal{S}_t(k)} d_v(\mathbf{x}_{t,i}, \mathbf{x}_{t,j})}{\sum_{i=1}^{n_t^v} w_{t,i}}. \tag{5}$$

The decrease in $L_t(k)$ after adding a candidate defines its marginal coverage gain. For non-medoid windows, higher similarity to medoid tokens leads to a stronger gain discount. Dynamic programming maximizes cumulative discounted gains under the total budget and per-window bounds; each window retains candidates in selection order up to its allocated budget.

Audio windows use a fixed retention ratio due to their low cross-window token similarity.

3.2 Depth-Adaptive Relay Compression

Guided by the observation in Section 2.2, DARC progressively shifts audiovisual token importance estimation from intra-modal anchor guidance toward text guidance according to information flow, ultimately removing all remaining audiovisual tokens in late layers.

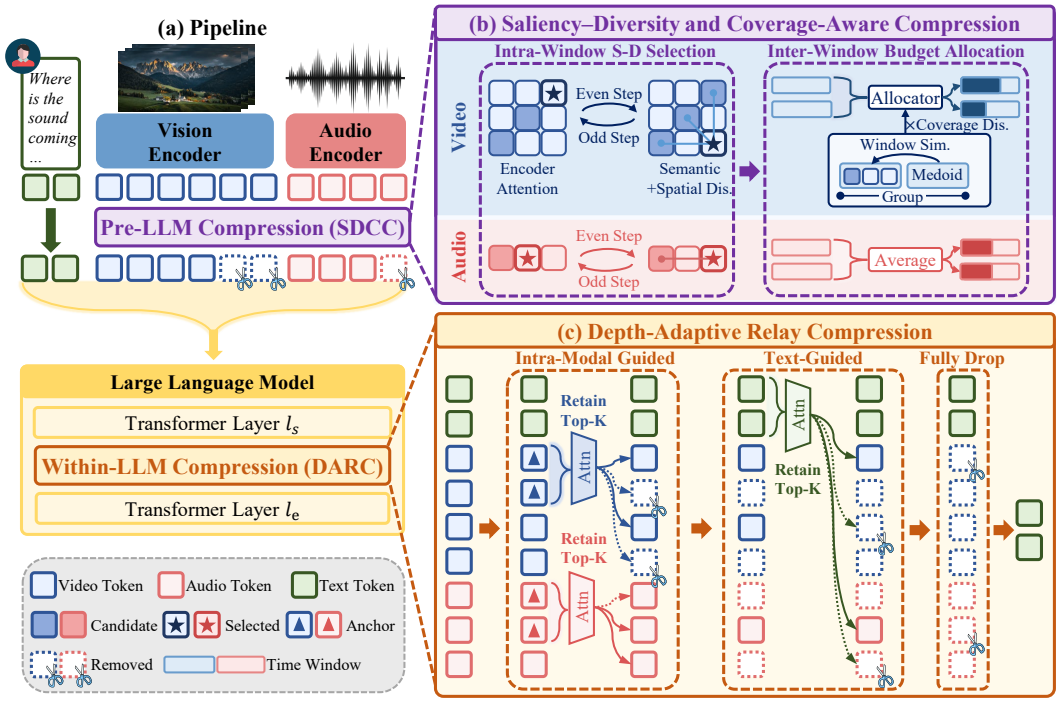


Figure 4: **Overview of OmniRelay. (a)** The pipeline integrating pre-LLM compression and within-LLM compression. **(b)** Details of Saliency–Diversity and Coverage-Aware Compression (SDCC). **(c)** Details of Depth-Adaptive Relay Compression (DARC).

Intra-Modal Guided Compression. At layer l_s, we use the video tokens retained from the medoid windows identified by SDCC and the last window tokens as video anchors, while using all retained audio tokens as audio anchors. At layer l, the surviving anchors of modality m serve as attention queries, with their indices forming the set $\mathcal{Q}_m^{(l)}$. For each token i of modality m, we compute its anchor-guided importance by averaging the attention weights it receives from these anchors across all heads:

$$s_{m,i}^{(l)} = \frac{1}{H|\mathcal{Q}_m^{(l)}|} \sum_{h=1}^{H} \sum_{q \in \mathcal{Q}_m^{(l)}} P_{q,i}^{(l,h)}. \tag{6}$$

For each modality m, we independently retain the top-$K_m^{(l)}$ tokens according to these scores, where $K_m^{(l)}$ is its token budget at layer l.

Text-Guided Compression. As audiovisual information is progressively transferred to text representations at layers l_t, we smoothly transition from intra-modal anchor guidance to text guidance. Since these information flow stages overlap across layers, we independently normalize the two importance scores and combine them through a weighted sum that increasingly favors text guidance at greater depths. Let $\mathcal{Q}_{\text{text}}^{(l)}$ denote the instruction token indices at layer l. All instruction tokens are retained throughout compression. For each audiovisual token i, we compute its text-guided importance score as

$$s_{\text{text},i}^{(l)} = \frac{1}{H|\mathcal{Q}_{\text{text}}^{(l)}|} \sum_{h=1}^{H} \sum_{q \in \mathcal{Q}_{\text{text}}^{(l)}} P_{q,i}^{(l,h)}. \tag{7}$$

We jointly rank video and audio tokens by the combined importance scores and retain the top-$K^{(l)}$ tokens, where $K^{(l)}$ is the total audiovisual token budget at layer l.

Late Modality Drop. Our information flow analysis indicates that deeper text representations have accumulated audiovisual information. Following SEATS (Xin et al., 2026), we remove all remaining

Table 2: **Performance comparison across various benchmarks under different retention ratios.** Each cell reports results for these three models in this order Qwen2.5-Omni-7B/Qwen2.5-Omni-3B/MiniCPM-o-2.6. The best and second-best compressed results for each model within each retention setting are highlighted in **bold** and underlined. Relative Performance denotes the mean performance relative to the full-token baseline. A dash (-) denotes unavailable results.

