EVOCOMP-2: LEARNING TO COMPRESS AUDIO-VIDEO TOKENS FOR OMNIMODAL LARGE LANGUAGE MODELS VIA TOPIC-GUIDED EVOLUTIONARY LABELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal Large Language Models (OmniLLMs) achieve strong performance on audio-video tasks but incur substantial computational overhead due to the large number of audio and video tokens. EvoComp proposes to adopt evolutionary search to construct token-retention supervision and train a lightweight compressor for visual token compression. In this work, we extend EvoComp to OmniLLMs and propose EvoComp-2 for efficient audio-video token compression. Extending EvoComp from static visual inputs to synchronized audio-video sequences introduces new challenges arising from temporal redundancy and query-dependent modality importance. To address them, EvoComp-2 introduces topic-guided temporal clustering that adapts a topic model to video frames and partitions audio-video sequences into semantically coherent intervals before evolutionary labeling, enabling more structured and efficient token selection. We further construct modality-level importance supervision to capture the relative contribution of audio and video to different queries. For compressor training, we introduce a temporal-window reweighting loss to emphasize information-rich temporal regions and a modality-level importance loss to learn query-dependent modality preference, together with the GHM loss inherited from EvoComp. Extensive experiments demonstrate that EvoComp-2 achieves favorable accuracy-efficiency trade-offs. On WorldSense, EvoComp-2 preserves 98.7% of the original accuracy while retaining only 35% of the audio-video tokens. In terms of inference efficiency, it achieves a 1.6x end-to-end speedup on DailyOmni with 3.3x token compression.

1 INTRODUCTION

Omnimodal Large Language Models (OmniLLMs) (Sun et al., 2024; Zhao et al., 2025; Xu et al., 2025a) have demonstrated increasingly strong capabilities in a broad spectrum of audio-video tasks, including audio-visual question answering (Tang et al., 2025), audio-visual understanding (Ghosh et al., 2026; Xing et al., 2026), cross-modal temporal reasoning (Xu et al., 2025b; Ghosh et al., 2026), and audio-visual temporal grounding (Xing et al., 2026; Team, 2026). Such capabilities, however, come at a substantial computational cost. OmniLLMs typically encode a video into densely sampled video frames and synchronized audio segments, producing extremely long sequences of audio and video tokens (Tao et al., 2026). These long multimodal contexts substantially increase the memory consumption and inference latency (Kong et al., 2025; Shao et al., 2026; Tao et al., 2026).

Fortunately, dense audio-video token sequences contain substantial redundancy. On the visual side, neighboring patches within a frame are often redundant, while adjacent frames exhibit strong temporal overlap (Jin et al., 2024). For audio streams, neighboring tokens often encode highly correlated speech or background sounds (Lin et al., 2025). Audio-video token compression therefore provides a promising approach to reduce OmniLLM inference cost without modifying the underlying Large Language Model (LLM). Recent studies use salient audio tokens to guide video token pruning (Tao et al., 2026), or extract visual anchors to guide audio token selection (Ding et al., 2026). Other approaches introduce query-aware compression (Yang et al., 2026). Despite their effectiveness, their token selection is not directly optimized according to the correctness of the final OmniLLM output.

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

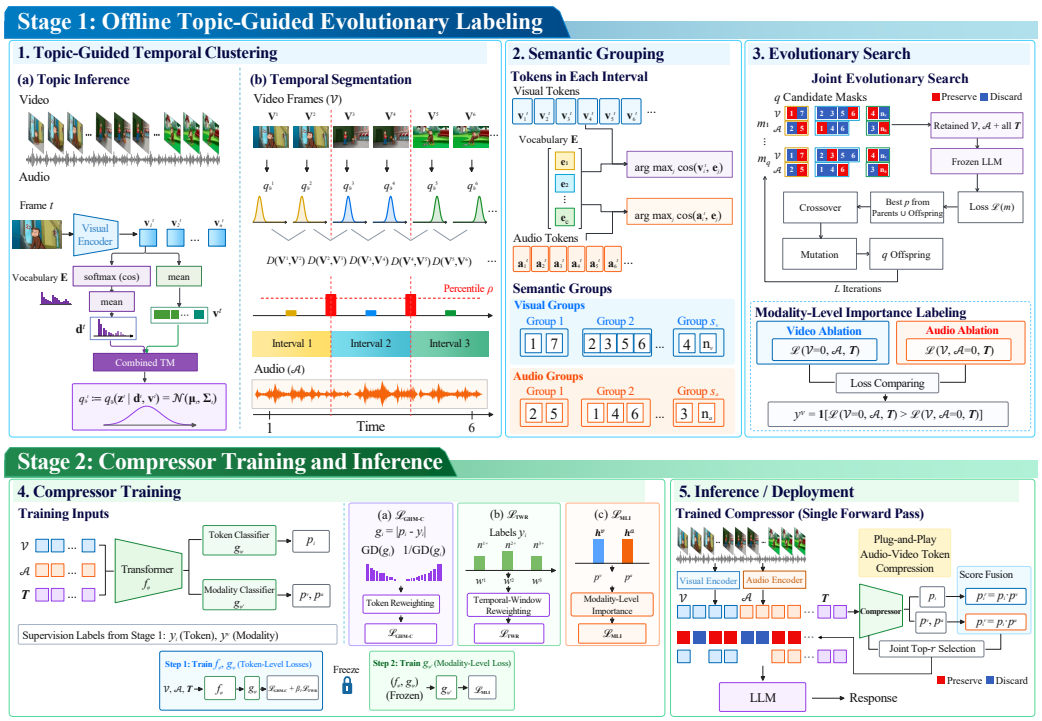


Figure 1: Overview of EvoComp-2. The framework consists of two stages: 1) offline topic-guided evolutionary labeling that constructs token- and modality-level supervision; and 2) compressor training and inference, where a lightweight compressor learns token- and modality-level importance and performs audio-video token selection for efficient OmniLLM inference.

EvoComp (Song et al., 2026) addresses this issue for visual token compression in Multimodal Large Language Models (MLLMs) by using evolutionary search to construct task-oriented supervision for training a lightweight compressor. Extending this output-oriented token compression from images to synchronized audio-video inputs, however, introduces several new challenges. First, video tokens exhibit substantial redundancy along both spatial and temporal dimensions. Second, informative evidence is often highly concentrated in only a few temporal windows. Third, the relative importance of audio and video varies across queries.

To address these challenges, we propose EvoComp-2, a learning-based audio-video token compression framework for OmniLLMs. Given audio-video and text inputs, a lightweight compressor jointly predicts token-level retention probabilities for audio-video tokens and modality-level importance scores that characterize the overall relative importance of the audio and video streams. The two predictions jointly determine which audio-video tokens should be retained.

To construct supervision for the compressor, we extend the evolutionary labeling strategy of EvoComp to synchronized audio-video inputs. For each training sample, we search for a binary mask over audio and video tokens, where each candidate mask is evaluated by the target OmniLLM loss after feeding the retained audio-video tokens together with all text tokens into the LLM. To incorporate the temporal structure and redundancy of audio-video sequences into the search space, we introduce topic-guided temporal clustering. It adapts a topic model (Srivastava & Sutton, 2017; Dieng et al., 2020; Bianchi et al., 2021) to video frames and partitions them into semantically coherent temporal segments based on their latent topic distributions and transfers the same partition to the synchronized audio stream. Within each temporal cluster, audio and video tokens are grouped separately according to semantic similarity, and the search is constrained to retain at most one token from each group. We further construct modality-level importance labels by separately suppressing the audio and video tokens and comparing the resulting OmniLLM losses.

Given the resulting supervision, we train the compressor with objectives designed for the highly imbalanced nature of audio-video token selection. Following EvoComp, we adopt the GHM loss (Li et al., 2019), which reweights the gradients of individual tokens to alleviate both class and difficulty

imbalance. We further introduce a temporal-window reweighting loss to encourage the compressor to focus on information-rich and task-critical segments. Finally, a modality-level importance loss supervises the predicted relative importance of the two modalities. Together, these objectives enable the compressor to jointly learn token-level importance, temporal saliency, and modality preference.

Our main contributions are summarized as follows:

- We propose EvoComp-2, an audio-video token compression framework for OmniLLMs that constructs supervision for training a compressor to jointly predict token-level retention and modality-level importance.
- We introduce a structured evolutionary labeling strategy with topic-guided temporal clustering and semantic grouping, enabling the search to identify task-relevant and non-redundant audio-video tokens. Modality-level labels are further derived by separately ablating the audio and video inputs and comparing the resulting increases in OmniLLM loss.
- We design a training objective that combines GHM, temporal-window reweighting, and modality-level importance losses to learn token, temporal, and modality importance.
- Extensive experiments demonstrate the effectiveness and efficiency of EvoComp-2 across diverse audio-video understanding and reasoning benchmarks. Under 1.8x to 3.3x audio-video token compression, it achieves 1.3x to 1.6x end-to-end speedups while preserving 97.5% to 92.5% of the uncompressed performance.

2 Related Work

Visual Token Compression for Images and Videos. Visual token compression has been widely studied for efficient MLLMs. For images, training-free methods prune or merge tokens according to attention, feature similarity, or representation diversity (Chen et al., 2024; Wen et al., 2025; Yang et al., 2025a; Zhang et al., 2025; Xing et al., 2024). Learning-based approaches instead train dedicated modules to compress visual tokens, among which EvoComp (Song et al., 2026) uses evolutionary search to construct token labels for training a lightweight compressor. For videos, additional temporal redundancy arises because adjacent frames often contain highly overlapping content. Existing methods therefore explicitly exploit inter-frame redundancy through progressive frame compression, temporal token merging, or temporally structured pruning (Jin et al., 2024; Tao et al., 2025; Shen et al., 2025; Huang et al., 2025; Hyun et al., 2025). Our method leverages topic modeling (Srivastava & Sutton, 2017; Dieng et al., 2020; Bianchi et al., 2021) to partition video frames into semantically coherent temporal segments, thereby grouping frames with similar semantic content more effectively.

