# AVOC: Enhancing Hour-Level Audio-Video Understanding in Omni-Modal LLMs via Retrieval-Inspired Token Compression

**Yijing Chen**[1] **Wenhui Tan**[1] **Xiaoyi Yu**[1] **Yuyue Wang**[1] **Xin Cheng**[1] **Kaisi Guan**[1]
**Hao Jiang**[2] **Xiangyang Li**[2] **Guojie Zhu**[2] **Ruihua Song**[1*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China [2]Huawei Inc.

## Abstract

Multimodal Large Language Models have achieved remarkable progress in short-form audio-video understanding, yet long-form audio-video comprehension remains challenged by limited context windows and severe information redundancy. To address these bottlenecks, we propose AVOC, a framework for long-form audio-video understanding in Omni-modal Large Language Models. AVOC introduces a learnable token compression module between the modality encoders and the LLM backbone. We reframe multimodal token compression as a top-$K$ retrieval problem: given a fixed context budget, the module must retrieve a compact subset of tokens that best supports answering the user query. We draw inspiration from three classical Information Retrieval criteria for selecting informative units from a large candidate pool: *relevance*, *importance*, and *diversity*. AVOC instantiates each criterion as a tailored mechanism for audio-video understanding, and integrates them into a unified retrieval-style compression pipeline. Experiments show that AVOC achieves state-of-the-art performance on long-form audio-video benchmarks, surpassing the second-best model by 4.9 and 5.5 points in average accuracy on OmniVideoBench and LVOmniBench, respectively. Moreover, AVOC maintains robust performance on Audio-Video Needle-in-a-Haystack task at durations up to one hour. Code and model are at github.com/YJCX330/AVOC.

## 1 Introduction

Multimodal Large Language Models (MLLMs) [41, 38, 8, 35, 10] have made remarkable progress in bridging vision, audio, and natural language. By integrating visual and audio encoders with large language models, existing methods perform well on short-form audio-video tasks such as audio-video question answering, video and audio captioning, and multimodal dialogue [3, 48, 23, 19]. However, real-world multimodal information (e.g., movies, meeting recordings, and tutorials) typically spans extremely long durations. This requires models not only to comprehend short-form events, but also to reason over and localize key information within hour-level audio-video contexts.

Despite the strong demand, endowing models with hour-level audio-video understanding capabilities still faces severe challenges. On the one hand, the limited context window of MLLMs cannot directly accommodate the massive token sequences produced by extremely long audio-video streams. On the other hand, raw audio-video streams exhibit substantial information redundancy, which not only wastes the precious context budget but also dilutes critical cues, degrading the model's understanding quality over long sequences. As illustrated in Figure 1, existing context-reduction strategies fall short on extremely long audio-video content. Content-agnostic sampling faces a fundamental trade-off: sparse sampling misses critical short-lived events, while dense sampling

*Corresponding author.

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

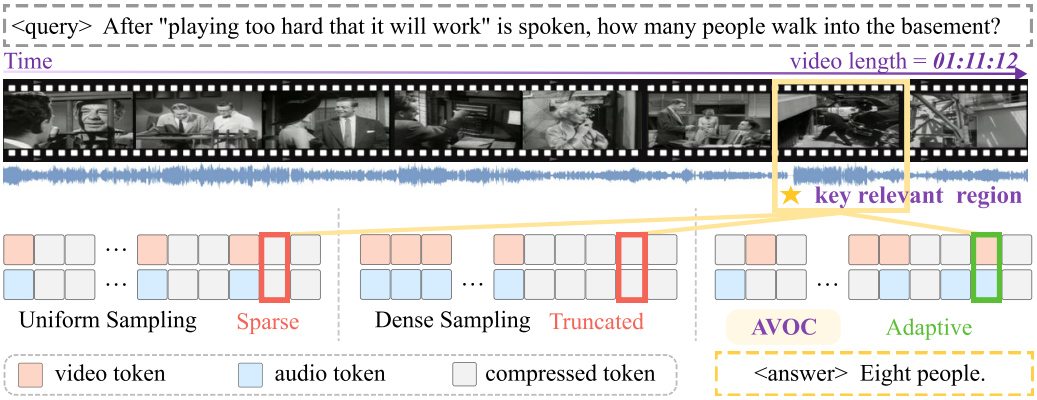


Figure 1: Comparison of context-reduction strategies for long-form audio-video understanding.

rapidly exhausts the context window, leading to severe sequence truncation. Recent omni-modal compression methods [36, 11] address this gap, but they usually adopt rigid asymmetric designs, where one modality drives compression of the other. As a result, important events may be discarded when the guiding modality provides weak or sparse signals.

To address the above problems, we propose a new framework called AVOC (Enhancing Hour-Level **A**udio-**V**ideo Understanding in **O**mni-Modal LLMs via Retrieval-Inspired Token **C**ompression). Our starting point is to reframe multimodal token compression as a *top-$K$ retrieval problem*: given a fixed context budget and a large pool of candidate tokens, the model must retrieve a compact subset which best supports answering the user query. This reformulation allows us to leverage classical Information Retrieval (IR) principles for selecting informative units under limited capacity budgets. Among the criteria that IR has long developed for ranking and selecting informative units, three are particularly relevant to our setting: query-conditioned *relevance*, which prioritizes units pertinent to the user query [28, 16]; query-agnostic *importance*, which captures intrinsic informativeness independent of any specific query [27]; and result *diversity*, which penalizes redundancy among the selected units [2, 9]. AVOC adapts these IR principles to long audio-video understanding via a learnable compression module that realizes each criterion with a tailored mechanism. *Relevance* is computed via text-guided cross-attention that conditions per-token scores on the user query. *Importance* is computed via bidirectional video-audio cross-attention within each temporal block, providing a query-agnostic signal that complements relevance when the textual query is sparse. *Diversity* is enforced through Temporal-Aware Maximal Marginal Relevance, which penalizes similarities within a local temporal window, suppressing redundant adjacent tokens while preserving recurring events that are temporally distant. Together, these three mechanisms yield a compact, informative token sequence under a tight context budget.

The main contributions of this paper can be summarized as follows:

- From a new perspective of multimodal token compression as a top-$K$ retrieval problem over multimodal tokens, we design a learnable compression module that instantiates three classical IR criteria with tailored mechanisms: text-guided cross-attention for query-conditioned *relevance*, bidirectional video-audio cross-attention within each temporal block for query-agnostic *importance*, and Temporal-Aware Maximal Marginal Relevance Selecting for local *diversity*.
- Built upon this compression module, we develop AVOC, an omni-modal large language model capable of processing hour-level audio-video streams, achieving both holistic comprehension and fine-grained retrieval over ultra-long multimodal content under a tight context budget.
- Extensive experiments demonstrate that AVOC achieves state-of-the-art performance on multiple long-form audio-video understanding benchmarks, surpassing the second-best method by 4.9 and 5.5 points in average accuracy on OmniVideoBench and LVOmniBench, respectively, and maintains robust accuracy on Audio-Video Needle-in-a-Haystack task at durations up to one hour.