Method	AVUT	WorldSense	DailyOmni	Video-MME	OmniVideo	Relative Performance
Full Tokens	64.76 / 61.88 / 62.40	46.63 / 46.44 / 43.63	62.91 / 61.24 / 60.40	65.37 / 63.48 / OOM	35.90 / 32.90 / 34.30	100.00% / 100.00% / 100.00%
25% Retention Ratio						
Random	60.09 / 57.27 / 56.17	43.54 / 42.56 / 41.27	56.22 / 53.97 / 56.73	64.56 / 61.03 / -	35.80 / 33.10 / 32.90	94.80% / 93.81% / 93.61%
FastV-om	58.65 / 54.21 / 52.31	42.97 / 43.44 / 40.10	56.31 / 55.05 / 54.55	64.30 / 61.19 / -	35.30 / 33.80 / **33.80**	93.78% / 94.03% / 91.15%
VisionZip-om	58.88 / 57.61 / 51.61	44.10 / 44.96 / 39.79	58.06 / 56.31 / 54.22	65.44 / 62.67 / -	35.80 / **34.00** / 32.60	95.52% / 96.79% / 89.68%
OmniZip	61.71 / 59.57 / 58.54	44.74 / 45.43 / 41.42	58.06 / 57.39 / 58.23	**66.56** / 63.41 / -	35.20 / 33.80 / 32.40	96.68% / 98.09% / 94.90%
SEATS	63.09 / - / 59.40	45.30 / - / 43.03	59.82 / - / 58.98	66.00 / - / -	**36.20** / - / 32.20	98.29% / - / 96.34%
OmniFocus	59.75 / 58.02 / 50.63	43.41 / 43.38 / 38.90	56.89 / 55.97 / 54.47	65.00 / 61.74 / -	35.90 / 33.20 / 32.40	95.04% / 95.35% / 88.73%
OmniRelay	**64.65 / 60.61 / 59.57**	**46.00 / 45.49 / 43.06**	**60.99 / 58.90 / 59.15**	**66.56 / 63.81** / -	36.00 / 33.90 / 33.10	**99.51% / 99.13% / 97.15%**
15% Retention Ratio						
Random	55.71 / 55.07 / 52.94	41.74 / 40.61 / 39.91	53.05 / 52.05 / 55.39	62.44 / 58.81 / -	33.60 / 31.60 / 32.00	89.80% / 90.03% / 90.33%
FastV-om	50.87 / 50.75 / 46.37	40.54 / 41.27 / 37.74	53.30 / 50.63 / 49.87	55.00 / 55.00 / -	33.60 / 33.70 / 32.40	85.59% / 88.53% / 84.46%
VisionZip-om	57.15 / 55.19 / 46.54	43.16 / 43.32 / 38.56	54.22 / 53.63 / 49.87	61.67 / 59.56 / -	34.50 / **34.40** / 32.20	91.49% / 93.69% / 84.85%
SEATS	59.75 / - / 56.23	44.33 / - / **41.65**	57.14 / - / 56.89	65.19 / - / -	**35.90** / - / 33.40	95.58% / - / 94.28%
OmniFocus	56.06 / 54.67 / 46.14	41.80 / 42.06 / 35.53	54.39 / 53.38 / 48.37	60.48 / 57.89 / -	34.10 / 32.90 / 31.30	90.03% / 91.46% / 81.68%
OmniRelay	**60.84 / 58.71 / 57.32**	**45.68 / 44.58 / 41.65**	**59.48 / 57.39 / 57.31**	**66.74 / 63.37** / -	**35.90** / 33.70 / **33.60**	**97.71% / 97.37% / 95.04%**
10% Retention Ratio						
Random	54.27 / 52.48 / 49.71	40.01 / 39.56 / 39.41	50.96 / 48.20 / 51.13	60.11 / 58.00 / -	32.30 / 30.00 / 32.70	86.51% / 86.25% / 87.49%
FastV-om	42.96 / 38.00 / 42.04	38.02 / 35.15 / 37.20	47.20 / 41.44 / 46.12	50.22 / 47.48 / -	30.40 / 29.30 / 32.10	76.88% / 73.72% / 80.64%
VisionZip-om	55.07 / 52.88 / 44.12	41.49 / 42.43 / 36.54	52.97 / 53.05 / 47.28	59.00 / 56.56 / -	32.70 / 31.20 / 31.60	87.91% / 89.48% / 81.22%
SEATS	58.65 / - / 51.44	43.03 / - / 39.91	56.06 / - / 52.63	63.48 / - / -	35.30 / - / 32.90	93.48% / - / 89.24%
OmniFocus	53.00 / 52.36 / 39.45	41.39 / 40.57 / 34.71	52.13 / 51.38 / 43.02	56.70 / 55.11 / -	33.80 / 31.10 / 31.30	86.87% / 87.44% / 76.31%
OmniRelay	**59.86 / 57.84 / 52.42**	**44.23 / 43.69 / 40.57**	**57.31 / 56.73 / 54.05**	**64.89 / 61.93** / -	**36.00 / 32.00 / 33.00**	**95.59% / 95.00% / 90.67%**

video and audio tokens after a designated layer l_e. Subsequent layers operate only on text tokens, eliminating further computation on audiovisual tokens.

4 Experiments

4.1 Experimental Settings

Benchmarks. We evaluate OmniRelay on five widely used audiovisual benchmarks: AVUT (Yang et al., 2025b), WorldSense (Hong et al., 2026), Daily-Omni (Zhou et al., 2025), Video-MME (Fu et al., 2025a), and OmniVideoBench (Li et al., 2026). These benchmarks span diverse video durations, domains and task types ranging from recognition, counting to reasoning.

Compared Methods. We compare OmniRelay with representative training-free token compression baselines. These include FastV-om and VisionZip-om, which adapt the vision-oriented methods FastV (Chen et al., 2024) and VisionZip (Yang et al., 2025a) to Omni-LLMs by applying their respective compression strategies to both visual and audio tokens. We also evaluate recent methods designed for Omni-LLMs: OmniZip (Tao et al., 2026), OmniFocus (Cao et al., 2026), and SEATS (Xin et al., 2026). Random samples tokens uniformly at random from each modality.

Implementation Details. We conduct experiments on Qwen2.5-Omni-3B/7B (Xu et al., 2025a) and MiniCPM-o-2.6 (Yao et al., 2024) as the backbone models. All experiments are conducted on NVIDIA RTX A6000 GPUs, each with 48 GB of memory. We set the maximum number of sampled frames to 768 for Video-MME, 384 for OmniVideoBench, and 128 for the other benchmarks. We evaluate all methods at the same layer-averaged audiovisual token retention ratios $R_{\text{mean}} \in \{25\%, 15\%, 10\%\}$. Further details are provided in Appendix C.

4.2 Main Results

We omit SEATS on Qwen2.5-Omni-3B due to its unsupported 36-layer architecture. MiniCPM-o-2.6 runs out of memory on VideoMME due to long token sequences from uncapped audio inputs.

Table 3: **Efficiency analysis.** FLOPs, prefill latency, peak GPU memory and accuracy score on DailyOmni. Each cell reports results in the order Qwen2.5-Omni-7B / Qwen2.5-Omni-3B / MiniCPM-o-2.6. A dash (-) denotes unavailable results.

Method	Retention	FLOPs (T) ↓	Prefill (s) ↓	Peak GPU Memory (GB) ↓	Score (%) ↑
Full Tokens	100%	73.2 / 37.4 / 73.2	1.141 / 0.575 / 1.165	24.63 / 14.66 / 19.92	62.91 / 61.24 / 60.40
Random	25%	15.0 / 6.8 / 14.7	0.281 / 0.144 / 0.395	22.04 / 12.24 / 17.52	56.22 / 53.97 / 56.73
FastV-om	25%	15.9 / 7.3 / 15.6	0.293 / 0.154 / 0.407	22.42 / 12.12 / 17.57	56.31 / 55.05 / 54.55
VisionZip-om	25%	15.0 / 6.8 / 14.7	0.276 / 0.141 / 0.396	22.03 / 12.23 / 17.45	58.06 / 56.31 / 54.22
OmniZip	25%	14.9 / 6.8 / 14.7	0.348 / 0.176 / 0.433	22.24 / 12.42 / 17.61	58.06 / 57.39 / 58.23
SEATS	25%	15.5 / – / 15.1	0.305 / – / 0.429	21.69 / – / 16.90	59.82 / – / 58.98
SEATS	15%	9.1 / – / 8.8	0.197 / – / 0.333	21.46 / – / 16.70	57.14 / – / 56.89
OmniFocus	25%	15.0 / 6.8 / 14.9	0.282 / 0.145 / 0.400	22.04 / 12.24 / 17.47	56.89 / 55.97 / 54.47
OmniFocus	15%	9.0 / 4.0 / 8.9	0.188 / 0.094 / 0.297	21.69 / 11.91 / 17.15	54.39 / 53.38 / 48.37
OmniRelay	25%	15.5 / 7.3 / 16.4	0.344 / 0.215 / 0.478	21.78 / 11.76 / 16.96	60.99 / 58.90 / 59.15
OmniRelay	15%	9.2 / 4.2 / 9.6	0.235 / 0.164 / 0.370	21.55 / 11.62 / 16.74	59.48 / 57.39 / 57.31
OmniRelay	10%	6.1 / 2.7 / 6.5	0.182 / 0.139 / 0.314	21.40 / 11.53 / 16.72	57.31 / 56.73 / 54.05

Overall Performance. As shown in Table 2, OmniRelay achieves SOTA relative performance across three Omni-LLMs and five benchmarks at all three retention ratios. Take the setting of Qwen2.5-Omni-7B as an example, at 25% retention, it preserves 99.51% of the full-token baseline performance with minimal degradation and maintains 97.71% and 95.59% at 15% and 10% retention. The gains are particularly pronounced over variants of image- and video-oriented methods, exceeding their mean relative performance by 4.86 percentage points at 25% retention. OmniRelay also outperforms the pre-LLM methods OmniZip and OmniFocus at their evaluated retention ratios, with a mean gain of 3.65% at 25% retention. Even compared with SEATS, the strongest competing method at all three settings, OmniRelay achieves gains of 1.22%, 2.13%, and 2.11%, respectively, with larger margins at the two lower retention ratios. Its corresponding gains over OmniFocus widen from 4.47% to 7.68% and 8.72%, indicating an increasing advantage under more aggressive compression.