Audio Token Compression. Audio sequences also contain substantial temporal redundancy, as neighboring tokens often encode highly correlated acoustic information (Bhati et al., 2025; Luo et al., 2026; Xiang et al., 2026). Existing methods reduce audio-token length through temporal pooling, segmentation, or similarity-based merging (Bhati et al., 2025; Luo et al., 2026). Speech-Prune (Lin et al., 2025) further performs query-aware token pruning by filtering speech tokens according to speech-text similarity. More recent methods such as HeadRouter (He et al., 2026) exploit task-dependent attention patterns to preserve important audio tokens across different acoustic and semantic tasks. These studies highlight the importance of both temporal redundancy and task relevance in audio token compression.

Audio-Video Token Compression for OmniLLMs. Recent work has extended token compression to synchronized audio-video inputs in OmniLLMs. OmniZip (Tao et al., 2026) uses salient audio tokens to guide dynamic video pruning, while OmniSIFT (Ding et al., 2026) extracts compact visual anchors to guide audio-token selection. OmniSelect (Yang et al., 2026) introduces query-aware compression by dynamically allocating audio and video token budgets according to their relevance to the textual query. Despite their effectiveness, these approaches rely on heuristic signals such as attention, feature similarity, cross-modal correspondence, or query relevance, which may not faithfully reflect the true contribution of each token to the downstream task. In contrast, EvoComp-2 directly uses the output loss of the target OmniLLM to construct task-oriented audio-video token labels through evolutionary search, while explicitly incorporating temporal structure, semantic redundancy, and modality importance into the labeling and learning process.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

An OmniLLM typically consists of modality-specific encoders followed by a shared LLM (Xu et al., 2025a;b; Team, 2026). Given a video, the visual and audio encoders produce synchronized visual and audio token sequences. Different OmniLLMs may adopt different video sampling and audio-video temporal tokenization strategies. To provide a model-agnostic formulation, we define a visual temporal unit and its aligned audio temporal unit as the basic units of temporal organization. Specifically, we denote the visual and audio tokens associated with the t-th unit as $\boldsymbol{V}^t = \{\boldsymbol{v}_i^t\}_{i=1}^n$ and $\boldsymbol{A}^t = \{\boldsymbol{a}_i^t\}_{i=1}^m$, respectively. For simplicity, we refer to each visual temporal unit as a frame, although it may correspond to one or multiple sampled frames. These token sequences $\mathcal{V} = \{\boldsymbol{V}^t\}_{t=1}^K$ and $\mathcal{A} = \{\boldsymbol{A}^t\}_{t=1}^K$, together with text tokens $\boldsymbol{T}$, are then temporally organized and jointly fed into the LLM for response generation. As video duration increases, the rapidly growing number of audio-video tokens incurs substantial prefill computation and inference latency (Tao et al., 2026). Our goal is therefore to compress these tokens while preserving the output quality of the target OmniLLM.

To this end, we propose EvoComp-2, a learning-based audio-video token compression framework for OmniLLMs. EvoComp-2 inserts a lightweight compressor between the modality-specific encoders and the LLM. The compressor takes the visual, audio, and text tokens as input and predicts token-level retention probabilities, together with query-dependent importance scores of the two modalities, to select a compact set of audio-video tokens. As illustrated in Figure 1, EvoComp-2 consists of two major stages. First, we construct task-oriented supervision labels offline for training the compressor. Second, we train the compressor on the resulting supervision. Once trained, only a single compressor forward pass is required to perform audio-video token compression. Since neither the modality-specific encoders nor the LLM needs to be modified or fine-tuned, EvoComp-2 supports plug-and-play integration with existing OmniLLMs.

3.2 TOPIC-GUIDED EVOLUTIONARY LABELING

Training the compressor requires reliable token-level supervision. We therefore extend the evolutionary labeling strategy of EvoComp (Song et al., 2026) to construct task-oriented labels from the existing OmniLLM training data (Geng et al., 2025; Cheng et al., 2025). EvoComp formulates visual token labeling for MLLMs as a search problem over binary visual-token retention masks. Each candidate mask is evaluated by retaining the selected visual tokens together with all text tokens and measuring the resulting task loss of the target MLLM. The loss is used as the fitness of the candidate, such that token selection is directly optimized toward model outputs. To reduce semantic redundancy and constrain the search space, EvoComp assigns each visual token to its nearest LLM vocabulary embedding according to cosine similarity, and groups tokens associated with the same vocabulary embedding. The evolutionary search is then constrained to select one token per group.

Directly extending this strategy to audio-video inputs, however, ignores their strong temporal structure. Tokens from nearby frames are often highly correlated, while tokens separated by major semantic transitions may describe entirely different events (Shen et al., 2025). Performing semantic grouping over the entire video may therefore fail to explicitly model temporal redundancy. We therefore introduce topic-guided temporal clustering to partition synchronized audio-video streams into semantically coherent temporal intervals before performing semantic grouping within each interval.

3.2.1 TOPIC-GUIDED TEMPORAL CLUSTERING

Since visual tokens exhibit meaningful semantic correspondence with the LLM vocabulary embeddings $\boldsymbol{E} = \{\boldsymbol{e}_i\}_{i=1}^c$, each token can be naturally represented as a distribution over vocabulary concepts, which can be further aggregated into a frame-level vocabulary distribution. This representation closely parallels the vocabulary distribution of a document in topic modeling, motivating us to adapt contextualized neural topic modeling (Bianchi et al., 2021) from natural language documents to video frames. Specifically, based on the cosine similarity $S_{ij}^t = \frac{\boldsymbol{v}_i^t \cdot \boldsymbol{e}_j}{\|\boldsymbol{v}_i^t\|_2 \|\boldsymbol{e}_j\|_2}$ between $\boldsymbol{v}_i^t$ and $\boldsymbol{e}_j$, we first compute a distribution $\boldsymbol{d}_i^t = \mathrm{softmax}(\{\tau \cdot S_{ij}^t\}_{j=1}^c)$ of $\boldsymbol{v}_i^t$ over $\boldsymbol{E}$, where $\tau > 1$ is a sharpness coefficient that suppresses weakly related vocabulary entries. We then average the distributions of all visual tokens in the frame $\boldsymbol{V}^t = \{\boldsymbol{v}_i^t\}_{i=1}^n$ to obtain its frame-level vocabulary distribution

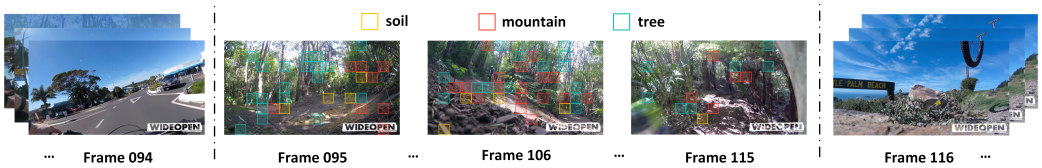


Figure 2: Visualization of a topic-guided temporal clustering and semantic grouping result. Three consecutive temporal intervals from a video are shown, separated by dashed lines. Within the middle interval, three example semantic groups associated with soil, mountain, and tree are highlighted.

$\boldsymbol{d}^t = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}_i^t$. Meanwhile, the visual tokens are averaged to obtain a contextualized frame representation $\boldsymbol{v}^t = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{v}_i^t$. We collect $(\boldsymbol{d}^t, \boldsymbol{v}^t)$ from all video frames in the training set and use them to train a Combined Topic Model (CombinedTM) (Bianchi et al., 2021). Analogous to the original CombinedTM, which combines a bag-of-words representation with a contextualized document embedding, our model takes the frame-level vocabulary distribution and contextualized frame representation as its two inputs. The model architecture and more training details can be found in Appendix A. After training, the topic model TM_{ϑ} maps each frame to the parameters of a Gaussian latent posterior $(\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2) = \mathrm{TM}_{\vartheta}(\boldsymbol{d}^t, \boldsymbol{v}^t)$, or equivalently, $q_{\vartheta}(\boldsymbol{z}^t \mid \boldsymbol{d}^t, \boldsymbol{v}^t) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t = \mathrm{diag}(\boldsymbol{\sigma}_t^2))$, where the latent variable $\boldsymbol{z}^t$ characterizes the topic composition of the t-th frame. Frames with similar semantic content are expected to exhibit similar latent topic distributions.