## 2 Related Work

**Long-Form Video Understanding in Vision Large Language Models.** Recent years have witnessed significant progress in extending Vision-Language Models (VLMs) to long-form video understanding [33, 45, 7, 34, 32, 20]. One line of research focuses on context window extension to ingest full token sequences [24, 45, 7, 6, 39], though this kind of approach is computationally prohibitive at long sequence lengths and fails to address the heavy information redundancy in video data. To reduce computational cost and redundancy, numerous compression-based methods have emerged [30, 17]. These methods generally fall into four underlying mechanisms [30]: transformation-based approaches that employ spatial or temporal pooling [26, 40]; similarity-based techniques that group and merge redundant tokens across consecutive frames [15, 20, 31, 33]; attention-based methods that prune tokens based on attention sparsity [4, 42, 44, 46]; and query-based strategies that utilize token distillation via dynamic memory banks or cross-modal token selection [33, 32, 31, 22]. Despite these advancements, current methodologies largely ignore the accompanying audio stream. In real-world multimodal content, such as movies, tutorials, and meetings, auditory signals carry irreplaceable semantic context. By remaining strictly reliant on compressed visual cues, existing long-video VLMs inevitably suffer from incomplete semantic comprehension, overlooking critical auditory information such as speech, environmental sounds, and music that is essential for holistic understanding.

**Unified Audio-Video Understanding in Omni-Modal Large Language Models.** To overcome the visual-centric limitations of VLMs, some recent research has shifted toward the development of Omni-Modal Large Language Models (OLLMs) capable of unified audio-video understanding. To compress the massive information generated by high-resolution video and continuous high-sampling-rate audio into a limited context window, initial OLLMs predominantly relied on content-agnostic operations such as sparse temporal subsampling, basic average pooling, or naive sequence truncation [41, 25, 43]. Lacking content awareness and offering limited compression ratios, these methods fail to enable models to comprehend extremely long audio-video content. To address these bottlenecks, recent studies have introduced dynamic token compression strategies to optimize context window utilization. OmniZip [36] utilizes salient audio tokens to capture information density and guide the pruning rate of corresponding video tokens. Conversely, OmniSIFT [11] indicates that human perception is visually anchored; it first prunes spatio-temporal video redundancy and then utilizes the resulting visual anchors to select informative audio tokens. Both OmniZip [36] and OmniSIFT [11] rely on a unidirectional dependency in which either video or audio serves as the dominant modality that drives the compression of the other. This risks destruction of key information when the dominant modality experiences sparsity. These gaps highlight the necessity for a symmetric and adaptive compression architecture that better models the relations across modalities and maximizes information density within the context window without restrictive asymmetric biases.

## 3 Methodology

To enable hour-level audio-video understanding in OLLMs, we introduce a dynamic compression module that jointly condenses continuous visual and auditory streams into a compact sequence of highly informative representations. As illustrated in Figure 2, this module is strategically positioned between the modality encoding stage and the large language model backbone.

### 3.1 Problem Formulation and a New Perspective

Following standard practice in OLLMs [41, 38, 10, 43], the video and audio streams are first encoded separately and grouped into temporal blocks with equal time duration, where each block concatenates the video and audio tokens from the time interval it covers. The blocks are then arranged sequentially to form a unified multimodal token sequence. Let $X = \{x_1, x_2, \ldots, x_N\}$ denote the full sequence of interleaved multimodal tokens, where each token $x_i$ carries a temporal block index $\tau_i$ and a modality label $m_i \in \{V, A\}$. Given a text query $T$ and a fixed token budget $K < N$, the goal of our compression module is to select a compact subset $S \subset X$ of size $K$ that best preserves the information needed for downstream reasoning.

This problem can be cast as a top-$K$ retrieval problem over multimodal tokens: each token plays the role of a candidate unit in a retrieval corpus, the text query $T$ serves as the search query, and the budget constraint $|S| = K$ corresponds to retrieving the top-$K$ units that best support answering the

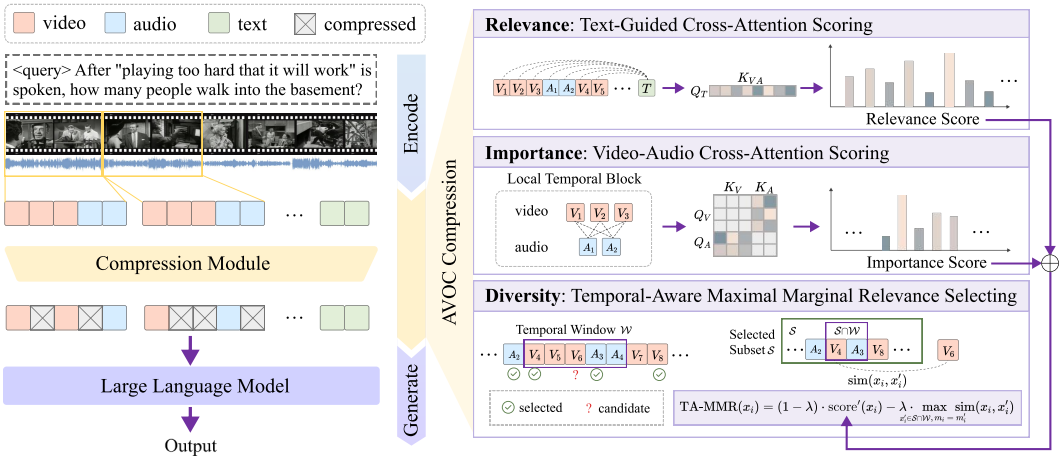


Figure 2: Overview of AVOC. The compression module condenses the interleaved video-audio token sequence into a compact subset before passing it to the LLM, guided by three retrieval-inspired criteria: relevance, importance and diversity.

query. This IR perspective lets us inherit several design principles that information retrieval has long developed for selecting informative units from a large candidate pool under capacity constraints. In particular, we adopt three well-established criteria from IR: query-conditioned *relevance* that prioritizes units pertinent to the user query, forming the basis of ranking from classical lexical matching to learned neural rankers [28, 16]; query-agnostic *importance* that captures intrinsic informativeness independent of any specific query, as exemplified by graph-centrality-style scores [27]; and result *diversity* that penalizes redundancy among the selected units, as exemplified by Maximal Marginal Relevance and related re-ranking schemes [2, 9].

Adapting these IR principles to the long audio-video setting, we identify three criteria that an informative retrieved token subset should satisfy: (i) *relevance*—retrieved tokens should carry information pertinent to the user query; (ii) *importance*—retrieved tokens should additionally reflect query-agnostic informativeness, complementing relevance when the textual query is sparse relative to the rich audio-video content; and (iii) *diversity*—retrieved tokens should contribute minimally overlapping information, so that each occupies the limited context budget with a distinct contribution.