Efficiency Analysis. OmniRelay also achieves strong computational efficiency across all three models. As shown in Table 3, at a 10% retention ratio on Qwen2.5-Omni-7B, OmniRelay achieves a 6.3× prefill speedup while reducing FLOPs by 91.7% and GPU memory usage by 13.1% relative to the full-token baseline. Regarding the performance–latency trade-off, OmniRelay at 10% retention outperforms OmniFocus at 25% retention and SEATS at 15% retention, while requiring only 64.5% and 92.4% of their prefill time. These results demonstrate a favorable performance–efficiency trade-off across different compression levels and model architectures.

4.3 Ablation Study

We ablate SDCC and DARC on Qwen2.5-Omni-7B using AVUT, WorldSense, and DailyOmni at 25% layer-averaged token retention, with 45% pre-LLM retention. All reported differences are absolute changes in mean relative performance across the three benchmarks.

Table 4: **Ablation study on compression strategy compatibility.** Performance of different pre- and within-LLM compression combinations on Qwen2.5-Omni-7B at 25% layer-averaged token retention.

ID	Pre-LLM	Within-LLM	AVUT	WorldSense	DailyOmni	Relative Perf.	Δ
Vanilla	–	–	64.76%	46.63%	62.91%	100.00%	–
1	Random	Ours	62.75%	45.21%	58.65%	95.69%	-2.79%
2	SEATS	Ours	63.61%	45.78%	60.23%	97.38%	-1.10%
3	OmniZip	Ours	62.63%	45.56%	59.65%	96.41%	-2.07%
4	OmniFocus	Ours	63.15%	45.37%	59.31%	96.36%	-2.12%
5	Ours	Random	62.69%	45.43%	58.73%	95.86%	-2.62%
6	Ours	SEATS	62.86%	45.84%	60.32%	97.08%	-1.40%
7	Ours	FastV-om	60.44%	44.86%	58.23%	94.03%	-4.45%
OmniRelay	Ours	Ours	**64.65%**	**46.00%**	**60.99%**	**98.48%**	–

Table 5: **Ablation study on individual components.** Results for (a) SDCC and (b) DARC on Qwen2.5-Omni-7B at 25% layer-averaged token retention.

(a) SDCC

Intra		Inter	Benchmark			Rel. Perf.
Sal.	Div.	Alloc.	AVUT	WorldSense	DailyOmni	
✓			63.49	45.65	59.65	96.92
	✓		63.55	45.81	59.90	97.20
✓	✓		63.44	45.56	60.48	97.27
✓		✓	64.07	45.96	60.23	97.75
	✓	✓	63.31	45.62	59.57	96.76
✓	✓	✓	**64.65**	**46.00**	**60.99**	**98.48**

(b) DARC

Guidance	AVUT	World Sense	Daily Omni	Rel. Perf.
Intra-Modal Only	62.28	45.78	60.65	96.92
Text Only	61.36	44.77	59.57	95.15
Original	**64.65**	**46.00**	**60.99**	**98.48**

Compression Strategy Compatibility. In Table 4, Within-LLM Random follows DARC's layer-wise retention schedule but selects tokens randomly, outperforming FastV-om by 1.83% and highlighting the benefit of depth-adaptive budgeting. SEATS is the strongest replacement for either stage, yet falls below the full setup by 1.10% when replacing SDCC and 1.40% when replacing DARC.

Component Ablations. For SDCC, Table 5(a) compares alternating selection with Saliency-only and Diversity-only variants. Replacing coverage-aware allocation with a fixed retention ratio for all video windows reduces performance by 1.21%. Full SDCC outperforms every two-component variant by at least 0.73%, supporting the complementary roles of selection and allocation. For DARC, Table 5(b) shows that using only intra-modal anchor guidance or only text guidance across all pruning layers reduces performance by 1.56% and 3.33%. These results support combining both guidance sources through DARC's depth-adaptive transition.

5 Related Work

Omni-modal Large Language Models. Omni-LLMs process visual and audio inputs within a unified framework, supporting a wide range of audiovisual understanding tasks (Yang et al., 2025b; Hong et al., 2026; Zhou et al., 2025; Fu et al., 2025a; Li et al., 2026). Qwen2.5-Omni (Xu et al., 2025a) and Qwen3-Omni (Xu et al., 2025b) temporally interleave audio and visual inputs and employ TMRoPE to encode their temporal alignment. MiniCPM-o 2.6 (Yao et al., 2024) adopts time-division multiplexing to partition parallel multimodal streams into periodic time slices for sequential streaming into the LLM. However, long audiovisual inputs produce large numbers of tokens, leading to substantial memory and computational costs.

Token Compression. Previous token compression methods either merge similar tokens (Bolya et al., 2022) or prune unimportant tokens (Chen et al., 2024; Zhang et al., 2024) in Vision-LLMs. Recent studies have further extended token compression to Omni-LLMs (Yang et al., 2026; Cao et al., 2026; Su et al., 2026a; Zhong et al., 2026). For instance, OmniZip (Tao et al., 2026) uses audio cues for pre-LLM audiovisual compression, while OmniSIFT (Ding et al., 2026) combines spatio-temporal video pruning with vision-guided audio selection. SEATS (Xin et al., 2026) applies stage-adaptive compression before and within the LLM, while OmniPack (Su et al., 2026b) combines structural pre-LLM compression with query-guided refinement within the LLM.

6 Conclusions

We present OmniRelay, a training-free pre&within-LLM token compression framework for Omni-LLMs. Our analysis reveals a depth-varying information relay that motivates global intra-modal guidance in early layers and text guidance in later layers. Following this relay, DARC progressively transitions between these guidance sources before removing the remaining audiovisual tokens. SDCC complements DARC by reducing local redundancy while preserving salient and diverse evidence across windows. Experiments across multiple Omni-LLMs and audiovisual understanding benchmarks demonstrate a strong performance–efficiency trade-off. Ablations support the complementary roles of saliency–diversity selection and coverage-aware allocation, as well as the benefit of depth-adaptive guidance. Overall, OmniRelay improves efficiency of Omni-LLMs.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar, clarity and presentation of text written by the authors. We did not use these tools for research ideation or execution, drafting substantive sections of the paper, generating synthetic datasets or proving mathematical claims. We reviewed and edited all AI-assisted revisions to ensure that they preserved the intended technical meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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. *arXiv preprint arXiv:2502.13923*, 2025.