We therefore measure the semantic transition between two adjacent frames using the Bhattacharyya distance (Bhattacharyya, 1943). Specifically, for frames $\boldsymbol{V}^t$ and $\boldsymbol{V}^{t+1}$, the distance is computed as

$$D(\boldsymbol{V}^t, \boldsymbol{V}^{t+1}) = \frac{1}{8}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t+1})^{\mathrm{T}}\left(\frac{\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}_{t+1}}{2}\right)^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_{t+1}) + \frac{1}{2}\ln\left(\frac{\det\left(\frac{\boldsymbol{\Sigma}_t + \boldsymbol{\Sigma}_{t+1}}{2}\right)}{\sqrt{\det \boldsymbol{\Sigma}_t \cdot \det \boldsymbol{\Sigma}_{t+1}}}\right). \quad (1)$$

For each video, we compute the Bhattacharyya distances between all adjacent frame pairs and a new temporal boundary is set if the distance is larger than the ρ-th percentile of these distances:

$$D(\boldsymbol{V}^t, \boldsymbol{V}^{t+1}) > \mathrm{Percentile}_{\rho}(\{D(\boldsymbol{V}^s, \boldsymbol{V}^{s+1})\}_{s=1}^{K-1}). \quad (2)$$

This partitions the video into temporal intervals, such that frames with similar latent topics tend to remain within the same interval. Compared with frame-feature-based segmentation (Jin et al., 2024; Shen et al., 2025), our topic-guided representation is less sensitive to superficial visual changes and more responsive to semantic transitions. We transfer the video-derived temporal boundaries to the synchronized audio stream and perform semantic grouping separately for visual and audio tokens within each interval. Specifically, two visual (audio) tokens $\boldsymbol{v}_i^t$ ($\boldsymbol{a}_i^t$) and $\boldsymbol{v}_k^{t'}$ ($\boldsymbol{a}_k^{t'}$) in the same temporal interval are assigned to the same semantic group if and only if

$$\arg\max_j \frac{\boldsymbol{v}_i^t \cdot \boldsymbol{e}_j}{\|\boldsymbol{v}_i^t\|_2 \|\boldsymbol{e}_j\|_2} = \arg\max_j \frac{\boldsymbol{v}_k^{t'} \cdot \boldsymbol{e}_j}{\|\boldsymbol{v}_k^{t'}\|_2 \|\boldsymbol{e}_j\|_2} \quad (\arg\max_j \frac{\boldsymbol{a}_i^t \cdot \boldsymbol{e}_j}{\|\boldsymbol{a}_i^t\|_2 \|\boldsymbol{e}_j\|_2} = \arg\max_j \frac{\boldsymbol{a}_k^{t'} \cdot \boldsymbol{e}_j}{\|\boldsymbol{a}_k^{t'}\|_2 \|\boldsymbol{e}_j\|_2}). \quad (3)$$

This partitions the visual and audio tokens into disjoint groups. As illustrated in Figure 2, topic-guided temporal clustering separates frames with distinct semantic content into different temporal intervals, while semantic grouping clusters vocabulary-level related tokens within each interval.

3.2.2 Evolutionary Search for Token-Level Labels

The evolutionary search is constrained to select exactly one token from each group and the objective is to find a global binary mask $\boldsymbol{m} = [\boldsymbol{m}^v, \boldsymbol{m}^a]$ over all audio-video tokens, where $\boldsymbol{m}^v = [\boldsymbol{m}^{v1}, \boldsymbol{m}^{v2}, \cdots, \boldsymbol{m}^{vs_v}]$, $\boldsymbol{m}^a = [\boldsymbol{m}^{a1}, \boldsymbol{m}^{a2}, \cdots, \boldsymbol{m}^{as_a}]$, $\boldsymbol{m}^{vi}$ and $\boldsymbol{m}^{aj}$ are sub-masks over the i-th group of visual tokens and j-th group of audio tokens, respectively, so that the resulting loss $\mathcal{L}(\boldsymbol{m})$ computed on the target response is minimized when the LLM receives the selected audio-video tokens together with all input text tokens. Following EvoComp, we initialize a population of q candidate masks $\boldsymbol{M} = \{\boldsymbol{m}_i\}_{i=1}^{q}$ and evaluate their losses $\{\mathcal{L}(\boldsymbol{m}_i)\}_{i=1}^{q}$ in parallel. The top-p candidates with the lowest losses are selected as parents $\boldsymbol{P} = \{\boldsymbol{m}_{i_k} = [\boldsymbol{m}_{i_k}^v, \boldsymbol{m}_{i_k}^a]\}_{k=1}^{p}$ for producing the next generation. New candidates are generated using two genetic operators:

Crossover: For a selected parent $\boldsymbol{m}_{i_k} = [\boldsymbol{m}_{i_k}^v, \boldsymbol{m}_{i_k}^a]$, crossover is independently performed for $\boldsymbol{m}_{i_k}^v$ and $\boldsymbol{m}_{i_k}^a$ with probability 0.9. If crossover is triggered for $\boldsymbol{m}_{i_k}^v$, a second parent $\boldsymbol{m}_{i_j} = [\boldsymbol{m}_{i_j}^v, \boldsymbol{m}_{i_j}^a]$ is sampled, and a new sub-mask $\boldsymbol{m}'^v$ is generated by taking the first half of sub-masks from $\boldsymbol{m}_{i_k}^v$

and the remaining half from $\boldsymbol{m}^v_{i_j}$. The same process is performed for $\boldsymbol{m}^a_{i_k}$ to obtain a new sub-mask $\boldsymbol{m}'^a$. These sub-masks are combined to form a new child global mask $\boldsymbol{m}' = [\boldsymbol{m}'^v, \boldsymbol{m}'^a]$.

Mutation: Given a mask $\boldsymbol{m}_{i_k} = [\boldsymbol{m}^v_{i_k}, \boldsymbol{m}^a_{i_k}]$, each sub-mask $\boldsymbol{m}^{vj}_{i_k}$ (or $\boldsymbol{m}^{aj}_{i_k}$) is independently mutated with probability 0.2. When mutation is activated, the 1 in $\boldsymbol{m}^{vj}_{i_k}$ (or $\boldsymbol{m}^{aj}_{i_k}$) is randomly shifted to a neighboring position.

The newly generated q candidates are evaluated using the same target OmniLLM loss, and the best p candidates from the parents and offspring are retained as new parents for the next generation. After L iterations, the candidate with the lowest loss is selected as the token-level supervision label.

3.2.3 Modality-Level Importance Labeling

Since the relative contribution of the two modalities varies substantially across questions, we construct modality-level supervision to characterize query-dependent modality importance. For each training sample, we alternately set all visual tokens $\mathcal{V}$ and audio tokens $\mathcal{A}$ to zero and compute the LLM output loss $\mathcal{L}(\mathcal{V} = \mathbf{0}, \mathcal{A}, \boldsymbol{T})$ and $\mathcal{L}(\mathcal{V}, \mathcal{A} = \mathbf{0}, \boldsymbol{T})$ on the target response. A larger loss indicates greater importance of the suppressed modality. We zero the tokens rather than removing them to avoid confounding modality importance with the effect of token count, since visual tokens typically far outnumber audio tokens. We then define the modality-level importance label as

$$y^v = \begin{cases} 1, & \text{if } \mathcal{L}(\mathcal{V} = \mathbf{0}, \mathcal{A}, \boldsymbol{T}) > \mathcal{L}(\mathcal{V}, \mathcal{A} = \mathbf{0}, \boldsymbol{T}), \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Algorithm 1 in Appendix B summarizes the construction of both token- and modality-level labels.

3.3 Compressor Training and Token Compression

After constructing the supervision data, we train the compressor to predict both the token-level evolutionary decisions and the modality-level importance judgments. Since only a small fraction of audio-video tokens are retained after compression, the token classification problem exhibits severe class and difficulty imbalance (Song et al., 2026): there are substantially more negative than positive labels, and many redundant tokens are trivially classified, whereas a small set of ambiguous tokens can be exceptionally difficult. Moreover, informative tokens are often concentrated in only a few temporal units, and modality importance varies across queries. We therefore employ three objectives: a GHM loss, a temporal-window reweighting loss, and a modality-level importance loss.

GHM Loss. Following EvoComp, we adopt the GHM loss (Li et al., 2019) for token-level classification to address class and difficulty imbalance. The encoder-only transformer $f_{\boldsymbol{\varphi}}$ takes $(\mathcal{V}, \mathcal{A}, \boldsymbol{T})$ as input and produces representations $\{\mathcal{H}^v, \mathcal{H}^a, \boldsymbol{H}\} = f_{\boldsymbol{\varphi}}(\mathcal{V}, \mathcal{A}, \boldsymbol{T})$, from which the token-level classifier $g_{\boldsymbol{\psi}}$ predicts retention probabilities $\{p_i\}_{i=1}^{(n+m)K} = g_{\boldsymbol{\psi}}(\mathcal{H}^v, \mathcal{H}^a)$. Given the evolutionary label $y_i \in \{0, 1\}$, the gradient norm g_i is defined as $|p_i - y_i|$. GHM estimates the gradient density $GD(g_i)$ that reflects the concentration of tokens whose gradient norm is close to g_i, and assigns the binary cross-entropy loss $\ell(p_i, y_i)$ of each token a weight inversely proportional to this density:

$$\mathcal{L}_{\text{GHM-C}}(\boldsymbol{\varphi}, \boldsymbol{\psi}) = \sum_{i=1}^{(n+m)K} \frac{1}{GD(g_i)} \ell(p_i, y_i). \tag{5}$$

This reweighting reduces the dominance of densely populated easy samples while preventing a small number of extremely hard samples from overwhelming the optimization.

