JEC-OMNI: TOKEN COMPRESSION FOR EFFICIENT OMNIMODAL LARGE LANGUAGE MODELS VIA JOINT ENTROPY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal Large Language Models (OmniLLMs) have demonstrated remarkable capabilities in audio-video understanding, but encoding continuous signals yields massive token sequences that impose enormous computational overhead and memory footprints. While training-free token compression offers a practical solution, existing omnimodal methods fail to precisely identify both intra- and cross-modal redundancies, struggling to stably maintain the full-token baseline accuracy under target token budgets. In this paper, we propose JEC-Omni, a novel training-free token compression framework that explicitly preserves high joint entropy for OmniLLMs. Our approach systematically eliminates redundancies through two core modules: MAST-P, which removes intra-frame spatial and inter-frame motion redundancies in the visual stream, and SICM-P, which leverages intra-modal temporal consistency and cross-modal semantic overlap to prune the audio stream. Extensive evaluations on the Qwen2.5-Omni and Qwen3-Omni architectures demonstrate that JEC-Omni achieves a superior trade-off between inference efficiency and accuracy. Remarkably, our method maintains 97% of the original model accuracy while retaining merely 25% of the raw audio-video tokens, reducing the end-to-end inference memory footprint by 1.1x to 1.8x and decreasing the Time-to-First-Token (TTFT) by 1.5x to 2.5x.

1 INTRODUCTION

Omnimodal Large Language Models (OmniLLMs) Xu et al. (2025a;b); Team et al. (2025); Cui et al. (2026); Tong et al. (2025); AI et al. (2025) have demonstrated remarkable capabilities in audio-video understanding Yang et al. (2025); Fu et al. (2025); Ge et al. (2025); Hong et al. (2025). They achieve this by integrating audio and video streams into a unified representation space for joint processing by Large Language Models (LLMs). However, encoding continuous audio-video signals yields massive token sequences, imposing enormous computational overhead and memory footprint on OmniLLMs.

To mitigate these bottlenecks, training-free token compression Shao et al. (2025) has emerged as a practical solution. By pruning or merging less informative tokens, this approach effectively reduces sequence lengths. Currently, most methods focus on video token compression Tao et al. (2025a); Chen et al. (2024); Huang et al. (2025); Chen et al. (2025) in Video Large Language Models (VideoLLMs) Bai et al. (2025); Li et al. (2024; 2025). However, due to distinct redundancy distributions of video and audio tokens, these video-centric methods cannot be directly applied to audio-video compression. While a few approaches Tao et al. (2025b); Li & Huang (2026) tailored for OmniLLMs have been proposed, they struggle to stably maintain the full-token baseline accuracy.

Maintaining accuracy under a target token budget requires eliminating redundancy while maximizing retained information. Formally, given the video and audio token distributions, V and A, their total information is governed by their joint entropy Kullback (1997):

$$H(V, A) = H(V) + H(A) - I(V; A), \tag{1}$$

where $H(V)$ and $H(A)$ denote the entropies of the video and audio modalities, respectively, and $I(V; A)$ is their mutual information. This formulation establishes that to maximize $H(V, A)$, one

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.

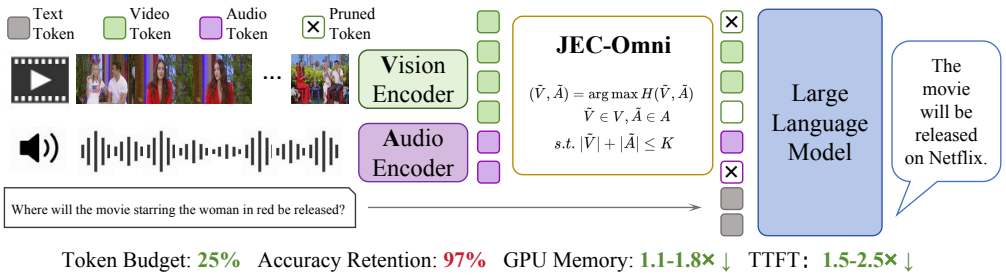

Figure 1: We introduce JEC-Omni, a novel training-free token compression method that preserves the joint entropy of the retained subset. Impressively, it maintains $> 97\%$ average downstream accuracy while retaining merely 25% of the raw audio-video tokens, substantially decreasing the end-to-end inference memory footprint and Time to First Token (TTFT).

must ***maximize $H(V)$ and $H(A)$ by removing intra-modal redundancy*** to preserve essential information, while simultaneously ***minimizing the mutual information $I(V;A)$ by discarding cross-modal redundancy***.