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

Shijie Cao, Qingyu Zhang, Boxi Yu, Yuzhong Zhang, Boxi Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, and Le Sun. Omnifocus: Query-guided modality-balanced token compression for omni-modal large language models. *arXiv preprint arXiv:2607.03050*, 2026.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

Yuchen Deng, Zidang Cai, Hai-Tao Zheng, Jie Wang, Feidiao Yang, and Yuxing Han. Omnirefine: Alignment-aware cooperative compression for efficient omnimodal large language models. *arXiv preprint arXiv:2605.12056*, 2026.

Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long MA, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Advances in Neural Information Processing Systems*, 38:75300–75320, 2026.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22654–22663. IEEE, 2025b.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Chao Gong, Depeng Wang, Zhipeng Wei, Ya Guo, Huijia Zhu, and Jingjing Chen. Echoing-pixels: Cross-modal adaptive token reduction for efficient audio-visual llms. *arXiv preprint arXiv:2512.10324*, 2025.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 52423–52443, 2026.

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Minji Kim, Taekyung Kim, and Bohyung Han. Map the flow: Revealing hidden pathways of information in videollms. In *International Conference on Learning Representations*, volume 2026, pp. 50839–50870, 2026.

Bingzhou Li and Tao Huang. Dash: Dynamic audio-driven semantic chunking for efficient omnimodal token compression. *arXiv preprint arXiv:2603.15685*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Zhenghao Song, Dingling Zhang, Ying He, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. In *International Conference on Learning Representations*, volume 2026, pp. 138214–138236, 2026.

Yeo Jeong Park, Hyemi Jang, Minseo Choi, Jongsun Lee, Jooyoung Choi, and Yongkweon Jeon. Omnidrop: Layer-wise token pruning for omni-modal llms via query-guidance. *arXiv preprint arXiv:2605.14458*, 2026.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Jinsen Su, Yongdong Luo, Yuexiao Ma, Yibo Hu, Meiguang Jin, and Xiawu Zheng. Omniscope: Modality-decoupled token compression for omnimodal large language models. *arXiv preprint arXiv:2607.23193*, 2026a.

Wanshun Su, Yang Shi, Feihu Liu, Ziwen Yu, Yan Min, Zhuoran Zhang, Qixun Wang, Haotian Wang, Shixuan Liu, Yuanxing Zhang, et al. Omnipack: Unified token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2608.03812*, 2026b.

Wish Suharitdamrong, Muhammad Awais, Xiatian Zhu, and Sara Atito. From senses to decisions: The information flow of auditory and visual perception in multimodal llms. *arXiv preprint arXiv:2606.10147*, 2026.

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

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18992–19001. IEEE, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17682–17692, 2026.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

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

Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pp. 453–470. Springer, 2024.

Zijie Xin, Jie Yang, Ruixiang Zhao, Tianyi Wang, Fengyun Rao, Jing Lyu, and Xirong Li. Stage-adaptive token selection for efficient omni-modal llms. *arXiv preprint arXiv:2605.20035*, 2026.

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. *arXiv preprint arXiv:2503.20215*, 2025a.

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

Morunliu Yang, Ruotao Xu, Le Li, Yue Wang, Jianxin Zhang, Juntao Li, Yihang Lou, Siwei Feng, and Peifeng Li. Omniselect: Dynamic modality-aware token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2605.18041*, 2026.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025a.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6580–6598, 2025b.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Yiming Zhong, Chang Nie, and Caifeng Shan. Omni-prune: Query-aware unified token pruning for efficient omnimodal large language models. *arXiv preprint arXiv:2607.23445*, 2026.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

APPENDIX

A BENCHMARK DETAILS

Table 6 summarizes the five benchmarks used in our evaluation. Together, they contain 4,565 videos and 9,803 question–answer pairs, covering diverse real-world scenarios and a broad range of temporal scales. These benchmarks vary substantially in their dependence on audio and visual evidence, enabling a comprehensive assessment of token compression methods across different video durations, modality dependencies, and reasoning requirements. In the following, we briefly introduce each benchmark.

Table 6: Statistics of the benchmarks used in our experiments.

Benchmark	#Videos	#QA Pairs	Duration (sec)
AVUT	691	1,734	69.1
WorldSense	1,662	3,172	140.7
DailyOmni	684	1,197	43.2
OmniVideoBench	628	1,000	384.2
VideoMME	900	2,700	1017.9

AVUT (Yang et al., 2025b). AVUT is an audio-centric video understanding benchmark comprising 2,662 videos and 11,609 question–answer pairs, designed to evaluate audio understanding and audio–visual alignment while minimizing textual shortcuts. Following the evaluation protocol of OmniZip (Tao et al., 2026), we evaluate on the human-annotated subset, AV-Human, which contains 691 videos and 1,734 question–answer pairs. Its questions cover audio information extraction, audio-event localization, and audio–visual correspondence, requiring models to effectively exploit auditory evidence and align it with relevant visual content. The average video duration of the evaluated AV-Human subset is approximately 69.1 seconds.

WorldSense (Hong et al., 2026). WorldSense comprises 1,662 synchronized audio–visual videos spanning 8 domains, with 3,172 expert-annotated multiple-choice questions across 26 cognitive tasks. Its core design principle is tight audio–visual coupling: every question requires jointly integrating visual and audio evidence, and removing either modality leads to a substantial accuracy drop. The average video duration is approximately 140.7 seconds.

DailyOmni (Zhou et al., 2025). DailyOmni comprises 684 videos collected from diverse daily-life scenarios, with 1,197 multiple-choice question–answer pairs across 6 major tasks. It focuses on temporally aligned audio–visual reasoning, covering event alignment, temporal ordering, cross-modal inference, and contextual reasoning, and therefore requires models to capture complementary information across synchronized audio and visual streams. The average video duration is approximately 43.2 seconds.

VideoMME (Fu et al., 2025a). VideoMME comprises 900 videos spanning 6 major domains, with 2,700 human-annotated multiple-choice question–answer pairs and video durations ranging from 11 seconds to 1 hour. It covers short-, medium-, and long-duration videos and supports visual, audio, and subtitle inputs, enabling evaluation across diverse temporal scales and modality configurations. In our experiments, we use audio–visual inputs without subtitles. The average video duration is approximately 1,017.9 seconds.

OmniVideoBench (Li et al., 2026). OmniVideoBench comprises 628 real-world audio–visual videos spanning 8 major categories and 68 fine-grained subcategories, with 1,000 manually verified question–answer pairs across 13 task types. Its questions are designed to require complementary evidence from both audio and visual modalities and cover diverse reasoning abilities, including temporal reasoning, spatial localization, counting, causal inference, and summarization. The average video duration is approximately 384.2 seconds.

B FLOPs Calculations

FLOPs Estimation. Following prior work (Tao et al., 2026), we evaluate the computational efficiency of token compression using the floating-point operations (FLOPs) introduced by vision and audio tokens during prefilling and decoding. We normalize the uncompressed multimodal sequence length to 10,000 tokens, with vision-to-audio token ratios of $288 : 50$ for Qwen2.5-Omni and $4096 : 750$ for MiniCPM-o-2.6. Let $n_l = n_l^{(v)} + n_l^{(a)}$ denote the number of vision and audio tokens retained at layer l, where d is the hidden dimension and m is the FFN intermediate dimension. Each Transformer layer is approximated as a multi-head attention (MHA) module followed by a feed-forward network (FFN). Under the convention that one multiply–accumulate operation is counted as one FLOP, the prefilling cost at layer l is approximated as $4n_l d^2 + 2n_l^2 d + 2n_l dm$. During autoregressive decoding, the cost of generating $\mathcal{R}$ tokens at layer l is approximated as $\mathcal{R}(4d^2 + 2dm) + 2\sum_{i=1}^{\mathcal{R}} d(n_l + i)$. We set $\mathcal{R} = 100$ in all experiments. For an LLM with T Transformer layers, the total FLOPs are:

$$\text{FLOPs} = \sum_{l=1}^{T} \left[4n_l d^2 + 2n_l^2 d + 2n_l dm + \mathcal{R}\left(4d^2 + 2dm + 2dn_l + d(\mathcal{R}+1)\right)\right]. \tag{8}$$

For methods without within-LLM token compression, $n_l = n$ for all layers, where $n = n^{(v)} + n^{(a)}$ denotes the total number of vision and audio tokens entering the LLM.