Guided by these three criteria, we design a learnable, retrieval-style compression module that instantiates each axis with a tailored mechanism, detailed in the following subsections: text-guided cross-attention scoring for **relevance** (3.2), bidirectional video-audio cross-attention scoring for **importance** (3.3), and Temporal-Aware Maximal Marginal Relevance selection for **diversity** (3.4). The first two stages produce per-token scores that play the role of a learned retrieval scorer, and the third stage performs a temporally aware diversity re-ranking over these scores, mirroring the common scoring-and-reranking pattern in IR.

### 3.2 Relevance: Text-Guided Cross-Attention Scoring

In the spirit of query-document scoring in IR, the text-guided cross-attention module treats the text query as the search query and the multimodal tokens as the candidate corpus, and computes a per-token relevance score that conditions token selection on the user query.

Let $E_{\text{va}} \in \mathbb{R}^{N\times d}$ denote the embedding matrix of all multimodal tokens in $X$, and let $E_{\text{text}} \in \mathbb{R}^{N_{\text{text}}\times d}$ denote the embeddings of the text query $T$. We project them into query and key spaces using two learnable projection matrices $W_q^{\text{rel}}, W_k^{\text{rel}} \in \mathbb{R}^{d\times d}$:

$$Q_{\text{text}} = E_{\text{text}} W_q^{\text{rel}}, \quad K_{\text{va}} = E_{\text{va}} W_k^{\text{rel}}. \tag{1}$$

We compute the cross-attention scores between the text queries and the multimodal tokens using scaled dot-product attention:

$$A^{\text{rel}} = \frac{Q_{\text{text}} \cdot K_{\text{va}}^T}{\sqrt{d}}, \tag{2}$$

where $d$ is the hidden dimension size. To determine the overall relevance of each multimodal token $x_i$, we average the attention logits received from all textual tokens $j$:

$$\mathrm{score}_{\text{rel}}(x_i) = \frac{1}{N_{\text{text}}} \sum_j A^{\text{rel}}_{j,i}. \tag{3}$$

### 3.3 Importance: Video-Audio Cross-Attention Scoring

While text-guided scoring captures the relation between multimodal tokens and the user query, the textual context is often sparse relative to the rich audio-video content of long-form videos: complex reasoning frequently depends on multimodal cues that are not explicitly mentioned in the query. To complement query-conditioned relevance, we additionally compute a query-agnostic *importance* score that estimates intrinsic informativeness, in the spirit of query-independent document priors in IR such as centrality- or popularity-based scores [27]. Concretely, we use bidirectional cross-modal attention within each temporal block as a learnable proxy for this informativeness: tokens that interact more strongly with the opposing modality receive higher importance scores.

For each temporal block, let $E_m \in \mathbb{R}^{N_m \times d}$ denote the embeddings of modality $m \in \{V, A\}$ within that block, and let $\bar{m}$ denote its opposing modality (i.e., $\bar{V} = A$, $\bar{A} = V$). We compute bidirectional cross-attention between the two modalities to capture cross-modal interaction.

Specifically, we first project the embeddings into query and key spaces using learnable projection matrices $W_q^{\text{imp}}, W_k^{\text{imp}} \in \mathbb{R}^{d \times d}$:

$$Q_m = E_m W_q^{\text{imp}}, \quad K_m = E_m W_k^{\text{imp}}, \quad m \in \{V, A\}. \tag{4}$$

The bidirectional cross-attention matrix from modality $\bar{m}$ to $m$ is then computed as:

$$A_{\bar{m}m} = \frac{Q_{\bar{m}} K_m^\top}{\sqrt{d}} \in \mathbb{R}^{N_{\bar{m}} \times N_m}. \tag{5}$$

We obtain a token-level importance score by averaging, for each token $x_i$, the attention logits it receives from all tokens $x_j$ of the opposing modality within the same temporal block:

$$\mathrm{score}_{\text{imp}}(x_i) = \frac{1}{N_{\bar{m}_i}} \sum_j (A_{\bar{m}_i m_i})_{j,i}. \tag{6}$$

With per-token relevance and importance scores in hand, we fuse them into a combined score that drives the subsequent selection. Since $\mathrm{score}_{\text{rel}}(x_i)$ and $\mathrm{score}_{\text{imp}}(x_i)$ arise from different attention mechanisms and modalities, their raw magnitudes are not directly comparable. To place them on a common scale, we apply Z-score normalization within each scoring method and modality, yielding $\mathrm{score}'_{\text{rel}}(x_i)$ and $\mathrm{score}'_{\text{imp}}(x_i)$, and define the combined per-token score as their average:

$$\mathrm{score}(x_i) = \frac{1}{2} \left( \mathrm{score}'_{\text{rel}}(x_i) + \mathrm{score}'_{\text{imp}}(x_i) \right). \tag{7}$$

### 3.4 Diversity: Temporal-Aware Maximal Marginal Relevance Selecting

Retrieving tokens purely by $\mathrm{score}(x_i)$ often yields severe redundancy, since the temporal continuity of natural audio-video streams causes high-scoring tokens to cluster within adjacent, highly similar segments—an issue analogous to the redundancy problem in top-$K$ retrieval results. Inspired by *result diversification* in IR, we therefore add a diversity-aware re-ranking stage on top of the per-token scores to ensure that the limited token budget is not wasted on redundant tokens.

A natural choice is to greedily select tokens via the Maximal Marginal Relevance (MMR) objective [2], a classical diversification method in information retrieval:

$$\mathrm{MMR}(x_i) = (1 - \lambda) \cdot \mathrm{s}(x_i) - \lambda \cdot \max_{x_{i'} \in S_{\text{select}}} \mathrm{sim}(x_i, x_{i'}), \tag{8}$$

where $\mathrm{s}(x_i)$ denotes the base score of token $x_i$, $S_{\text{select}}$ denotes the set of tokens already selected in previous iterations, and $\lambda \in [0, 1]$ balances the base score against redundancy with the selected tokens.

However, conventional MMR is designed in a time-agnostic manner. Applying MMR to long-form audio-video introduces significant bias; for instance, it would incorrectly suppress a semantically similar but temporally distinct event occurring at the end of the video simply because a similar event happened at the beginning. Therefore, we propose Temporal-Aware MMR (TA-MMR). In contrast to MMR, TA-MMR constrains the novelty calculation to a local temporal window. For a candidate token $x_i$ at temporal index $\tau_i$, the objective function is formulated as:

$$\text{TA-MMR}(x_i) = (1-\lambda)\cdot \text{score}'(x_i) - \lambda \cdot \max_{x_{i'} \in S_{\text{select}} \cap \text{Window}(\tau_i)} \text{sim}(x_i, x_{i'}), \tag{9}$$

where $\text{Window}(\tau_i) = [\tau_i - W, \tau_i + W]$ defines the local temporal scope centered at $\tau_i$ with radius $W$, and $\text{sim}(\cdot,\cdot) \in [-1,1]$ denotes the mean-centered cosine similarity between token representations. The similarity is restricted to tokens of the same modality, as features from different modalities reside in heterogeneous spaces, where directly computing similarity would yield unreliable redundancy estimates. And $\text{score}'(x_i) \in [0,1]$ is obtained by min–max normalizing $\text{score}(x_i)$ across tokens of modality $m_i$. By only penalizing similarities within the local context, TA-MMR suppresses informationally repetitive adjacent tokens while preserving similar but temporally distinct events across the hour-level duration.