Temporal-Window Reweighting Loss. Since informative tokens are often concentrated in a few temporal units, we assign larger training weights to units containing more positively labeled tokens. Specifically, for the t-th temporal unit, we compute its importance score from the number of retained-token labels $n^{t+} = \sum_{i \in \mathcal{I}^t} y_i$, where $\mathcal{I}^t = \{i \mid y_i \in t\text{-th temporal unit}\}$. It is converted into a normalized temporal weight $w^t = \alpha + (1 - \alpha) \cdot n^{t+}/|\mathcal{I}^t|$, where α is set to 0.1 by default to provide a minimum weight for units without any positive labels. The temporal-window reweighting loss is then computed as

$$\mathcal{L}_{\text{TWR}}(\boldsymbol{\varphi}, \boldsymbol{\psi}) = \frac{1}{(n+m)K} \sum_{t=1}^{K} w^t \sum_{i \in \mathcal{I}^t} \ell(p_i, y_i). \tag{6}$$

In this way, the compressor is encouraged to focus more on information-rich and task-critical temporal regions rather than being dominated by the large number of uninformative units.

Table 1: Comparison of token compression methods on WorldSense with Qwen2.5-Omni-7B under different audio-video token retention ratios.

Method	Retained Ratio	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.
Full Tokens	100%	52.4	50.1	48.5	44.6	43.8	41.6	41.6	47.3	46.8
Random	55%	47.1	47.0	44.4	41.2	40.0	40.1	40.1	46.3	43.6
FastV	50%	48.8	47.4	44.2	44.1	41.2	38.3	40.0	46.6	44.3
DyCoke (V&A)	50%	48.4	49.9	46.7	41.4	39.9	40.8	40.2	46.5	44.6
OmniZip	45%	49.4	51.1	45.6	43.9	40.1	40.8	41.9	46.7	45.9
OmniSelect	45%	45.9	49.2	45.0	42.0	44.6	40.8	**43.3**	**49.0**	45.1
OmniZip	35%	48.3	49.5	**47.6**	42.5	40.1	40.2	42.3	46.3	45.3
OmniSelect	35%	43.7	47.9	43.8	42.0	43.1	**42.9**	**43.3**	46.3	44.1
EvoComp-2	35%	**49.6**	**51.5**	46.4	**44.6**	**46.1**	39.1	41.9	48.0	**46.2**

Table 2: Performance of EvoComp-2 and other methods on audio-video understanding and reasoning tasks using Qwen2.5-Omni-7B under different token retention ratios.

Method	Retained Ratio	ShortVid-Bench	AVUT							DailyOmni						
			EL	OR	OM	IE	CC	CM	Avg.	Con.	Event	AV Event	Com.	Inf.	Rea.	Avg.
Full Tokens	100%	70.5	38.2	67.8	59.6	85.6	44.1	66.7	64.5	57.0	60.1	50.8	71.8	78.6	78.9	64.2
Random	55%	68.3	38.2	64.9	55.6	80.1	34.7	65.0	61.0	52.3	49.7	41.5	71.0	72.7	71.4	56.9
FastV	50%	68.0	34.1	64.3	57.1	77.6	36.4	56.4	58.4	53.9	54.6	44.1	66.4	76.6	76.6	59.7
DyCoke (V&A)	50%	68.5	38.8	67.2	**58.2**	81.9	39.0	62.4	62.0	50.3	52.9	45.0	70.2	77.9	76.6	59.5
OmniZip	45%	69.9	38.4	67.2	56.9	**85.3**	42.4	66.0	63.0	53.9	52.0	46.2	70.2	77.3	76.6	60.0
OmniSelect	45%	69.9	34.9	**74.0**	**58.2**	77.0	40.7	64.8	62.6	53.4	53.9	46.6	**72.5**	76.6	76.6	60.7
EvoComp-2	45%	**70.8**	**40.6**	72.4	**58.2**	**85.3**	**44.9**	**68.1**	**65.6**	**54.4**	**55.6**	**50.0**	66.4	**79.9**	**77.1**	**61.7**
OmniZip	30%	**70.2**	**36.5**	69.5	54.1	84.4	40.7	64.0	62.3	47.2	47.4	43.7	63.4	76.6	73.7	56.0
OmniSelect	30%	69.8	**36.5**	68.1	55.4	71.2	35.6	62.3	59.1	48.2	50.0	**46.2**	67.2	74.7	73.1	57.4
EvoComp-2	30%	69.3	**36.5**	**69.8**	**57.4**	**85.0**	**42.4**	**66.7**	**64.0**	**52.3**	**52.6**	45.0	**69.5**	**77.3**	**75.4**	**59.4**

Modality-Level Importance Loss. In addition to predicting individual token importance, we train a modality-level classifier $g_{\boldsymbol{\psi}'}$ to estimate the query-dependent modality importance. We average the visual and audio representations in $\mathcal{H}^v$ and $\mathcal{H}^a$ to obtain modality representations $\boldsymbol{h}^v$ and $\boldsymbol{h}^a$. The classifier $g_{\boldsymbol{\psi}'}$ takes $(\boldsymbol{h}^v, \boldsymbol{h}^a)$ as input and predicts their importance probabilities $\{p^v, p^a\} = g_{\boldsymbol{\psi}'}(\boldsymbol{h}^v, \boldsymbol{h}^a)$. Given the modality-level label y^v, the modality-level importance loss is computed as

$$\mathcal{L}_{\textbf{MLI}}(\boldsymbol{\varphi}, \boldsymbol{\psi}') = \frac{1}{2}(\ell(p^v, y^v) + \ell(p^a, 1 - y^v)). \tag{7}$$

This objective encourages the classifier to estimate the overall contribution of each modality.

During the training, we first optimize $f_{\boldsymbol{\varphi}}$ and $g_{\boldsymbol{\psi}}$ with $\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{\psi}) = \mathcal{L}_{\textbf{GHM-C}}(\boldsymbol{\varphi}, \boldsymbol{\psi}) + \beta_1 \mathcal{L}_{\textbf{TWR}}(\boldsymbol{\varphi}, \boldsymbol{\psi})$, where β_1 balances the two terms. This allows the compressor to learn reliable token-level representations and retention probabilities. We then freeze them and train $g_{\boldsymbol{\psi}'}$ using $\mathcal{L}_{\textbf{MLI}}(\boldsymbol{\varphi}, \boldsymbol{\psi}')$. This two-stage training decouples modality-level objective from token-level compression learning, while allowing $g_{\boldsymbol{\psi}'}$ to exploit the learned audio-video representations for modality importance estimation.

Inference-Time Token Compression. At inference, the compressor predicts the token-level retention probabilities and modality importance probabilities in a single forward pass. We combine the two predictions as $p_i^r = p_i \cdot p^v$ and $p_i^r = p_i \cdot p^a$ for visual and audio tokens, respectively. Given a target retention ratio r, we jointly rank all visual and audio tokens by p_i^r and retain the top-r fraction. The selected audio-video tokens, together with all text tokens, are then forwarded to the LLM for subsequent reasoning and generation.

4 EXPERIMENTS

In this section, we conduct extensive experiments to evaluate EvoComp-2 from four aspects. First, we compare EvoComp-2 with existing token compression methods on various audio-video tasks. Second, we examine the cross-model transferability of the compressor. Third, we evaluate practical inference efficiency by comparing the prefill latency and task performance of EvoComp-2 with other methods under different compression ratios. Finally, we conduct ablation studies to investigate the effectiveness of topic-guided temporal clustering and the proposed training objectives.

Table 3: Cross-model transfer performance of EvoComp-2 on Qwen2.5-Omni-3B using the compressor trained with Qwen2.5-Omni-7B under different token retention ratios.

Method	Retained Ratio	ShortVid-Bench	DailyOmni						
			Con.	Event	AV Event	Com.	Inf.	Rea.	Avg.
Full Tokens	100%	69.4	55.4	56.2	53.8	71.0	79.2	74.3	62.8
Random	55%	67.9	47.7	44.4	45.4	64.1	71.4	63.4	53.6
FastV	50%	68.0	**51.8**	51.6	49.6	65.7	75.3	**71.4**	58.7
DyCoke (V&A)	50%	67.4	48.2	49.4	47.5	65.7	72.7	65.1	55.9
OmniZip	45%	68.5	47.2	47.7	46.6	64.1	**76.6**	69.7	56.1
OmniSelect	45%	68.1	50.8	49.7	**52.1**	65.7	74.0	69.1	58.1
EvoComp-2	45%	**70.0**	51.3	**53.3**	50.8	**67.9**	74.0	69.1	**59.1**
OmniZip	30%	**68.9**	46.6	46.1	42.9	59.5	68.2	**68.0**	53.1
OmniSelect	30%	66.8	46.6	45.4	**48.3**	62.6	**72.1**	65.1	54.4
EvoComp-2	30%	68.6	**50.8**	**47.1**	43.3	**67.2**	70.1	65.1	**54.7**

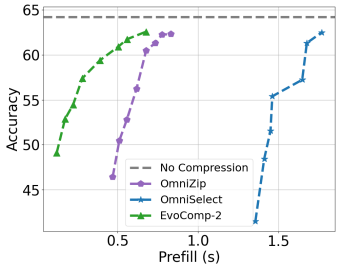


Figure 3: Accuracy–latency of Qwen2.5-Omni-7B on DailyOmni under different token retention ratios.

4.1 Experimental Setup

Models and Training Data. We conduct experiments on Qwen2.5-Omni-7B/3B (Xu et al., 2025a). Qwen2.5-Omni-7B serves as the primary target OmniLLM for label construction and compressor training, while Qwen2.5-Omni-3B is used to evaluate the cross-model transferability of the learned compressor. The compressor consists of a single-layer transformer followed by token- and modality-level linear classifiers. Following EvoComp (Song et al., 2026), the transformer is initialized from the first LLM layer in the target OmniLLM. We replace its causal attention with bidirectional attention and adopt a residual connection between its input and output. To construct the training supervision, we use a subset of LongVALE (Geng et al., 2025) together with the Video-Holmes dataset (Cheng et al., 2025). More implementation details are provided in Appendix C.