C Experimental Details

Model Configurations. We evaluate all methods on Qwen2.5-Omni-7B, Qwen2.5-Omni-3B, and MiniCPM-o-2.6 using their official pretrained checkpoints. Unless otherwise specified, all experiments focus on audio–visual understanding and disable speech generation. No additional fine-tuning is performed for the evaluated training-free token compression methods. For MiniCPM-o-2.6, our implementation does not provide an explicit mapping from visual tokens to two-dimensional spatial coordinates. We therefore disable the spatial-distance term in OmniRelay and use feature distance alone for visual token selection.

Table 7: Visual–audio retention ratios under different pre-LLM retention ratios.

Pre-LLM Retention Ratio	Qwen2.5-Omni-7B/3B	MiniCPM-o-2.6
25%	20%-55%	20%-50%
15%	10%-45%	10%-40%
10%	6%-35%	5%-35%

Token Budget Allocation. For each target pre-LLM retention ratio R, we jointly allocate visual and audio token budgets using model-specific retention ratios to ensure comparable overall multimodal token budgets across compression methods. The notation R_v–R_a denotes the retention ratios applied to visual and audio tokens, respectively. As shown in Table 7, the allocations differ slightly between Qwen2.5-Omni-7B/3B and MiniCPM-o-2.6 because the two model families produce different numbers of visual and audio tokens for the same input.

Retention Ratio Definition. Let N_v and N_a denote the numbers of visual and audio tokens before compression, and let R_v and R_a denote the corresponding modality-specific retention ratios. We define the overall multimodal retention ratio as

$$R = \frac{R_v N_v + R_a N_a}{N_v + N_a}.$$

Thus, R represents the fraction of visual and audio tokens retained after modality-specific compression, weighted by their original token counts. Text tokens are always preserved and are excluded from the calculation of R. Because the relative numbers of visual and audio tokens differ across model backbones, the same target R may correspond to different R_v–R_a allocations, as shown in Table 7.

Evaluation Setup. We use consistent input preprocessing and generation settings across all methods evaluated on the same benchmark. Table 8 summarizes the benchmark-specific evaluation settings.

Table 8: **Evaluation settings.** Input configurations for the five benchmarks. All methods use identical preprocessing and generation settings on the same benchmark.

Benchmark	Modal	Sampling FPS	Max Frames	Max Pixels
AVUT	Audio + Video	2.0	128	100,352
WorldSense	Audio + Video	2.0	128	100,352
DailyOmni	Audio + Video	2.0	128	100,352
VideoMME	Audio + Video	2.0	768	100,352
OmniVideoBench	Audio + Video	2.0	384	100,352

All evaluations use a batch size of 1, greedy decoding, and audio–visual inputs without subtitles. Videos are sampled at 2 FPS, and the maximum number of sampled frames is set according to the benchmark: 128 for AVUT, WorldSense, and DailyOmni, 768 for VideoMME, and 384 for OmniVideoBench. For visual preprocessing, both the minimum and maximum pixel budgets are set to $128 \times 28 \times 28 = 100{,}352$ pixels. The maximum output length is set to 100 tokens for all benchmarks.

C.1 Reproduction Details

For pre-LLM compression methods, the modality-specific retention ratios are applied directly before the multimodal tokens enter the LLM. For within-LLM compression methods, we report the average retention ratios across all LLM layers. Minor deviations from the nominal target budgets may arise from integer rounding, window partitioning, and dynamic token allocation. Random, FastV-om (Chen et al., 2024), VisionZip-om (Yang et al., 2025a), OmniZip (Tao et al., 2026), SEATS (Xin et al., 2026) and OmniFocus (Cao et al., 2026) all compress both visual and audio tokens. Unless otherwise specified, all baseline configurations follow their official implementations under the unified evaluation protocol described above. OmniSIFT (Ding et al., 2026) is excluded from our evaluation because it requires training, while OmniPack (Su et al., 2026b) and OmniDrop (Park et al., 2026) are excluded because their code is not publicly available.

FastV-om (Chen et al., 2024). FastV performs a single attention-based pruning step before the K-th LLM block using the last text token as the query. We use FastV-om, which independently selects visual and audio tokens. Since we report layer-averaged retention ratios, the post-pruning retention ratio is computed as $r_m = (LR_m - K)/(L - K)$. We use $K = 2$ for R25 and R15, and $K = 1$ for R10 because the target retention ratios fall below the minimum achievable value with $K = 2$.

VisionZip-om (Yang et al., 2025a). We apply VisionZip after the model's native visual pooling or token-merging stage to avoid interfering with the original architecture. Visual token selection is performed locally within two consecutive visual temporal grids. We use VisionZip-om, which applies the same Top-k strategy to audio using 50-token windows. For MiniCPM-o-2.6, each frame's Resampler outputs are treated as a one-dimensional token sequence, and two consecutive frames form one visual window.

OmniZip (Tao et al., 2026). OmniZip first compresses audio tokens and then adjusts the visual budget according to audio information density, followed by alternating DPC-KNN and temporal-difference-based visual selection. We use the default contextual ratio of 0.05 and $G = 3$. OmniZip is evaluated only at R25 because its video-group discard ratio is capped at 75%, preventing it from satisfying the lower visual budgets required at R15 and R10.

SEATS (Xin et al., 2026). SEATS performs pre-LLM window-wise audiovisual compression followed by progressive text-guided pruning at predefined LLM layers and removes all remaining audiovisual tokens at a late layer. We report its modality-specific retention ratio averaged across all

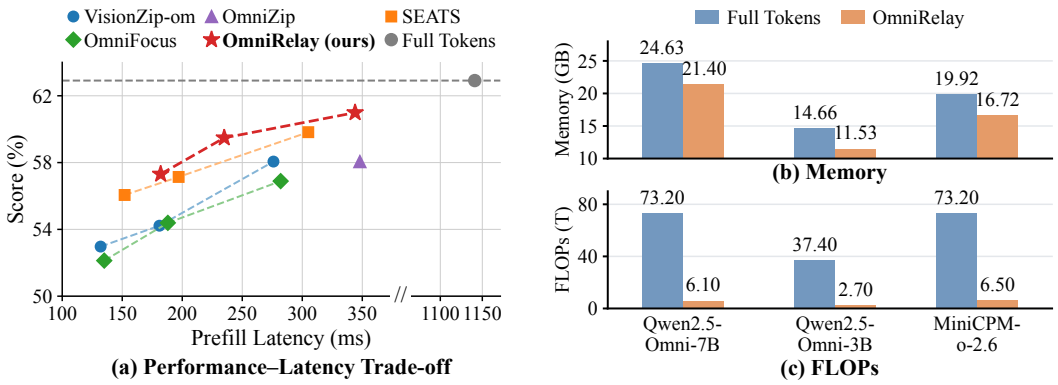


Figure 5: **Efficiency on DailyOmni.** (a) Performance versus prefill latency on Qwen2.5-Omni-7B. (b–c) Peak GPU memory and theoretical FLOPs across three backbones, comparing the full-token baseline with OmniRelay at 10% retention.