Given a total budget $K$, we further split it into a modality-aware budget $(K_{\text{video}}, K_{\text{audio}})$ with $K_{\text{video}} + K_{\text{audio}} = K$, and perform TA-MMR selection independently within each modality. This decoupled allocation prevents either modality from dominating the context budget and ensures a balanced cross-modal representation. For each modality $m \in \{V, A\}$ with budget $K_m$, we iteratively select the token maximizing the TA-MMR objective and add it to $S_{\text{select}}$, until $K_m$ tokens of that modality have been chosen. The selected tokens from both modalities together form the final set $S$, which is re-ordered by temporal index $\tau$ before being passed to the LLM.

# 4 Experiments

## 4.1 Implementation details.

**Model Configuration and Training.** We build our model upon MiniCPM-o 4.5 [10], initializing the architecture with pre-trained MiniCPM-o 4.5 checkpoints, with the compression module initialized randomly. The model is trained on 40k samples drawn from datasets including AVSD [1], How2 [29], FineVideo [12], ChronusAV [5], and LongVILA_sft [7]. This diverse collection encompasses a wide range of tasks, including audio-video speech recognition, video and audio captioning, and audio-video question answering. For video preprocessing, we sample 1 frame per second (FPS) for videos up to 320 seconds, and uniformly sample 320 frames for longer videos. Each temporal block aligns the video tokens from one sampled frame with the audio tokens from the corresponding time interval, spanning 1 second for videos up to 320 seconds and $L/320$ seconds for longer videos of duration $L$.

The training process is divided into two stages. In Stage 1, we disable the compression module and fine-tune only the LLM on 20k samples for one epoch. In Stage 2, we activate the randomly initialized compression module and jointly train it with the LLM using the remaining 20k samples for one epoch. The vision encoder, audio encoder, and adapter remain frozen throughout. This two-stage training approach aligns the LLM with the training data distribution in Stage 1, ensuring it provides a stable, high-quality gradient signal to the compression module in Stage 2. For the compression module, we use a learning rate of $5 \times 10^{-5}$, whereas the LLM is fine-tuned with a more conservative rate of $5 \times 10^{-6}$. To enable the model to adapt to different compression ratios, the token retention ratio is randomly sampled from a range of 0.1 to 1.0 during each Stage 2 training iteration.

**Differentiable Top-$k$ via Gumbel-Softmax.** A key challenge in training the compression module is the non-differentiability of the $K$-selection process. To enable end-to-end gradient propagation from the next-token prediction loss back to the projection layers of the compression module, we implement a differentiable top-$K$ selection strategy based on Gumbel-Softmax [14]. We utilize a Straight-Through Estimator during the forward pass: a hard $K$-hot mask is generated to select the discrete set of informative tokens for the subsequent LLM processing. In the backward pass, gradients are propagated through the continuous Gumbel-Softmax relaxations, bypassing the non-differentiable selection. We set the Gumbel-Softmax temperature to 1.0 to maintain a balance between sampling exploration and selection accuracy. Note that TA-MMR is disabled during training and only activated at inference, as its greedy iterative selection is incompatible with parallel differentiable top-$K$.

Table 1: Performance comparison on long-form audio-video understanding benchmarks. All compared models use LLM backbones at the 7–8B scale. All results are reported as accuracy (%). The best and second-best results are marked in **bold** and underlined, respectively. The last row reports the absolute improvement of AVOC over the second-best result in each column.

| Method | WorldSense (up to 10min) | OmniVideoBench (up to 30min) | | LVOmniBench (up to 90min) | | |
|---|---|---|---|---|---|---|
| | Avg. | (10,30] min | Avg. | Medium | High | Avg. |
| VideoLLaMA2-7B [8] | 25.4 | 28.3 | 29.2 | 26.8 | 28.2 | 27.2 |
| Baichuan-Omni-1.5 [21] | - | 32.4 | 30.7 | - | - | - |
| HumanOmni-7B [47] | 47.1 | 29.3 | 30.5 | - | - | - |
| Qwen2.5-Omni-7B [41] | 45.4 | 26.7 | 29.3 | 29.9 | 28.3 | 32.0 |
| video-SALMONN 2+ 7B [35] | <u>50.9</u> | <u>34.6</u> | <u>37.4</u> | 30.2 | 26.7 | 32.7 |
| MiniCPM-o 2.6 [10] | 44.3 | 26.2 | 29.7 | 33.8 | 26.6 | 32.5 |
| MiniCPM-o 4.5 [10] | 50.3 | 31.6 | 36.9 | <u>34.1</u> | 25.1 | <u>34.8</u> |
| OmniZip [36] | 46.3 | 32.3 | 36.1 | 31.8 | <u>30.1</u> | 32.5 |
| AVOC (Ours) | **52.6** | **39.8** | **42.3** | **41.3** | **35.5** | **40.3** |
| $\Delta$ over 2nd-best | +1.7 | +5.2 | +4.9 | +7.2 | +5.4 | +5.5 |

### 4.2 General Long-Form Audio-Video Understanding Evaluation

**Evaluation Settings.** We evaluate AVOC on three long-form audio-video benchmarks: WorldSense [13], OmniVideoBench [18], and LVOmniBench [37]. WorldSense (up to 10 min) covers diverse real-world scenarios, and we report its average accuracy. OmniVideoBench (up to 30 min) emphasizes audio-video reasoning with strong modality complementarity; we report results on its ultralong "(10, 30] min" subset as well as the overall average. LVOmniBench (10–90 min) stratifies questions into Low, Medium, and High difficulty tiers according to factors such as average video duration and information granularity; we report accuracy on the Medium and High subsets together with the average across all tiers.

When evaluating AVOC, we follow the same video preprocessing pipeline used during training: videos are sampled at 1 FPS for those shorter than 320 seconds and uniformly sampled 320 frames for longer videos, while the accompanying audio stream is input in full. We activate the compression module, and adopt a fixed global token budget of $K = 10240$, with a modality token budget allocation ratio of $K_{\text{video}} : K_{\text{audio}} = 2{:}1$, matching the video-heavy information density in the target benchmarks. The TA-MMR diversity weight $\lambda = 0.15$ and the local temporal window radius $W = 3$. For the baselines, we use the official configurations for each model and evaluate using the maximum permissible number of frames and audio length.