However, existing omnimodal compression frameworks exhibit limitations in eliminating both intra- and cross-modal redundancies. For intra-modal video compression, they overlook the redundancy caused by the spatial movement of objects over time. For cross-modal compression, they either compress each modality independently Su et al. (2026), or leverage cross-modal signals only for temporal alignment Li & Huang (2026); Deng et al. (2026), modality-level budget allocation Yang et al. (2026), or retention cues rather than reduction Tao et al. (2025b), ultimately failing to identify and prune information already conveyed across modalities.

To overcome these limitations, we propose **JEC-Omni**, a novel training-free token compression framework that explicitly preserves high joint entropy for OmniLLMs. It achieves this via two core modules: (1) **MAST-P** (Motion-Aware Spatio-Temporal Token Pruning) for the visual stream, which leverages H.264-inspired paradigms Wiegand et al. (2003) to eliminate intra-frame spatial and inter-frame motion redundancies; and (2) **SICM-P** (Synergistic Intra- and Cross-Modal Token Pruning) for the audio stream, which evaluates intra-modal temporal consistency alongside cross-modal semantic overlap with the retained video tokens. Together, these mechanisms effectively suppress mutual information and retain essential joint information, ensuring stable accuracy under a target token budget.

Extensive experiments demonstrate that our approach achieves a superior trade-off between inference efficiency and accuracy. Evaluated on the Qwen2.5-Omni Xu et al. (2025a) and Qwen3-Omni Xu et al. (2025b) architectures, our framework successfully maintains 97% of the original model accuracy while retaining merely 25% of the raw audio-video tokens. It reduces the end-to-end inference memory footprint by $1.1\times$ to $1.8\times$ and decreases the Time-to-First-Token (TTFT) by $1.5\times$ to $2.5\times$, significantly lowering deployment costs.

2 Related Work

2.1 Omnimodal Large Language Models

The evolution of Large Language Models (LLMs) has rapidly progressed from text-only Brown et al. (2020) and image-text Li et al. (2024); Bai et al. (2025); Li et al. (2025) systems to sophisticated Omnimodal Large Language Models (OmniLLMs) Xu et al. (2025b;a); Team et al. (2025); AI et al. (2025). Unlike traditional VideoLLMs that focus exclusively on visual inputs, OmniLLMs employ dedicated visual and audio encoders Radford et al. (2021); Tschannen et al. (2025); Baevski et al. (2020) to transform continuous audio-video streams into token sequences for subsequent understanding and reasoning by the LLM. Recent advanced architectures, such as the dense model Qwen2.5-Omni and the Mixture-of-Experts (MoE) based Qwen3-Omni, have demonstrated formidable capabilities in complex audio-video understanding tasks Yang et al. (2025); Fu et al. (2025); Ge et al. (2025); Hong et al. (2025). However, the introduction of massive audio and video tokens poses significant challenges for the practical deployment of OmniLLMs.

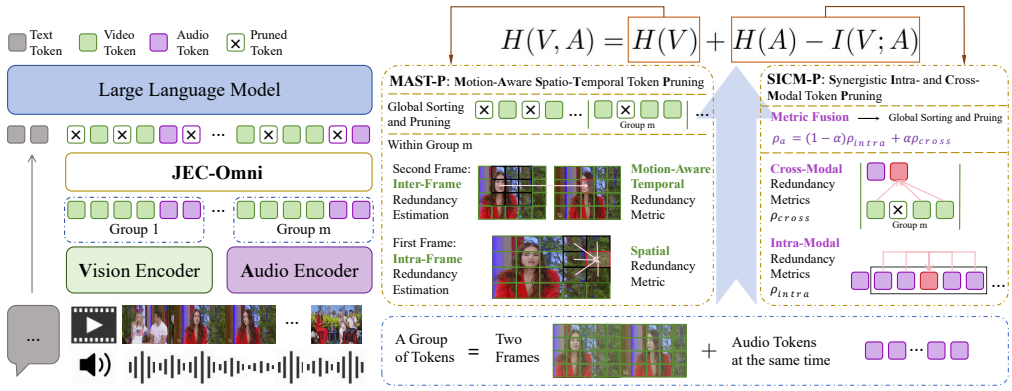

Figure 2: The **JEC-Omni** framework incorporates **MAST-P** and **SICM-P** to compress video and audio tokens, respectively, thereby preserving critical uni-modal features and removing cross-modal redundancies