LLM layers. We evaluate SEATS on Qwen2.5-Omni-7B and MiniCPM-o-2.6 using its specified 28-layer pruning schedule. We exclude Qwen2.5-Omni-3B because its 36-layer architecture is not supported by the SEATS implementation used in our evaluation.

OmniFocus (Cao et al., 2026). OmniFocus dynamically assigns chunk-level audiovisual budgets according to text-query relevance and selects tokens using cross-modal association and intra-modal peakness. In our settings, the target visual discard ratios exceed the implementation's 75% dynamic-allocation range, causing all visual chunks to use the same global discard ratio, while query-guided dynamic allocation remains active on the audio branch.

C.2 Additional Efficiency Results

Table 9 extends the main-text results with a detailed comparison on DailyOmni across Qwen2.5-Omni-7B, Qwen2.5-Omni-3B, and MiniCPM-o-2.6 at 25%, 15%, and 10% token retention. We report theoretical FLOPs, prefill latency, prefill speedup, peak GPU memory, and accuracy. Speedups are calculated relative to each backbone's full-token baseline.

Figure 5(a) illustrates the performance–latency trade-off on Qwen2.5-Omni-7B. At 10% retention, OmniRelay outperforms OmniFocus at 25% retention and SEATS at 15% retention while requiring less prefill time. Across the overlapping latency range shown in the figure, OmniRelay achieves higher accuracy than all other compared compression methods at comparable prefill latency. Figure 5(b–c) further compares computational and memory costs across the three backbones. At 10% retention, OmniRelay reduces theoretical FLOPs by approximately 91.7%, 92.8%, and 91.1%, and peak GPU memory by 13.1%, 21.3%, and 16.1%, respectively, relative to the corresponding full-token baselines. These results demonstrate consistent efficiency gains across different model architectures.

C.3 Detailed Performance on WorldSense and AVUT

Tables 10 and 11 report accuracy at 25% token retention across Qwen2.5-Omni-7B, Qwen2.5-Omni-3B, and MiniCPM-o-2.6, following each benchmark's category definitions: eight video domains for WorldSense and six sub-tasks for AVUT. OmniRelay shows particular strength in preserving fine-grained matching capabilities, outperforming the strongest competing compression results by 4.08% in Object Matching on Qwen2.5-Omni-7B and 3.53% in OCR Matching on Qwen2.5-Omni-3B. On Qwen2.5-Omni-7B, it ranks first in five WorldSense domains and second in the remaining three. On AVUT, it ranks first in four sub-tasks and second in one. Its overall accuracy on this backbone exceeds the strongest competing method, SEATS, by 0.70% on WorldSense and 1.56% on AVUT, preserving 98.65% and 99.83% of the respective full-token baseline accuracy. Overall, OmniRelay achieves the highest overall accuracy among the compared compression methods on both benchmarks across all three backbones.

Table 9: **Detailed efficiency analysis.** Efficiency and accuracy comparison on DailyOmni across Qwen2.5-Omni-7B, Qwen2.5-Omni-3B, and MiniCPM-o-2.6. Each cell reports results for these three models in this order, separated by slashes. Speedup is calculated relative to the corresponding full-token prefill latency. A dash (-) denotes unavailable results.

Method	FLOPs (T) ↓	Prefill Time (s) ↓	Speedup ↑	Peak GPU Memory (GB) ↓	Score (%) ↑
Full Tokens	73.2 / 37.4 / 73.2	1.141 / 0.575 / 1.165	1.0× / 1.0× / 1.0×	24.627 / 14.661 / 19.916	62.91 / 61.24 / 60.40
Retention Ratio: 25%					
Random	15.0 / 6.8 / 14.7	0.281 / 0.144 / 0.395	4.1× / 4.0× / 2.9×	22.041 / 12.236 / 17.517	56.22 / 53.97 / 56.73
FastV-om	15.9 / 7.3 / 15.6	0.293 / 0.154 / 0.407	3.9× / 3.7× / 2.9×	22.424 / 12.119 / 17.573	56.31 / 55.05 / 54.55
VisionZip-om	15.0 / 6.8 / 14.7	0.276 / 0.141 / 0.396	4.1× / 4.1× / 2.9×	22.030 / 12.225 / 17.453	58.06 / 56.31 / 54.22
OmniZip	14.9 / 6.8 / 14.7	0.348 / 0.176 / 0.433	3.3× / 3.3× / 2.7×	22.236 / 12.420 / 17.611	58.06 / 57.39 / 58.23
SEATS	15.5 / - / 15.1	0.305 / - / 0.429	3.7× / - / 2.7×	21.691 / - / 16.900	59.82 / - / 58.98
OmniFocus	15.0 / 6.8 / 14.9	0.282 / 0.145 / 0.400	4.0× / 4.0× / 2.9×	22.041 / 12.236 / 17.474	56.89 / 55.97 / 54.47
OmniRelay	15.5 / 7.3 / 16.4	0.344 / 0.215 / 0.478	3.3× / 2.7× / 2.4×	21.780 / 11.755 / 16.958	60.99 / 58.90 / 59.15
Retention Ratio: 15%					
Random	9.0 / 4.0 / 8.7	0.187 / 0.095 / 0.298	6.1× / 6.1× / 3.9×	21.687 / 11.905 / 17.198	53.05 / 52.05 / 55.39
FastV-om	10.1 / 4.6 / 9.8	0.198 / 0.101 / 0.308	5.8× / 5.7× / 3.8×	22.424 / 12.119 / 17.573	53.30 / 50.63 / 49.87
VisionZip-om	9.0 / 4.0 / 8.7	0.181 / 0.090 / 0.301	6.3× / 6.4× / 3.9×	21.674 / 11.894 / 17.129	54.22 / 53.63 / 49.87
SEATS	9.1 / - / 8.8	0.197 / - / 0.333	5.8× / - / 3.5×	21.461 / - / 16.700	57.14 / - / 56.89
OmniFocus	9.0 / 4.0 / 8.9	0.188 / 0.094 / 0.297	6.1× / 6.1× / 3.9×	21.687 / 11.905 / 17.154	54.39 / 53.38 / 48.37
OmniRelay	9.2 / 4.2 / 9.6	0.235 / 0.164 / 0.370	4.9× / 3.5× / 3.1×	21.550 / 11.620 / 16.739	59.48 / 57.39 / 57.31
Retention Ratio: 10%					
Random	6.1 / 2.7 / 5.8	0.135 / 0.070 / 0.256	8.5× / 8.2× / 4.6×	21.513 / 11.744 / 17.040	50.96 / 48.20 / 51.13
FastV-om	6.7 / 3.0 / 6.4	0.147 / 0.074 / 0.260	7.8× / 7.8× / 4.5×	22.335 / 12.069 / 17.481	47.20 / 41.44 / 46.12
VisionZip-om	6.1 / 2.7 / 5.8	0.132 / 0.066 / 0.250	8.6× / 8.7× / 4.7×	21.506 / 11.737 / 16.975	52.97 / 53.05 / 47.28
SEATS	6.2 / - / 5.8	0.152 / - / 0.283	7.5× / - / 4.1×	21.342 / - / 16.694	56.06 / - / 52.63
OmniFocus	6.1 / 2.7 / 6.0	0.135 / 0.070 / 0.262	8.5× / 8.2× / 4.4×	21.514 / 11.744 / 16.995	52.13 / 51.38 / 43.02
OmniRelay	6.1 / 2.7 / 6.5	0.182 / 0.139 / 0.314	6.3× / 4.1× / 3.7×	21.402 / 11.531 / 16.716	57.31 / 56.73 / 54.05

Table 10: **Detailed performance on WorldSense.** Accuracy (%) across video domains at 25% token retention. Gray model-name rows report full-token baselines. Overall denotes question-weighted accuracy across the full benchmark. The best and second-best compressed results within each model are bold and underlined, respectively; ties share the same rank.