**Performance.** As shown in Table 1, AVOC consistently achieves state-of-the-art performance across all three benchmarks, with absolute gains of 1.7–7.2 points over the second-best results. Two observations are worth highlighting. First, AVOC's advantage scales with video duration: compared to the relatively shorter WorldSense (+1.7), AVOC achieves larger gains on the much longer OmniVideoBench (+4.9 on Avg., +5.2 on (10, 30] min subset) and LVOmniBench (+5.5 on Avg., up to +7.2 on medium difficulty subset). This confirms that our compression module is particularly effective for ultra-long audio-video understanding, where context-window pressure and information redundancy are most severe. Second, AVOC delivers consistent improvements over OmniZip, which is the most directly comparable token-compression baseline, with gains of +6.3 on WorldSense, +6.2 on OmniVideoBench Avg., and +7.8 on LVOmniBench Avg. This validates the effectiveness of our compression design.

### 4.3 Audio-Video Needle-in-a-Haystack

**Evaluation Settings.** To assess the fine-grained retrieval capability of AVOC over long audio-video streams, we construct an Audio-Video Needle-in-a-Haystack (AV-NIAH) evaluation. For each audio-video haystack, we inject a "needle" carrying a secret keyword—a randomly generated 6 digit numeric string—at a controlled temporal position. The needle is instantiated in two modalities: (i) a *vision needle*, rendered as a caption "The secret word is `<needle>`" overlaid on a single video frame; and (ii) an *audio needle*, synthesized via text-to-speech reading "The secret word is `<needle>`" and spliced into the audio stream. During evaluation, videos are sampled at 1 FPS and fed into the model together with the full accompanying audio stream. We evaluate the visual and auditory needles separately, prompting the model with the query "What is the secret number?" and requiring

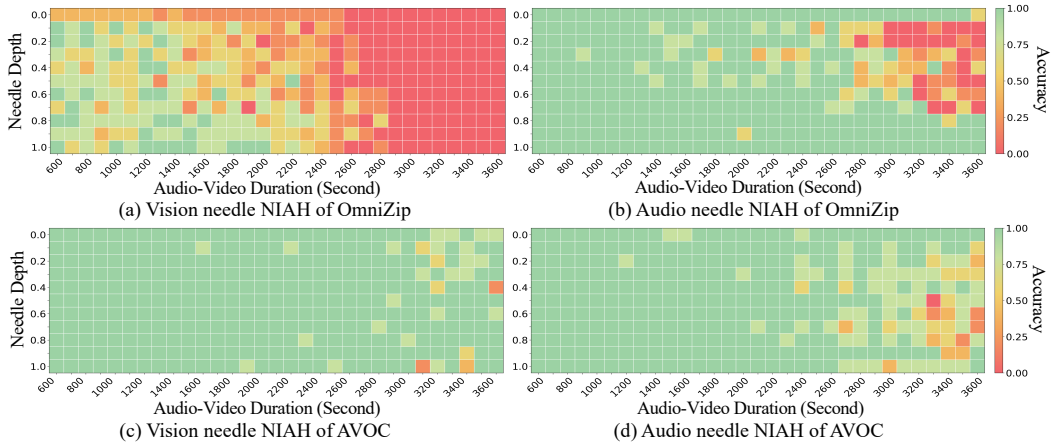


Figure 3: Audio-Video Needle-in-a-Haystack results. Each cell reports retrieval accuracy at a given audio-video duration (x-axis) and relative needle depth (y-axis).

Table 2: Ablation study on the compression components of AVOC. TGS: Text-Guided cross-attention Scoring; VAS: Video-Audio cross-attention Scoring; TA-MMR: Temporal-Aware Maximal Marginal Relevance. "Random" replaces the scoring-and-selection procedure with uniform random sampling under the identical token budget and modality allocation.

| Variant | Component | | | OmniVideoBench | | LVOmniBench | | |
|---|---|---|---|---|---|---|---|---|
| | TGS | VAS | TA-MMR | (10,30] min | Avg. | Medium | High | Avg. |
| Random Selection | – | – | – | 27.5 | 37.2 | 33.7 | 32.5 | 35.1 |
| w/o TGS | | ✓ | ✓ | 38.2 | 41.4 | 39.9 | 34.4 | 39.1 |
| w/o VAS | ✓ | | ✓ | 35.0 | 41.1 | 40.1 | 34.9 | 39.5 |
| w/o TA-MMR | ✓ | ✓ | | 36.5 | 41.8 | 39.2 | 32.9 | 38.7 |
| AVOC (full) | ✓ | ✓ | ✓ | **39.8** | **42.3** | **41.3** | **35.5** | **40.3** |

it to localize and extract the digit string from the target modality. We iterate over various needle depths (where the needle is placed) and audio-video lengths (up to 3600 seconds) to measure the performance, and report accuracy as the exact-match rate between the predicted and ground-truth digit strings. The more detailed evaluation setup is in the Appendix A.1.

**Performance.** As shown in Figure 3, OmniZip exhibits a clear duration-induced collapse: its accuracy degrades substantially beyond 2000s on the vision needle (Figure 3a) and beyond 3000s on the audio needle (Figure 3b). In contrast, AVOC maintains high retrieval accuracy across the entire duration-depth grid for both modalities, with only minor degradation appearing at isolated cells beyond 3000s (Figure 3c, d). These results demonstrate AVOC's capability in ultra-long audio-video context modeling. Additional AV-NIAH results on more baselines are provided in the Appendix A.2.

### 4.4 Ablation Studies

**Effect of Compression Components.** To validate the effectiveness of each component in our compression module, we conduct a series of ablation studies on OmniVideoBench and LVOmniBench. All ablated variants in Table 2 adopt the same default hyperparameter configuration as in Section 4.2, isolating the effect of compression components. As shown in Table 2, using random selection yields a substantial performance drop compared to our full model. This indicates that, under tight token budgets, which tokens are retained matters far more than how many, confirming the necessity of a content-aware compression mechanism for ultra-long audio-video understanding. Beyond this, removing any single component consistently degrades performance, demonstrating that relevance, importance, and diversity contribute complementary rather than redundant signals.

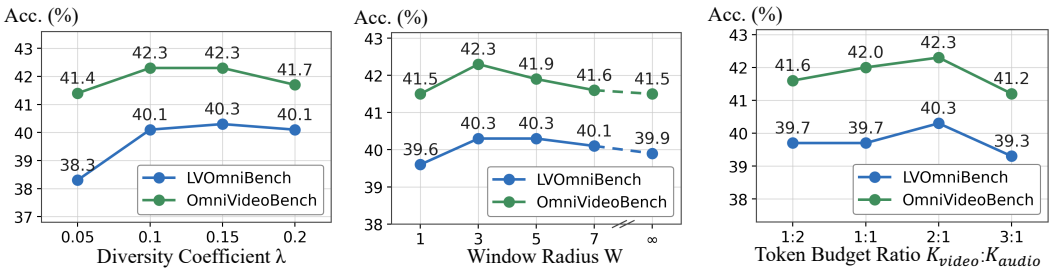


Figure 4: Ablation on the diversity coefficient $\lambda$ (left) and the local window radius $W$ (right) of TA-MMR.