Baselines and Evaluation Benchmarks. We compare EvoComp-2 with Random Selection, attention-based FastV (Chen et al., 2024), temporal-redundancy-based DyCoke (Tao et al., 2025), audio-guided OmniZip (Tao et al., 2026), and query-aware OmniSelect (Yang et al., 2026). We evaluate these methods on WorldSense (Hong et al., 2026), ShortVid-Bench (Ge et al., 2025), AVUT (Yang et al., 2025b), and DailyOmni (Zhou et al., 2025), covering diverse audio-video understanding and reasoning tasks. More details can be found in Appendix D.

4.2 Main Results

Compression Performance. We first compare EvoComp-2 with existing token compression methods using Qwen2.5-Omni-7B under different audio-video token retention ratios. As shown in Table 1, EvoComp-2 preserves 98.7% of the uncompressed performance on WorldSense with only 35% of audio-video tokens retained. Notably, this result is even better than those of OmniZip and OmniSelect when retaining 45% of the audio-video tokens, indicating that EvoComp-2 can preserve more task-relevant information under substantially more aggressive compression. Table 2 further reports the results on ShortVid-Bench, AVUT, and DailyOmni at token retention ratios of 45% and 30%. EvoComp-2 achieves the best performance in most tasks. The qualitative visualizations in Appendix E illustrate how EvoComp-2 adjusts audio-video token selection according to query-dependent modality importance. Overall, these results show that EvoComp-2 can retain task-critical audio-video information more effectively than existing heuristic compression strategies.

Cross-Model Transferability. We next investigate whether the compressor learned for one OmniLLM can be directly transferred to another model without additional training. Specifically, we apply the compressor trained with Qwen2.5-Omni-7B directly to Qwen2.5-Omni-3B. Since the audio-video token representations of the two models have different feature dimensions, we employ adaptive average pooling to align the input representations with the dimensionality expected by the trained compressor. Table 3 reports the transfer results on ShortVid-Bench and DailyOmni. Without any additional training on Qwen2.5-Omni-3B, EvoComp-2 achieves strong performance at both 45% and 30% token retention ratios and remains competitive with or superior to the compared compression methods. These results suggest that the learned token-selection patterns are not tightly coupled to a specific target model and can generalize across OmniLLMs of different scales. Such cross-model transferability substantially improves the flexibility of EvoComp-2, as a trained compressor can potentially be reused for related OmniLLMs without repeatedly training.

Inference Efficiency. Finally, we evaluate the practical inference efficiency of EvoComp-2 by jointly considering compression overhead, LLM prefill latency, and task accuracy. We compare EvoComp-2 with OmniZip and OmniSelect under different token retention ratios, with all experiments conducted on an NVIDIA A100-80GB GPU. Figure 3 presents the trade-off between accuracy and LLM prefill (including compression overhead) on DailyOmni. EvoComp-2 consistently achieves higher accuracy under comparable prefill latency, demonstrating a more favorable accuracy-efficiency trade-off than OmniZip and OmniSelect. Without token compression, the LLM prefill latency is 1.20 s and the total inference latency is 2.17 s. With EvoComp-2, these are reduced to 0.39 s and 1.36 s under 3.3x token compression, respectively, corresponding to a 3.1x prefill speedup and a 1.6x end-to-end speedup, while retaining 92.5% of the original task performance. Under a milder 1.8x token compression, EvoComp-2 preserves 97.5% of the original accuracy while achieving a 1.3x end-to-end speedup. These results show that the lightweight compressor overhead is substantially outweighed by the reduction in LLM computation, yielding meaningful end-to-end acceleration across different compression levels.

4.3 Ablation Studies

Effectiveness of Topic-Guided Temporal Clustering. To evaluate the effectiveness of the proposed topic-guided temporal clustering strategy, we compare EvoComp-2 with two alternative schemes: 1) *EvoComp-2 w/o TGTC*: removing topic-guided temporal clustering and separately performing semantic grouping over all visual and audio tokens before evolutionary search; 2) *EvoComp-2 w/ DPC-KNN*: applying DPC-KNN (Du et al., 2016) independently to cluster the visual and audio tokens and retaining the resulting cluster centers. As shown in Table 4, the proposed topic-guided temporal clustering consistently achieves the best performance, while the DPC-KNN variant performs the worst. These results indicate that explicitly modeling the temporal semantic structure of audio-video sequences is important for constructing high-quality supervision. Compared with global semantic grouping, our method avoids grouping semantically similar tokens across distinct temporal events, while also preserving the temporal organization of the video.

Effectiveness of Training Objectives. We further examine the contribution of the three training objectives used in EvoComp-2. We consider three variants: 1) *EvoComp-2 w/* $\mathcal{L}_{CE} + \mathcal{L}_{TWR} + \mathcal{L}_{MLI}$: replacing the GHM loss with cross-entropy loss; 2) *EvoComp-2 w/* $\mathcal{L}_{GHM\text{-}C} + \mathcal{L}_{MLI}$: removing the temporal-window reweighting loss; 3) *EvoComp-2 w/* $\mathcal{L}_{GHM\text{-}C} + \mathcal{L}_{TWR}$: removing the modality-level importance loss. The results in Table 4 show that each component contributes to the final performance. Replacing $\mathcal{L}_{\text{GHM-C}}$ with $\mathcal{L}_{\text{CE}}$ consistently degrades accuracy, confirming the importance of addressing class and difficulty imbalance in token retention prediction. Removing either the $\mathcal{L}_{\text{TWR}}$ or the $\mathcal{L}_{\text{MLI}}$ also leads to performance degradation, demonstrating the benefits of explicitly modeling temporal saliency and modality preference. In particular, removing $\mathcal{L}_{\text{TWR}}$ causes a larger accuracy drop, indicating that emphasizing information-rich temporal regions is especially important for training. The three objectives provide complementary supervision for learning token-level, temporal, and modality-level importance.

Table 4: Ablation study on different label construction strategies and training objectives using Qwen2.5-Omni-7B.

Method	DailyOmni	
Retained Ratio	45%	30%
EvoComp-2 w/o TGTC	55.0	50.3
EvoComp-2 w/ DPC-KNN	52.1	47.1
EvoComp-2 w/ $\mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{TWR}} + \mathcal{L}_{\text{MLI}}$	53.0	49.3
EvoComp-2 w/ $\mathcal{L}_{\text{GHM-C}} + \mathcal{L}_{\text{MLI}}$	55.2	51.9
EvoComp-2 w/ $\mathcal{L}_{\text{GHM-C}} + \mathcal{L}_{\text{TWR}}$	61.4	58.8
EvoComp-2	**61.7**	**59.4**

5 Conclusion

In this work, we present EvoComp-2, an audio-video token compression framework for OmniLLMs. EvoComp-2 introduces topic-guided temporal clustering and semantic grouping to construct structured search spaces, and leverages evolutionary search guided by the output loss of the target OmniLLM to generate task-oriented token-level supervision. We further model query-dependent modality importance and train a lightweight compressor with GHM loss, temporal-window reweighting, and modality-level importance supervision. Extensive experiments demonstrate that EvoComp-2 achieves favorable accuracy-efficiency trade-offs, strong cross-model transferability, and substantial inference acceleration. Overall, EvoComp-2 provides an effective and flexible solution for efficient audio-video compression in OmniLLMs.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to develop theoretical models or conceptual frameworks, clean or reformat datasets. Other potential uses, including generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, and proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools to summarize existing literature, identify relevant literature, draft parts of the paper, and search for information. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Saurabhchand Bhati, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Towards audio token compression in large audio language models. *arXiv preprint arXiv:2511.20973*, 2025.

Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distribution. *Bulletin of the Calcutta Mathematical Society*, 35:99–110, 1943.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Short Papers*, pp. 759–766, 2021.

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, 2024.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=aPnQpmwHW7.

Mingjing Du, Shifei Ding, and Hongjie Jia. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems*, 99:135–145, 2016.

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. Long-VALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18959–18969. IEEE, 2025.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Siddharth Gururani, Hanrong Ye, Pritam Biswas, Yuanhang Su, Ehsan Hosseini-Asl, et al. Audio-Visual Flamingo: Open audio-visual intelligence for long and complex videos. *arXiv preprint arXiv:2607.16107*, 2026.

Peize He, Yaodi Luo, Xiaoqian Liu, Xuyang Liu, Jiahang Deng, Yaosong Du, Bangyu Li, Xiyan Gui, Yuxuan Chen, and Linfeng Zhang. HeadRouter: Dynamic head-weight routing for task-adaptive audio token pruning in large audio language models. *arXiv preprint arXiv:2604.23717*, 2026.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In *International Conference on Learning Representations*, 2026.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. arXiv:2412.16117.

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. Multi-granular spatio-temporal token merging for training-free acceleration of video llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23990–24000. IEEE, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13700–13710. IEEE, 2024.

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

Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8577–8584, 2019.

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.

Jiale Luo, Xiaoyu Liang, and Haoji Hu. Locality matters for training-free audio token compression in audio-language models. *arXiv preprint arXiv:2605.25179*, 2026.

Kele Shao, Keda TAO, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. A survey of token compression for efficient multimodal large language models. *Transactions on Machine Learning Research*, 2026.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. FastVID: Dynamic density pruning for fast video large language models. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2503.11187.