Method	Daily Life	Music	Culture & Politics	Sports	Film & TV	Games	Performance	Tech & Science	Overall
Qwen2.5-Omni-7B	48.33	47.04	50.81	41.86	44.06	41.63	43.07	51.84	46.63
Random	43.77	46.31	48.22	38.84	41.42	37.77	40.82	47.96	43.54
FastV-om	42.10	44.58	47.90	39.30	**43.27**	38.63	38.20	47.35	42.97
VisionZip-om	44.83	46.31	48.54	40.23	40.11	38.63	41.20	49.18	44.10
OmniZip	46.35	45.57	49.51	40.23	41.95	38.63	**41.95**	49.39	44.74
SEATS	**46.66**	46.80	50.49	**42.79**	42.22	39.48	39.70	49.39	45.30
OmniFocus	45.59	46.06	47.90	38.14	39.84	38.20	41.20	46.53	43.41
OmniRelay	**46.66**	**47.54**	**52.75**	40.93	43.01	**42.06**	41.57	**50.61**	**46.00**
Qwen2.5-Omni-3B	44.38	45.81	52.10	44.65	45.12	43.78	42.32	52.24	46.44
Random	41.95	44.33	44.66	40.00	40.37	38.20	39.70	48.16	42.56
FastV-om	42.55	43.10	47.25	39.53	43.27	38.20	40.45	50.20	43.44
VisionZip-om	**44.07**	**45.32**	49.19	41.40	**44.33**	**42.06**	40.82	50.41	44.96
OmniZip	**44.07**	43.84	**51.13**	42.79	**44.33**	40.34	41.57	**52.65**	45.43
OmniFocus	43.01	45.07	45.95	40.47	41.16	**42.06**	40.07	47.55	43.38
OmniRelay	43.77	44.58	50.81	**43.26**	43.80	40.34	**44.19**	51.63	**45.49**
MiniCPM-o-2.6	39.67	46.06	45.95	40.70	44.59	44.21	47.19	45.10	43.63
Random	36.17	45.32	44.34	39.77	43.01	40.77	43.82	41.63	41.27
FastV-om	36.47	44.09	41.10	36.28	41.95	42.06	41.95	41.02	40.10
VisionZip-om	37.39	44.09	41.10	35.81	39.84	40.34	40.45	41.43	39.79
OmniZip	37.39	46.55	40.45	39.30	43.27	41.63	44.94	41.63	41.42
SEATS	38.30	47.29	44.01	40.23	**45.91**	**45.92**	**46.07**	42.45	43.03
OmniFocus	37.54	43.10	39.81	36.28	41.16	38.20	39.70	37.14	38.90
OmniRelay	**39.21**	**47.54**	**44.98**	**40.93**	45.12	42.06	45.69	**42.65**	**43.06**

D CASE STUDY

Figure 6 compares OmniRelay and SEATS on a question requiring visual identification at a speech-specified moment. The audio provides the temporal cue, while the corresponding visual content determines the answer, requiring evidence from both modalities. In the pre-LLM stage, SDCC groups similar consecutive video windows and assigns larger budgets to medoid windows. This allocation reduces redundancy among similar adjacent windows while preserving representative visual evidence, rather than assigning the same retention ratio to every window. Within the LLM, at the same overall retention ratio, OmniRelay retains up to 33 tokens per window in the question-

Table 11: **Detailed performance on AVUT.** Accuracy (%) across sub-tasks at 25% token retention. Gray model-name rows report full-token baselines. Overall denotes question-weighted accuracy across the full benchmark. The best and second-best compressed results within each model are bold and underlined, respectively; ties share the same rank.

Method	Character Matching	Content Counting	Event Location	Information Extraction	OCR Matching	Object Matching	Overall
Qwen2.5-Omni-7B	68.11	42.37	35.88	86.20	68.81	59.44	64.76
Random	62.83	39.83	38.24	83.13	61.41	52.55	60.09
FastV-om	54.68	37.29	**40.00**	82.82	66.88	50.77	58.65
VisionZip-om	59.47	36.44	37.06	82.82	61.41	52.55	58.88
OmniZip	63.79	38.98	37.06	83.74	68.81	53.06	61.71
SEATS	64.51	38.98	39.41	**86.20**	69.13	55.10	63.09
OmniFocus	55.16	38.98	37.65	84.97	66.88	53.83	59.75
OmniRelay	**66.19**	**42.37**	38.82	85.58	**70.10**	**59.18**	**64.65**
Qwen2.5-Omni-3B	61.63	45.76	30.59	84.97	65.92	58.16	61.88
Random	58.03	42.37	27.06	82.52	57.56	52.81	57.27
FastV-om	45.56	**44.07**	**32.35**	83.44	58.20	48.47	54.21
VisionZip-om	58.75	39.83	25.29	84.36	60.77	51.02	57.61
OmniZip	**61.15**	43.22	28.82	84.05	61.74	54.08	59.57
OmniFocus	55.16	40.68	**32.35**	**84.97**	63.67	50.51	58.02
OmniRelay	59.95	42.37	29.41	**84.97**	**67.20**	**54.85**	**60.61**
MiniCPM-o-2.6	70.26	42.37	27.65	70.25	68.17	64.03	62.40
Random	62.59	40.68	28.82	65.64	60.77	54.34	56.17
FastV-om	60.67	38.98	**31.76**	59.20	53.70	49.49	52.31
VisionZip-om	52.76	41.53	26.47	65.03	53.38	51.79	51.61
OmniZip	64.27	42.37	31.18	66.26	**64.63**	57.91	58.54
SEATS	**65.47**	39.83	30.59	69.94	63.02	**59.69**	59.40
OmniFocus	50.84	43.22	27.06	59.82	55.63	51.28	50.63
OmniRelay	65.23	**44.92**	28.82	**70.86**	63.02	59.18	**59.57**

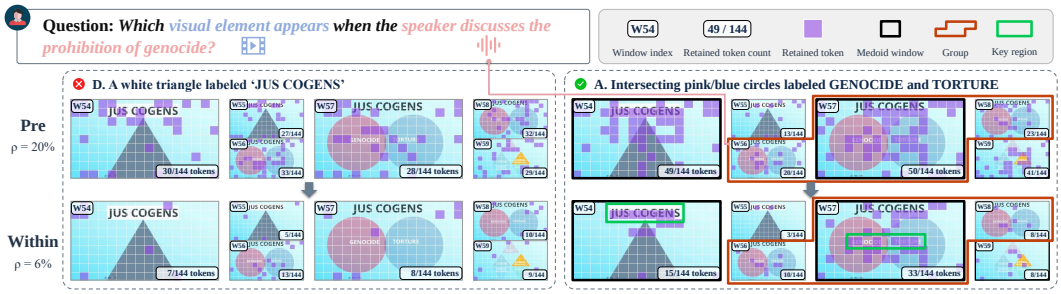


Figure 6: **Case study.** Token retention on Qwen2.5-Omni-7B after pre-LLM compression (top) and LLM layer 20 (bottom). Green boxes mark key visual evidence; orange outlines indicate OmniRelay's groups of similar consecutive windows. Display sizes are for visualization only.

relevant windows W56–W58, compared with a maximum of 13 for SEATS. Beyond the difference in token counts, OmniRelay fully preserves the marked key region containing "GENOCIDE" and "TORTURE," retaining the visual details needed to distinguish the answer options. OmniRelay correctly identifies the intersecting circles bearing these terms, whereas SEATS incorrectly selects the white triangle. This case illustrates how representative content preservation before the LLM and task-relevant selection within the LLM can jointly support audiovisual reasoning under compression.