Figure 5: Ablation on modality token budget ratio.

Table 3: Latency comparison between AVOC and its backbone MiniCPM-o 4.5. $\rho$ denotes the token retention ratio.

| Model | $\rho$ | Time to First Token (s) | Prefilling (s) | Compression Module (s) |
|---|---|---|---|---|
| MiniCPM-o 4.5 | – | $26.271 \pm 0.146$ | $4.453 \pm 0.002$ | – |
| AVOC | 1.0 | $28.226 \pm 0.237$ | $4.442 \pm 0.004$ | $1.834 \pm 0.007$ |
| | 0.5 | $25.015 \pm 0.105$ | $2.088 \pm 0.002$ | $0.929 \pm 0.003$ |
| | 0.1 | $22.549 \pm 0.196$ | $0.497 \pm 0.001$ | $0.260 \pm 0.001$ |

**Effect of TA-MMR Hyperparameters.** We examine the two key hyperparameters of TA-MMR: the diversity coefficient $\lambda$ and the local window radius $W$. As shown in Figure 4, both hyperparameters exhibit a unimodal trend on OmniVideoBench and LVOmniBench, peaking at $\lambda = 0.15$ and $W = 3$. Setting $\lambda$ or $W$ too small leaves adjacent duplicated tokens unpenalized, whereas too large $\lambda$ or $W$ biases selection toward merely dissimilar tokens. Notably, when $W \to \infty$, TA-MMR degenerates into the standard MMR, and the observed performance drop empirically validates the necessity of our temporal-window design.

**Effect of Modality Token Budget Ratio.** We investigate the impact of the modality token budget allocation $K_{\text{video}} : K_{\text{audio}}$. As shown in Figure 5, performance on both OmniVideoBench and LVOmniBench peaks at 2:1. An audio-leaning allocation (e.g., 1:2 or 1:1) under-represents the dense visual cues, whereas an overly video-skewed allocation (3:1) starves the audio stream of speech, environmental sounds, and music that carry irreplaceable semantic information. These results indicate that a moderately video-leaning budget best matches the information density of real-world long-form audio-video content.

### 4.5 Efficiency Analyses

We compare the latency of AVOC against its backbone MiniCPM-o 4.5, which shares the identical architecture except for the compression module. For a fair comparison, both models are fed with the same 10-minute video uniformly sampled to 128 frames, together with the full accompanying audio stream. We evaluate AVOC under three token retention ratios ($\rho \in \{1.0, 0.5, 0.1\}$) to examine how latency varies with compression aggressiveness. All measurements are performed on a single NVIDIA A800 GPU with BF16 precision and flash-attention2, and we report the average time over multiple runs to mitigate measurement noise.

As summarized in Table 3, the compression module introduces only modest overhead. Even at $\rho = 1.0$ where no token is dropped, the compression module adds merely 1.834 s, yielding a slight Time-to-First-Token increase over the backbone. This confirms that the compression module is computationally lightweight relative to the LLM forward pass. More importantly, reducing the retention ratio yields substantial prefilling speedups: prefilling latency drops from 4.453 s (backbone) to 2.088 s at $\rho = 0.5$ and 0.497 s at $\rho = 0.1$, a nearly $9\times$ reduction. Meanwhile, the compression module's own cost scales nearly proportionally with $\rho$ (1.834 s, 0.929 s, 0.260 s for $\rho = 1.0, 0.5, 0.1$), so more aggressive compression incurs smaller compression-time overhead, making AVOC particularly suitable for hour-level audio-video understanding scenarios that demand high compression ratios.

## 5 Conclusion

We presented AVOC, a framework that enhances hour-level audio-video understanding in Omni-Modal LLMs through a learnable token compression module. Drawing on classical principles from information retrieval, the module instantiates three complementary criteria: text-guided relevance, bidirectional video-audio importance, and Temporal-Aware Maximal Marginal Relevance for local diversity. These criteria jointly guide the selection of a compact, informative token subset under a tight context budget. Experiments on multiple long-form audio-video benchmarks show that AVOC achieves state-of-the-art performance, surpassing the second-best method by up to 5.5 points in average accuracy, and maintains robust retrieval on Audio-Video Needle-in-a-Haystack task at durations up to one hour. We hope AVOC offers a step toward Omni-Modal LLMs capable of reasoning over the rich, hour-long multimodal content that pervades real-world applications.

## Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (No. 62276268) and Huawei Inc. The authors declare no competing interests.

## References


[1] Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio visual scene-aware dialog. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7558–7567. Computer Vision Foundation / IEEE, 2019.

[2] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

[3] Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, and Liyun Ru. Jointavbench: A benchmark for joint audio-visual reasoning evaluation. *arXiv preprint arXiv:2512.12772*, 2025.

[4] 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 Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, Lecture Notes in Computer Science, pages 19–35. Springer, 2024.

[5] Yijing Chen, Yihan Wu, Kaisi Guan, Wenhui Tan, Yuchen Ren, Yuyue Wang, Ruihua Song, and Liyun Ru. Chronusomni: Improving time awareness of omni-modal large language models. In *European Conference on Computer Vision*, pages 54–73. Springer, 2026.

[6] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[7] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Yihui He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[9] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In Sung-Hyon Myaeng, Douglas W. Oard, Fabrizio Sebastiani, Tat-Seng Chua, and Mun-Kew Leong, editors, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008*, pages 659–666. ACM, 2008.

[10] Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

[11] Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, Yuanxing Zhang, Jiaheng Liu, Qiang Liu, Pengfei Wan, and Liang Wang. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *CoRR*, abs/2602.04804, 2026.

[12] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`, 2024.

[13] 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.

[14] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[15] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13700–13710. IEEE, 2024.

[16] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics, 2020.

[17] Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models–from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2026.

[18] Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Wentao Wang, Zhenghao Song, Dingling Zhang, et al. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*, 2025.

[19] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19108–19118, 2022.

[20] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. *CoRR*, abs/2501.00574, 2025.

[21] 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*, 2025.

[22] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, Lecture Notes in Computer Science, pages 323–340. Springer, 2024.

[23] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models. *CoRR*, abs/2409.15272, 2024.

[24] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

[25] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *CoRR*, abs/2502.04328, 2025.

[26] Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 12585–12602. Association for Computational Linguistics, 2024.

[27] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bring order to the web. In *Proc. of the 7th International World Wide Web Conf.–1998*, 1999.

[28] Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, 2009.

[29] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, et al. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018.

[30] 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. *CoRR*, abs/2507.20198, 2025.

[31] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. Longvu: Spatiotemporal adaptive compression for long video-language understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research, 2025.

[32] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 26160–26169. Computer Vision Foundation / IEEE, 2025.

[33] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 18221–18232. IEEE, 2024.