Jiafei Song, Fengwei Zhou, Jin Qu, Wenjin Jason Li, Tong Wu, Gengjian Xue, Zhikang Zhao, Daomin Wei, Yichao Lu, and Bailin Na. EvoComp: Learning visual token compression for multimodal large language models via semantic-guided evolutionary labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3532–3542, 2026.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=BybtVK9lg`.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-SALMONN: Speech-enhanced audio-visual large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47198–47217. PMLR, 21–27 Jul 2024.

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

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

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Jian Liu, and Huan Wang. OmniZip: Audio-guided dynamic token compression for fast omnimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17682–17692, 2026.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025.

Bajian Xiang, Tingwei Guo, Xuan Chen, and Yang Han. Do we need distinct representations for every speech token? unveiling and exploiting redundancy in large speech language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15069–15087, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.742. URL https://aclanthology.org/2026.findings-acl.742/.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Zhenghao Xing, Ruiyang Xu, Yuxuan Wang, Jinzheng He, Ziyang Ma, Qize Yang, Yunfei Chu, Jin Xu, Junyang Lin, Chi-Wing Fu, and Pheng-Ann Heng. Native active perception as reasoning for omni-modal understanding. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3dLPVNhFZh.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

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

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

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

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Audio-centric video understanding benchmark without text shortcut. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6569–6587, 2025b.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857, 2025.

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.

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

APPENDIX

A MODEL ARCHITECTURE AND TRAINING DETAILS OF COMBINED TOPIC MODEL

As introduced in Section 3.2.1, we adapt CombinedTM (Bianchi et al., 2021) to model the latent semantic topics of video frames. For each frame, the model takes as input the frame-level vocabulary distribution $\boldsymbol{d}^t$ and the contextualized frame representation $\boldsymbol{v}^t$. The two inputs are first independently projected to 900-dimensional representations through separate fully connected layers. The projected features are concatenated and passed through a Softplus activation, followed by a fully connected layer that reduces the feature dimension to 900 and another Softplus activation. We then apply four additional fully connected layers, each preserving the 900-dimensional hidden representation and followed by a Softplus activation. Finally, two separate fully connected layers with Batch Normalization (BN) predict the posterior mean $\boldsymbol{\mu}_t$ and variance $\boldsymbol{\sigma}_t^2$ (after taking the exponential function). Both outputs have a dimension of $R = 2000$, corresponding to the number of latent topics. This follows the variational inference formulation of CombinedTM, which augments the ProdLDA (Srivastava & Sutton, 2017) inference network with contextualized representations.

Following CombinedTM and ProdLDA, we train the topic model by minimizing the negative evidence lower bound, consisting of a KL-divergence term and a reconstruction term. The KL-divergence term regularizes the approximate posterior toward the topic Gaussian prior. The approximate posterior for frame t is defined as $q_{\vartheta}(\boldsymbol{z}^t \mid \boldsymbol{d}^t, \boldsymbol{v}^t) = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t = \mathrm{diag}(\boldsymbol{\sigma}_t^2))$. The topic Gaussian prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0 = \mathrm{diag}(\boldsymbol{\sigma}_0^2))$ is parameterized by mean $\boldsymbol{\mu}_0$ and variance $\boldsymbol{\sigma}_0^2$, which are initialized as $\boldsymbol{\mu}_0 = 0$ and $\boldsymbol{\sigma}_0^2 = 1 - 1/R$ respectively, following the default CombinedTM configuration. These prior parameters are optimized jointly with the inference network by the KL-divergence term

$$\mathcal{L}_{\mathbf{KL}} = \frac{1}{2}(\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_t) + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_t)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_t) - R + \ln(\frac{\det \boldsymbol{\Sigma}_0}{\det \boldsymbol{\Sigma}_t})). \tag{8}$$

To compute the reconstruction term, a latent topic representation is sampled according to $q_{\vartheta}(\boldsymbol{z}^t \mid \boldsymbol{d}^t, \boldsymbol{v}^t)$ using the reparameterization trick and converted into topic proportions through $\boldsymbol{\theta}^t = \mathrm{softmax}(\boldsymbol{z}^t)$, where $\boldsymbol{z}^t = \boldsymbol{\mu}_t + \boldsymbol{\sigma}_t \odot \boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. The topic proportions are then passed through the ProdLDA decoder to reconstruct the frame-level vocabulary distribution, producing $\hat{\boldsymbol{d}}^t = \mathrm{softmax}(\mathrm{BN}(\boldsymbol{\theta}^{t\mathrm{T}}\boldsymbol{W}))$, where the ProdLDA decoder is parameterized by a learnable topic-vocabulary matrix $\boldsymbol{W}$. Each row of $\boldsymbol{W}$ captures the association between one latent topic and the vocabulary dimensions. Thus, the reconstruction is determined by a weighted combination of the learned topic-vocabulary representations, where the weights are given by the frame-specific topic proportions $\boldsymbol{\theta}^t$. The reconstruction term encourages the latent topic representation to reconstruct the original frame-level vocabulary distribution:

$$\mathcal{L}_{\mathbf{rec}} = -\boldsymbol{d}^t \cdot \log \hat{\boldsymbol{d}}^t. \tag{9}$$

The final training objective is $\mathcal{L}_{\mathbf{TM}} = \mathcal{L}_{\mathbf{KL}} + \mathcal{L}_{\mathbf{rec}}$.

For training data, we use the same subset of LongVALE (Geng et al., 2025) and the Video-Holmes dataset (Cheng et al., 2025) employed for evolutionary label construction. We first use the visual encoder of the target OmniLLM to extract visual tokens for each sampled frame. The frame-level vocabulary distribution and contextualized frame representation are then computed following Section 3.2.1, with the sharpness coefficient τ set to 500, yielding paired inputs $(\boldsymbol{d}^t, \boldsymbol{v}^t)$ for topic-model training. A small portion of these pairs is held out as the validation set. We train CombinedTM on a single NVIDIA A100-80GB GPU with a batch size of 2048 and a learning rate of 0.002 using the Adam optimizer. Training runs for at most 30 epochs and is terminated early if the validation loss does not improve for three consecutive epochs. The trained topic model is subsequently fixed and used to obtain frame-level latent Gaussian representations for topic-guided temporal clustering.

B TOPIC-GUIDED EVOLUTIONARY LABELING ALGORITHM

Algorithm 1 summarizes the complete procedure of topic-guided evolutionary labeling. We first partition the synchronized audio-video tokens into semantically coherent temporal intervals using the trained topic model and perform semantic grouping within each interval. We then initialize

Algorithm 1 Topic-Guided Evolutionary Labeling

Input: Visual tokens $\mathcal{V}$; Audio tokens $\mathcal{A}$; Text tokens $\boldsymbol{T}$; Vocabulary embeddings $\boldsymbol{E}$; Pretrained LLM; Trained topic model TM_{ϑ}; Population size q; Number of parents p; Number of iterations L;
Output: Binary mask $\boldsymbol{m}^*$ indicating selected visual and audio tokens; Modality-level importance label y^v;

Partition $\mathcal{V}$ and $\mathcal{A}$ into temporal intervals according to Eq. (2);
Partition visual and audio tokens in each interval into subsets according to Eq. (3);
Randomly initialize a population of binary masks $\boldsymbol{M}^{(0)} = \{\boldsymbol{m}_i = [\boldsymbol{m}_i^v, \boldsymbol{m}_i^a]\}_{i=1}^q$, where $\boldsymbol{m}_i^v = [\boldsymbol{m}_i^{v1}, \boldsymbol{m}_i^{v2}, \cdots, \boldsymbol{m}_i^{vs_v}]$ and $\boldsymbol{m}_i^a = [\boldsymbol{m}_i^{a1}, \boldsymbol{m}_i^{a2}, \cdots, \boldsymbol{m}_i^{as_a}]$, such that each $\boldsymbol{m}_i^{vj}$ and $\boldsymbol{m}_i^{ak}$ is one-hot;
Evaluate losses $\{\mathcal{L}(\boldsymbol{m}_i)\}_{i=1}^q$ for $\boldsymbol{M}^{(0)}$ in parallel;
Select top-p candidates with lowest losses as parents $\boldsymbol{P}^{(0)}$;
for $1 \le l \le L$ **do**
  Initialize a new population $\boldsymbol{M}^{(l)} = \emptyset$;
  while $|\boldsymbol{M}^{(l)}| < q$ **do**
    Randomly sample a parent $\boldsymbol{m}_{\text{parent}}$ from $\boldsymbol{P}^{(l-1)}$;
    Generate $\boldsymbol{m}_{\text{crossover}}$ by performing crossover on $\boldsymbol{m}_{\text{parent}}$;
    Generate $\boldsymbol{m}_{\text{mutation}}$ by performing mutation on $\boldsymbol{m}_{\text{crossover}}$;
    if $\boldsymbol{m}_{\text{mutation}}$ is not in $\boldsymbol{M}^{(0)}$ **then**
      Add $\boldsymbol{m}_{\text{mutation}}$ to $\boldsymbol{M}^{(0)}$ and $\boldsymbol{M}^{(l)}$;
    end if
  end while
  Evaluate losses for $\boldsymbol{M}^{(l)}$ in parallel;
  Select top-p candidates from both $\boldsymbol{P}^{(l-1)}$ and $\boldsymbol{M}^{(l)}$ with lowest losses as parents $\boldsymbol{P}^{(l)}$;
end for
Evaluate losses $\mathcal{L}(\mathcal{V} = \boldsymbol{0}, \mathcal{A}, \boldsymbol{T})$ and $\mathcal{L}(\mathcal{V}, \mathcal{A} = \boldsymbol{0}, \boldsymbol{T})$ in parallel;
Determine modality-level importance label y^v according to Eq. (4);
return $\boldsymbol{m}^* = \arg\min_{\boldsymbol{m} \in \boldsymbol{P}^{(L)}} \mathcal{L}(\boldsymbol{m})$ and y^v

a population of binary retention masks and evaluate each candidate using the output loss of the target OmniLLM. The candidates with the lowest losses are iteratively evolved through crossover and mutation, while retaining the best-performing masks as parents for the next generation. After the evolutionary search, the mask with the lowest loss is selected as the token-level retention label. Finally, we separately suppress the visual and audio tokens and compare the resulting losses to construct the modality-level importance label.