E Parameter Ablation

We conduct parameter ablations on the feature–position distance weights λ_v and λ_a in pre-LLM SDCC, and the Gaussian centers (μ_1, μ_2) and width σ controlling the depth-dependent anchor-guided and text-guided importance profiles in within-LLM DARC. Experiments are conducted on Qwen2.5-Omni-7B using AVUT, WorldSense, and DailyOmni at a layer-averaged token retention ratio of $R_{\text{mean}} = 25\%$. We vary one parameter setting at a time while keeping the others fixed.

Table 12: **Parameter ablation.** Results on Qwen2.5-Omni-7B across AVUT, WorldSense, and DailyOmni at 25% layer-averaged token retention.

Parameter	Value	AVUT	WorldSense	DailyOmni	Relative Perf.	Δ
λv	0.10	63.38	46.19	60.23	97.56%	-0.92%
	0.15	63.26	45.87	60.15	97.22%	-1.26%
	0.20	**64.65**	46.00	**60.99**	**98.48%**	0.00%
	0.25	63.67	**46.63**	60.65	98.24%	-0.24%
	0.30	63.78	46.25	60.57	97.98%	-0.50%
λa	0.10	64.01	46.03	60.15	97.72%	-0.76%
	0.15	64.30	46.09	60.40	98.05%	-0.43%
	0.20	**64.65**	46.00	**60.99**	**98.48%**	0.00%
	0.25	64.01	46.15	60.15	97.81%	-0.67%
	0.30	64.13	**46.28**	60.40	98.10%	-0.38%
(μ_1, μ_2)	$(8, 13)$	64.30	46.03	60.57	98.09%	-0.39%
	$(9, 14)$	63.73	46.12	60.48	97.82%	-0.66%
	(10,15)	**64.65**	46.00	**60.99**	**98.48%**	0.00%
	$(11, 16)$	64.07	46.19	60.82	98.22%	-0.26%
	$(12, 17)$	63.78	**46.37**	60.40	97.98%	-0.50%
σ	3	64.42	46.15	60.15	98.02%	-0.46%
	4	64.13	45.90	60.65	97.96%	-0.52%
	5	**64.65**	46.00	**60.99**	**98.48%**	0.00%
	6	64.07	46.31	60.65	98.22%	-0.26%
	7	64.13	**46.41**	60.07	98.01%	-0.47%

Feature–Position Distance Weights. As shown in Table 12, both modalities achieve their best overall performance at $\lambda_v = \lambda_a = 0.20$, despite using different positional structures: two-dimensional spatial distance for video and one-dimensional temporal distance for audio. This suggests that the same mixing weight provides an effective balance between feature dissimilarity and positional separation for both modalities under the evaluated setting. Averaged over the four non-default values, relative performance decreases by 0.73 and 0.56 percentage points for λ_v and λ_a, respectively, indicating greater sensitivity to the video distance weight.

Depth-Dependent Guidance Parameters. Varying (μ_1, μ_2) and σ yields average relative performance drops of only 0.45 and 0.43 percentage points, respectively, across the four non-default settings. Compared with the distance weights, these parameters have a smaller effect on overall performance, indicating that DARC remains stable across the tested guidance locations and widths.

Default Configuration. The default configuration achieves the best results on AVUT and DailyOmni among the tested settings, while WorldSense benefits from some alternative values. For example, $\lambda_v = 0.25$ improves WorldSense from 46.00 to 46.63, but reduces performance on the other two benchmarks. We therefore adopt a shared configuration of $\lambda_v = \lambda_a = 0.20$, $(\mu_1, \mu_2) = (10, 15)$, and $\sigma = 5$, rather than selecting separate parameter values for each benchmark. Overall, these ablations show that OmniRelay maintains stable performance across the tested parameter ranges.

F Additional Related Work

F.1 Video Large Language Models

Video large language models aim to model temporally structured visual information within the limited context budgets of LLMs. Early approaches primarily relied on cross-modal connectors to project sampled video frames into continuous representations compatible with the language model. Video-LLaMA (Zhang et al., 2023) introduces a Video Q-Former to aggregate temporal information across frames and align video representations with the LLM embedding space. LLaVA-OneVision (Li et al., 2024) provides a unified visual modeling framework for handling images, multiple images, and videos within a single model. As video duration increases, subsequent work has increasingly focused on selective processing and long-context video understanding. Q-Frame (Zhang et al., 2025) performs query-aware frame selection to retain task-relevant temporal evidence, while

LongVLM (Weng et al., 2024) hierarchically models local and global temporal information for long-video understanding. Qwen2-VL (Wang et al., 2024) and Qwen2.5-VL (Bai et al., 2025) employ dynamic visual representations and temporally aware positional encoding to accommodate videos of varying resolutions and durations. Overall, video LLMs have evolved from uniformly processing sampled frames to selectively retaining task-relevant information over longer temporal contexts, motivating token-level redundancy reduction.

F.2 Omni-modal Large Language Models

Omni-modal large language models extend this setting to jointly modeling text, visual, and audio streams with different sampling rates and temporal resolutions. In addition to modeling temporal dependencies within each modality, these models must also preserve temporal correspondence between synchronized audio and visual streams. Existing approaches typically achieve this through multimodal sequence construction and cross-modal interaction. Qwen2.5-Omni (Xu et al., 2025a) and Qwen3-Omni (Xu et al., 2025b) temporally interleave audio and visual representations and use temporally aligned positional encodings to model their cross-modal correspondence. MiniCPM-o 2.6 (Yao et al., 2024) organizes continuous multimodal streams through time-division multiplexing, while video-SALMONN (Sun et al., 2024) employs dedicated connectors to integrate speech, general audio, and visual representations. As audiovisual inputs become longer, LLMs receive increasingly long sequences of synchronized yet redundant modality tokens, making efficient modeling of both intra-modal dependencies and cross-modal complementarity increasingly important.

F.3 Token Compression

To alleviate the growing redundancy and computational burden of long multimodal sequences, token compression reduces sequence length by removing or merging redundant tokens. Visual token compression methods leverage representation similarity (Bolya et al., 2022), attention-based importance (Chen et al., 2024), and text relevance (Zhang et al., 2024) to guide merging or pruning. Video token compression further targets both spatial and temporal redundancy in video sequences. PruneVid (Huang et al., 2025) jointly prunes spatially and temporally redundant representations, DyCoke (Tao et al., 2025) dynamically compresses repetitive tokens across frames, and FrameFusion (Fu et al., 2025b) combines similarity-based merging with importance-based pruning. LongVU (Shen et al., 2024) leverages inter-frame dependencies and textual queries to adaptively compress long videos. With the emergence of Omni-LLMs, token compression has expanded from visual inputs to synchronized audiovisual streams through cross-modal guidance. At the pre-LLM stage, OmniZip (Tao et al., 2026) leverages audio saliency and temporal correspondence to guide visual token compression, whereas OmniSIFT (Ding et al., 2026) uses retained visual representations to identify informative audio tokens. OmniDrop (Park et al., 2026) performs text-guided layer-wise audiovisual pruning, SEATS (Xin et al., 2026) dynamically adjusts the compression strategy across different processing stages, and OmniPack (Su et al., 2026b) combines structured pre-LLM compression with query-guided refinement within the LLM.

G Limitations

Adapting OmniRelay to different backbones requires configuring heuristic hyperparameters. One example is the layer at which all remaining audiovisual tokens are removed. The spatial positions needed for distance calculations must also be extracted according to each backbone's token organization. Furthermore, window-based grouping and anchor guidance may underrepresent brief but relevant changes within otherwise similar consecutive windows. Future work could explore event-aware grouping and more adaptive anchor selection.