[34] Wenhui Tan, Xiaoyi Yu, Jiaze Li, Yijing Chen, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Msjoe: Jointly evolving mllm and sampler for efficient long-form video understanding. *arXiv preprint arXiv:2602.22932*, 2026.

[35] 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*, 2025.

[36] 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. *CoRR*, abs/2511.14582, 2025.

[37] Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, et al. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms. *arXiv preprint arXiv:2603.19217*, 2026.

[38] Qwen Team. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025.

[39] Hongchen Wei and Zhenzhong Chen. Visual context window extension: A new perspective for long video understanding. In Cathal Gurrin, Klaus Schoeffmann, Min Zhang, Luca Rossetto, Stevan Rudinac, Duc-Tien Dang-Nguyen, Wen-Huang Cheng, Phoebe Chen, and Jenny Benois-Pineau, editors, *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 4281–4289. ACM, 2025.

[40] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXIII*, Lecture Notes in Computer Science, pages 453–470. Springer, 2024.

[41] 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. *CoRR*, abs/2503.20215, 2025.

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19792–19802, 2025.

[43] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, Yuming Lou, Dong Yang, Zhijian Liu, Yukang Chen, Ambrish Dantrey, Ehsan Jahangiri, Sreyan Ghosh, Daguang Xu, Ehsan Hosseini-Asl, Danial Mohseni-Taheri, Vidya Murali, Sifei Liu, Yao Lu, Oluwatobi Olabiyi, Yu-Chiang Frank Wang, Rafael Valle, Bryan Catanzaro, Andrew Tao, Song Han, Jan Kautz, Hongxu Yin, and Pavlo Molchanov. Omnivinci: Enhancing architecture and data for omni-modal understanding LLM. *CoRR*, abs/2510.15870, 2025.

[44] Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. Vscan: Rethinking visual token reduction for efficient large vision-language models. *Trans. Mach. Learn. Res.*, 2026, 2026.

[45] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *Trans. Mach. Learn. Res.*, 2025, 2025.

[46] Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20857–20867, 2025.

[47] Jiaxing Zhao, Qize Yang, Yixing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao Fu, Xihan Wei, Liefeng Bo, et al. Humanomni: A large vision-speech language model for human-centric video understanding. *arXiv preprint arXiv:2501.15111*, 2025.

[48] 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.

# A Additional Details and Results on Audio-Video Needle-in-a-Haystack

## A.1 Evaluation Setting Details.

We provide additional details on the construction and evaluation protocol of the Audio-Video Needle-in-a-Haystack (AV-NIAH) task introduced in Section 4.3.

**Haystack source.** We use a long-form audio-video clip drawn from LVOmniBench with a total duration exceeding 60 minutes as the haystack. To construct samples of varying lengths, we truncate the clip to target durations ranging from 100s to 3600s with a step size of 100s, yielding 36 duration settings in total. The accompanying audio stream is preserved in alignment with the truncated video.

**Needle generation.** Each needle carries a secret keyword instantiated as a randomly generated 6-digit numeric string. We sample 5 independent needles in total, and report the average accuracy across these 5 samples for each (duration, depth) cell to mitigate the variance introduced by individual needle realizations. The needle is rendered in two modalities, evaluated separately:

(1) Vision needle. The keyword is rendered as the caption "The secret word is <needle>". After sampling the video at 1 FPS, we overlay the keyword on a single sampled frame at the target depth.

(2) Audio needle. The audio clip reading 'The secret word is <needle>' is synthesized using Qwen3-TTS. To ensure the audio needle blends naturally into the haystack rather than standing out as an acoustic outlier, we normalize the loudness of the synthesized audio clip to match the average volume of the haystack audio stream before splicing it into the target temporal position.

**Duration–depth grid.** For each duration setting, we vary the relative needle depth (i.e., the normalized temporal position of the needle within the clip) over 11 evenly spaced values from 0.0 to 1.0 with a step size of 0.1. This yields a 36 × 11 (duration × depth) evaluation grid for each modality. At each cell, we average over the 5 needle samples described above.

**Inference protocol.** AV-NIAH does not impose a maximum frame number or audio length cap. For both AVOC and all baselines, videos are uniformly sampled at 1 FPS and the full accompanying audio stream is fed into the model, ensuring that the needle is never discarded by preprocessing-stage subsampling. For AVOC, the compression module is activated with global token budget $K = 25000$, modality allocation $K_{\text{video}} : K_{\text{audio}} = 2{:}1$, TA-MMR diversity coefficient $\lambda = 0.15$, local temporal window radius $W = 3$. The model is prompted with the query *"What is the secret number?"* and asked to localize and extract the digit string from the target modality. Vision and audio needles are evaluated in independent runs.

## A.2 Additional Baseline Results.

To provide a more comprehensive view of fine-grained retrieval capability across long audio-video durations, we extend the AV-NIAH evaluation in Section 4.3 to two additional baselines: MiniCPM-o 4.5 [10] (the backbone of AVOC) and Qwen2.5-Omni-7B [41]. The complete results across all four models are presented in Figure 6, where each row corresponds to one model and the two columns report the vision needle and audio needle retrieval accuracy, respectively.

**MiniCPM-o 4.5.** As shown in Figure 6(a, b), MiniCPM-o 4.5 exhibits a severe and immediate context-window collapse: its successful retrieval is restricted to durations below approximately 300 seconds for both vision and audio needles, and drops to near-zero across the entire duration-depth grid beyond this threshold. This collapse stems from the rigid context-window constraint of the backbone, which is unable to accommodate the dense token sequences produced by hour-level audio-video streams without aggressive content-agnostic truncation.

**Qwen2.5-Omni.** As shown in Figure 6(c, d), Qwen2.5-Omni extends the effective retrieval range substantially compared to MiniCPM-o 4.5, but still exhibits a clear duration-induced degradation. On the vision needle, accuracy degrades noticeably beyond 1500s and collapses to near-zero beyond 2300s. The audio needle is comparatively more robust, sustaining moderate accuracy up to around 2500s before degrading sharply at longer durations.