C Implementation Details

C.1 Evolutionary Label Construction

We construct the labeled training data using a subset of LongVALE (Geng et al., 2025) and the Video-Holmes dataset (Cheng et al., 2025). LongVALE provides temporally localized descriptions for different segments of each video. We extract video segments with durations between 10 and 200 seconds and use the original Qwen2.5-Omni-7B (Xu et al., 2025a) to generate a description for each segment, which serves as the target response when evaluating candidate masks during evolutionary search. We do not directly use the descriptions provided by LongVALE because Qwen2.5-Omni-7B is not trained specifically on this dataset, and its output can differ substantially from the reference description even when all audio-video tokens are retained. Consequently, using the original annotations as targets would introduce a mismatch between the target response and the intrinsic generation behavior of the target OmniLLM, making the loss less reliable for measuring the quality of a candidate token subset. Using the response generated by the uncompressed target model instead provides a more consistent supervision signal for identifying tokens that preserve the model's original output. For the same reason, we also use the responses generated by the original Qwen2.5-Omni-7B as target responses for Video-Holmes.

For evolutionary label construction, we set the population size q to 32, the number of selected parents p to 8, and the number of evolutionary iterations L to 5. During topic-guided temporal clustering, a temporal boundary is introduced when the Bhattacharyya distance (Bhattacharyya, 1943) between two adjacent frames exceeds the 90th percentile of all adjacent-frame distances within the corresponding video. To further reduce the evolutionary search space, we rank the semantic groups of

each modality according to the number of tokens they contain and perform evolutionary optimization only over the top-ranked visual and audio semantic groups, with the number of optimized groups set to 35% of the total number of visual and audio tokens, respectively. The masks associated with all remaining groups are fixed to zero throughout the search. This design substantially reduces the number of variables explored by the evolutionary algorithm while forcing the search to identify the most informative tokens from highly redundant semantic groups, thereby providing the compressor with supervision for retaining the most critical audio-video tokens.

C.2 Compressor Training

Following EvoComp (Song et al., 2026), we initialize the encoder-only transformer $f_{\boldsymbol{\varphi}}$ with the parameters of the first LLM layer of Qwen2.5-Omni-7B. The training is conducted on two NVIDIA A100-80GB GPUs. As described in Section 3.3, we first train the encoder-only transformer $f_{\boldsymbol{\varphi}}$ and the token-level classifier $g_{\boldsymbol{\psi}}$ using the GHM loss $\mathcal{L}_{\textbf{GHM-C}}(\boldsymbol{\varphi}, \boldsymbol{\psi})$ and the temporal-window reweighting loss $\mathcal{L}_{\textbf{TWR}}(\boldsymbol{\varphi}, \boldsymbol{\psi})$. For the GHM loss, we adopt the unit-region approximation (Li et al., 2019) that partitions the range of gradient norms into discrete regions and estimates the gradient density from the number of samples falling into each region. We use 50 unit regions in our implementation. The coefficient β_1 for balancing the GHM loss and the temporal-window reweighting loss is set to 0.3. We train the model for 10 epochs with a batch size of 32 using the AdamW optimizer. The initial learning rate is set to 1×10^{-5} and decayed according to a linear learning-rate schedule. We then freeze the trained $f_{\boldsymbol{\varphi}}$ and $g_{\boldsymbol{\psi}}$ and optimize only the modality-level classifier $g_{\boldsymbol{\psi}'}$ using the modality-level importance loss $\mathcal{L}_{\textbf{MLI}}(\boldsymbol{\varphi}, \boldsymbol{\psi}')$. Since both classifiers predict importance from the contextualized audio-video representations, we initialize $g_{\boldsymbol{\psi}'}$ with the learned parameters of the token-level classifier $g_{\boldsymbol{\psi}}$ rather than training it from scratch. We train $g_{\boldsymbol{\psi}'}$ for one epoch with a batch size of 32 using AdamW, with the learning rate set to 1×10^{-6}. This training strategy enables the compressor to first learn robust token-level retention and then learn modality-level importance without interfering with the learned token representations.

D Evaluation Benchmarks

WorldSense. WorldSense (Hong et al., 2026) is a benchmark designed to evaluate real-world omnimodal understanding with strongly coupled audio and visual information. It contains 1,662 synchronized audio-video clips spanning 8 broad domains and 67 fine-grained categories, together with 3,172 multiple-choice questions covering 26 tasks from basic perception to high-level reasoning. A key characteristic of WorldSense is that its questions are explicitly designed to require joint audio-visual understanding, such that either modality alone is often insufficient to determine the correct answer. The benchmark therefore provides a comprehensive evaluation of whether an OmniLLM can preserve and integrate complementary information from both modalities.

ShortVid-Bench. ShortVid-Bench (Ge et al., 2025) focuses on the structured understanding of real-world user-generated short videos, which typically exhibit fast pacing, dense visual content, rich audio signals, and rapid semantic transitions. The released benchmark contains 1,000 human-curated multiple-choice samples and evaluates six dimensions: temporal reasoning and localization, affective intent classification, creator intent understanding, narrative comprehension, humor and meme interpretation, and creative innovation analysis. These tasks go beyond surface-level perception and require models to jointly exploit visual and audio cues to understand temporal structure, narrative, and high-level intent, making ShortVid-Bench particularly suitable for evaluating compression under information-dense short-video scenarios.

AVUT. AVUT (Yang et al., 2025b) is an audio-centric video understanding benchmark designed to evaluate both audio understanding and audio-visual interaction while mitigating the text-shortcut problem. We evaluate on the AV-Human partition of AVUT. It contains 698 videos and 1,734 human-annotated question-answer pairs, covering audio information extraction, audio content counting, audio event localization, audio-visual character matching, audio-visual object matching, and audio-visual text matching. The questions are designed to require genuine audio or audio-visual understanding rather than relying on textual cues alone, making AV-Human particularly suitable for evaluating whether token compression preserves critical acoustic information and cross-modal evidence.

DailyOmni. DailyOmni (Zhou et al., 2025) targets audio-visual reasoning with explicit cross-modal temporal alignment in everyday scenarios. It comprises 684 real-world videos and 1,197 multiple-choice questions, with 30-second and 60-second clips covering six task families: audio-visual event alignment, event sequence understanding, reasoning, inference, comparison, and context understanding. Its construction explicitly identifies temporally aligned audio and visual events and filters questions that can be answered from text alone, thereby placing particular emphasis on fine-grained temporal relationships across modalities. DailyOmni therefore provides a challenging evaluation of whether compressed audio-video representations retain the synchronized evidence required for cross-modal temporal reasoning.

E Visualization of Query-Dependent Audio-Video Token Selection

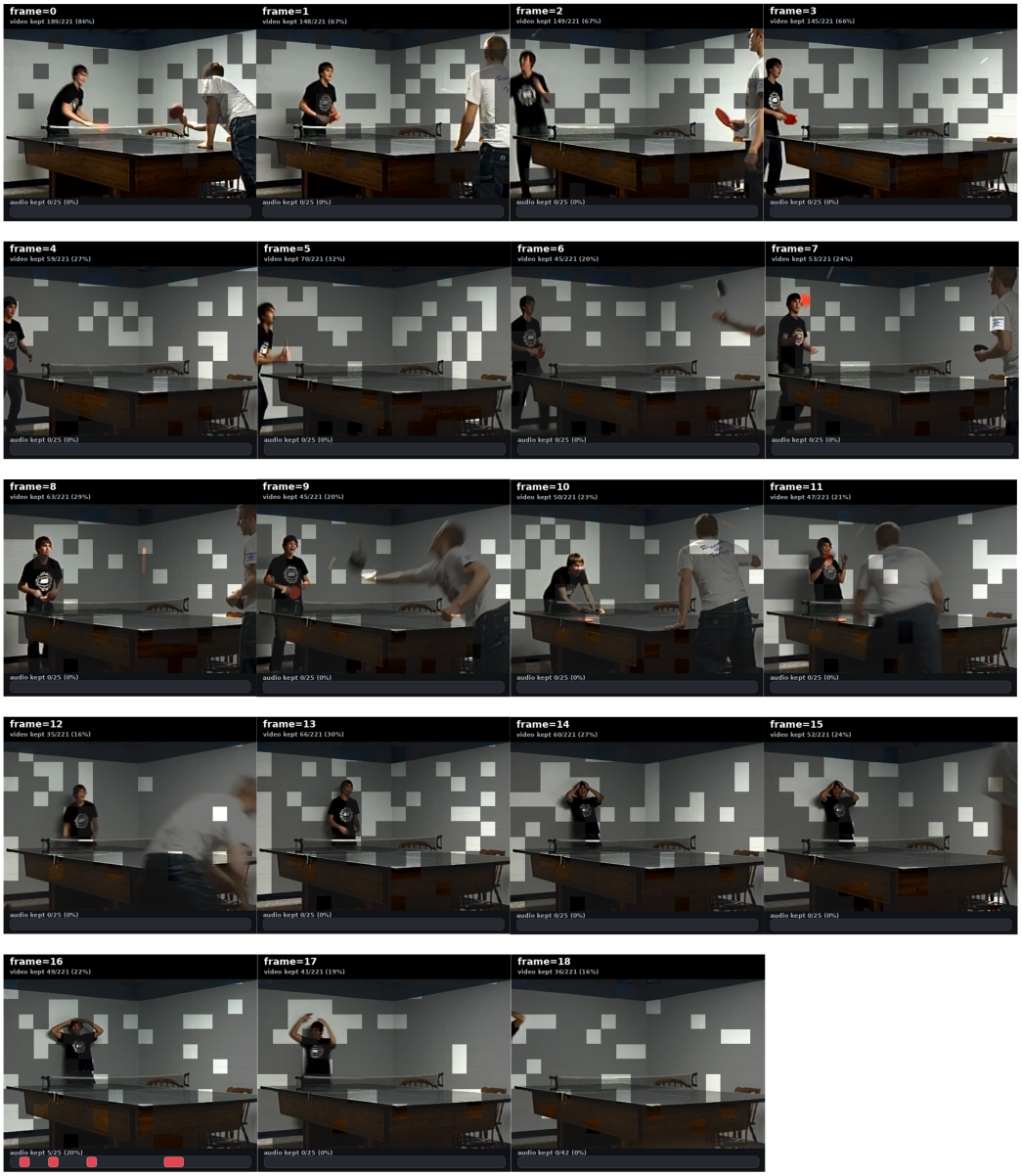
We further visualize the audio-video tokens retained by EvoComp-2 on three representative samples from DailyOmni at a 30% token retention ratio, illustrating how token selection varies with query-dependent modality importance. In the following figures, bright regions overlaid on the video frames denote retained visual tokens, while the bars below the frames represent temporally aligned audio tokens, with red segments indicating the retained audio tokens.