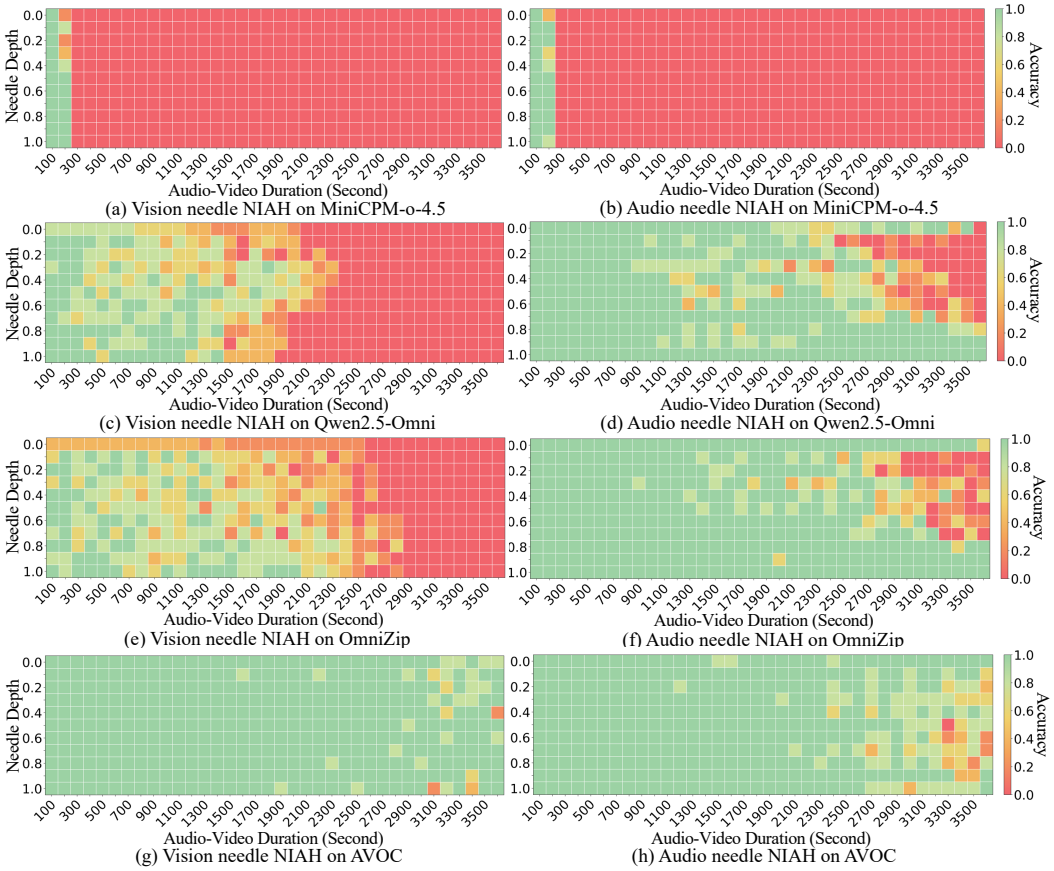


Figure 6: Extended Audio-Video Needle-in-a-Haystack results across four models. Each row corresponds to one model, and each cell reports retrieval accuracy at a given audio-video duration (x-axis, 100s–3600s) and relative needle depth (y-axis, 0.0–1.0). The left and right columns report vision needle and audio needle retrieval accuracy, respectively.

**OmniZip.** As shown in Figure 6(e, f), OmniZip pushes the vision-needle effective retrieval range, achieving a moderate improvement over Qwen2.5-Omni. However, as the duration approaches one hour, the vision-needle accuracy degrades substantially across nearly all depths, and the audio needle exhibits lower accuracy particularly beyond 3000s at shallow-to-mid depths.

**AVOC.** In contrast, as shown in Figure 6(g, h), AVOC maintains consistently high retrieval accuracy across the entire 100s–3600s duration range and across all needle depths for both vision and audio needles, with only minor degradation appearing at isolated cells beyond 3000s. This demonstrates AVOC's robust fine-grained information localization capability over hour-level audio-video streams.

## B Generalization to an Additional Backbone

To evaluate the generalizability of our method, we adapt the compression module of AVOC to Qwen2.5-Omni. We compare the resulting model with the original Qwen2.5-Omni and OmniZip, and the latter also performs compression on the same backbone.

**Evaluation settings.** All three models use identical input settings: video frames sampled at 2 FPS with a maximum of 768 frames, and the full accompanying audio stream. The original Qwen2.5-Omni baseline processes these inputs without additional token compression. For the two compression-based variants, OmniZip and our method, we use the same audio-video token budget of 32,768.

Table 4: Performance comparison using Qwen2.5-Omni as the backbone. Results are reported as accuracy (%).

| Model | WorldSense | OmniVideoBench | LVOmniBench |
|---|---|---|---|
| Qwen2.5-Omni | 45.4 | 29.3 | 32.0 |
| OmniZip | 46.3 | 36.1 | 32.5 |
| Qwen2.5-Omni + Our Compression Module | **53.9** | **39.5** | **37.4** |

**Performance.** As shown in Table 4, Qwen2.5-Omni equipped with our compression module achieves the highest accuracy across all three benchmarks. Compared with OmniZip under matched input settings and token budgets, our method improves accuracy by 7.6, 3.4, and 4.9 percentage points on WorldSense, OmniVideoBench, and LVOmniBench, respectively. These results demonstrate that the compression module of AVOC is also effective when integrated into Qwen2.5-Omni, providing additional evidence of its generalizability across omni-modal backbones.

# C Limitations

Despite the strong empirical results, AVOC has several limitations that point to directions for future work. First, the compression module operates in an offline manner: the entire audio-video stream must be available before scoring and selection can be performed, which prevents AVOC from being directly applied to streaming scenarios where tokens arrive incrementally. Extending the framework to causal or chunk-wise online settings is a natural next step. Second, our experiments are conducted at a single, relatively small parameter scale; whether the proposed compression mechanism scales gracefully to larger backbones remains to be verified. Third, the modality token budget allocation ratio $K_{\text{video}} : K_{\text{audio}}$ is currently set as a fixed hyperparameter. A content-adaptive allocation that dynamically adjusts the ratio based on per-sample audio-video information density could further improve robustness across diverse multimodal distributions.

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We list our three main contributions in Section 1.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: In Appendix.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: The paper does not contain theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Section 4 describe the architecture, hyperparameters, datasets, and training procedure.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Our code is publicly available. The training and evaluation datasets are public.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Training/test details are reported in Section 4.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: Due to computational cost, we report single-run results on the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: All trainings were run on 8× A800 80GB GPUs. Each main experiment takes 24 hours.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and the research conforms to it in all respects.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: AVOC enables Omni LLMs to comprehend hour-level audio-video content under a tight context budget, which can benefit applications such as accessibility (e.g., generating descriptions of long-form videos for visually or hearing impaired users), educational content summarization, and efficient analysis of meeting recordings or instructional videos. On the negative side, improved long-form audio-video understanding could in principle be misused for large-scale automated surveillance or non-consensual analysis of personal recordings. We note, however, that our work focuses on token compression for existing Omni LLM backbones rather than introducing new generative or recognition capabilities, so the marginal contribution to such misuse risks is limited. We do not release any new pretrained generative model that could directly enable harmful synthesis.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The released AVOC checkpoint adapts MiniCPM-o 4.5 for long-form audio-video understanding through token compression. We consider the additional misuse risk from this adaptation limited.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: CC-BY 4.0 and Apache-2.0.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

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

13. **New assets**

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

    Answer: [Yes]

    Justification: Our project repository links to the released AVOC checkpoint and documents the software environment, data formats, training procedure, and inference commands.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subject research.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were only used for writing assistance and did not impact the methodology.

    Guidelines:

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