Video-Oriented Example. As shown in Figure 4, the question asks about the temporal order of actions performed by the person in the black shirt and therefore primarily depends on visual evidence. EvoComp-2 retains informative visual tokens across the relevant frames while removing almost all audio tokens. This example shows that the compressor can prioritize the visual modality when the query can be answered mainly from visual events.

Audio-Oriented Example. Figure 5 presents an example whose answer depends primarily on information provided by the audio narration, specifically the reason why the Celtics' win is characterized as "gritty". Accordingly, EvoComp-2 retains a large proportion of audio tokens. This demonstrates its ability to shift the token budget toward audio when acoustic or spoken information is more important for answering the query.

Audio-Video Collaboration Example. As illustrated in Figure 6, answering the question requires jointly understanding the commentator's narration and the corresponding visual events to determine which player dives first. EvoComp-2 therefore retains informative tokens from both audio and video, resulting in a more balanced allocation between the two modalities. This example highlights the ability of EvoComp-2 to preserve complementary cross-modal evidence when both modalities are necessary for reasoning.

Video Oriented



Question:
Which of the following sequences accurately reflects the order of events involving the person in the black shirt?

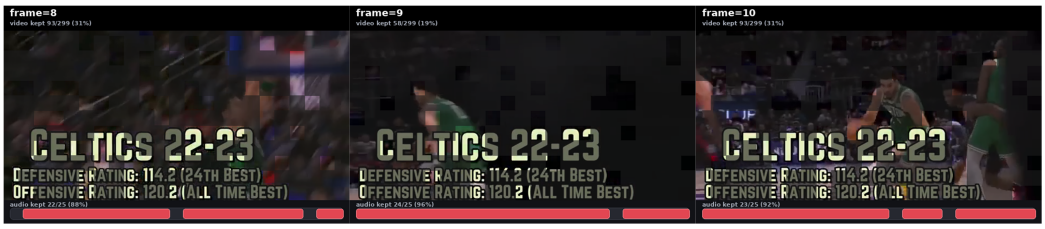
Options:
A. Puts hands on head, hits ball, laughs.
B. Moves away from table, puts hands on head, hits ball.
C. Hits ball, laughs, puts hands on head.
D. Laughs, hits ball, moves away from table.

EvoComp-2's Answer:
C. Hits ball, laughs, puts hands on head. ✅

Figure 4: Video-oriented token selection by EvoComp-2 on DailyOmni at a 30% token retention ratio. The query primarily relies on visual evidence, and EvoComp-2 retains informative visual tokens while pruning most audio tokens.

Audio Oriented

...



...best offensive rating of all time. The player in green successfully makes a shot into the basket.



On Saturday, without Jaylen Brown, Malcolm Brogdon, and Al Horford,



the Seas picked up a gritty road W in Motor City.



The player with the ball is trying to pass the ball to his teammate...

...

Question:

What audio-described factor contributed most to the Celtics' win being characterized as a 'gritty' performance in Motor City?

Options:

A. Jason Tatum extending his 30+ point scoring streak

B. Missing three key rotation players due to injury

C. Achieving the highest offensive rating in NBA history

D. Overcoming a 15-point fourth quarter deficit

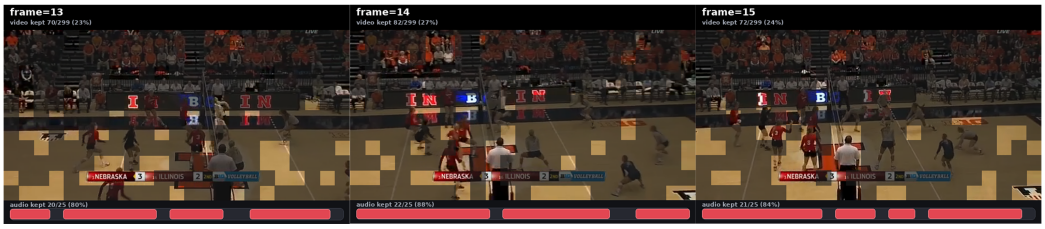
EvoComp-2's Answer:

B. Missing three key rotation players due to injury ✅

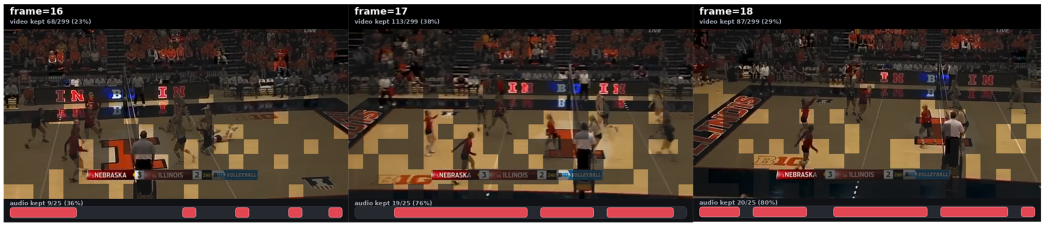
Figure 5: Audio-oriented token selection by EvoComp-2 on DailyOmni at a 30% token retention ratio. The query primarily relies on audio evidence, leading EvoComp-2 to retain substantially more audio tokens.

Audio-Video Collaboration

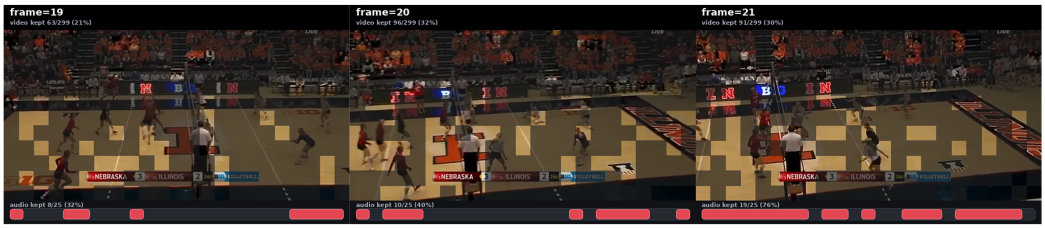
...



Pulter tries to go for the back corner, but Albright was there and then Becky with another swing

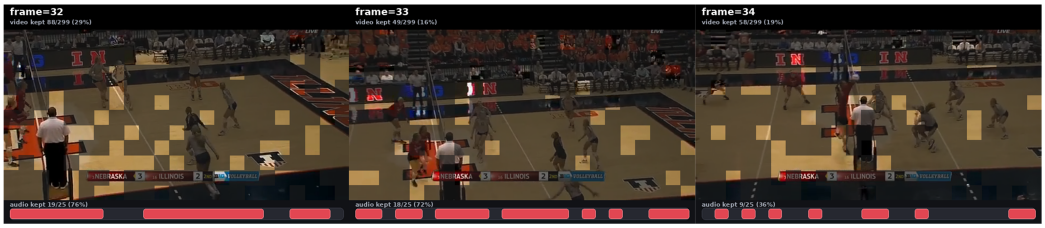


Connor dives and puts it over.



Rally will continue.

...



Donnelly the dig. Again, Donnelly dives for that ball and keeps it in play.

...

Question:

Based on the commentator's narration and visual events, which player dived first to save the ball during the rally?

Options:

A. Donnelly, followed by Connor after a long rally

B. Donnelly made two consecutive saves

C. Connor, followed by Donnelly

D. Both players dove simultaneously after a block

EvoComp-2's Answer:

C. Connor, followed by Donnelly ✅

Figure 6: Audio-video collaborative token selection by EvoComp-2 on DailyOmni at a 30% token retention ratio. The query requires complementary audio and visual evidence, and EvoComp-2 retains informative tokens from both modalities.