2.2 Token Compression

To alleviate the computational burden of massive sequence lengths, training-free token compression has emerged as a highly effective strategy. By pruning or merging uninformative tokens before they enter the LLM's core transformer blocks, these methods can significantly accelerate inference and reduce memory usage. Historically, token compression has been extensively explored within visual contexts, including image Chen et al. (2024); Bolya et al. (2022) and video compression Tao et al. (2025a); Chen et al. (2024); Huang et al. (2025); Chen et al. (2025). However, directly applying these video-centric techniques to omnimodal tokens is suboptimal, as they fail to account for the distinct redundancy distributions of audio and video signals.

Recently, a few pioneering frameworks have attempted to address token compression specifically for OmniLLMs. OmniZipTao et al. (2025b) introduces an audio-guided dynamic token compression framework. It adopts a "listen-to-prune" paradigm that utilizes audio to dynamically guide video token pruning. DASH Li & Huang (2026) aligns token compression with semantic structure. It treats audio embeddings as semantic anchors and detects boundary candidates to form dynamic segments. However, current methods Su et al. (2026); Deng et al. (2026); Yang et al. (2026); Tao et al. (2025b); Li & Huang (2026) fail to precisely identify intra- and cross-modal redundancies, making them unable to maintain joint entropy under a specified token budget, which consequently leads to unstable accuracy preservation. This is the core problem our work aims to address.

3 Method

In this section, we present the methodology of our proposed token compression framework. We first introduce the architectural foundation of OmniLLMs in Section 3.1, focusing primarily on the formulation and composition of the input audio and video token sequences. Subsequently, we detail the formal algorithmic design of the proposed JEC-Omni framework for effectively reducing this sequence length.

3.1 Background on OmniLLMs

Before being processed by the LLM backend Brown et al. (2020); Bai et al. (2023), continuous audio-video inputs must be transformed into discrete token sequences. Given a video accompanied by synchronized audio, the visual and audio streams are first sampled at their respective frame rates and sampling frequencies. This sampling process yields the raw video input $X_v \in \mathbb{R}^{T\times H\times W\times 3}$ and the raw audio input $X_a \in \mathbb{R}^{L_a}$, where T, H, and W denote the number of sampled video frames, frame height, and frame width, respectively, and L_a represents the discrete length of the sampled audio signal. These raw inputs are subsequently mapped through a dedicated vision encoder $f_v(\cdot)$ and an audio encoder $f_a(\cdot)$ to generate their corresponding token embeddings:

$$Z_v = f_v(X_v), \quad Z_a = f_a(X_a), \tag{2}$$

where $Z_v \in \mathbb{R}^{N_v \times h \times w \times D}$ and $Z_a \in \mathbb{R}^{N_a \times D}$. Specifically, N_v and N_a represent the temporal token counts for the video and audio streams, respectively. For the visual features, h and w denote the spatial height and width of the token grid, while D signifies the hidden dimension of the joint multimodal representation space.

In practice, the token counts per unit time for video and audio streams are fixed but different. We define κ as the temporal ratio, representing the number of audio tokens corresponding to a single video temporal index. Thus, the relationship between the total audio tokens N_a and video tokens N_v is given by $N_a = \kappa \cdot N_v$. The token sequences Z_v and Z_a can be formulated as sets of temporal segments indexed by t:

$$Z_v = \left\{ Z_v^t \right\}_{t=1}^{N_v}, \tag{3}$$

$$Z_a = \left\{ Z_a^{(t-1)\kappa+1}, Z_a^{(t-1)\kappa+2}, \ldots, Z_a^{t\kappa} \right\}_{t=1}^{N_v}, \tag{4}$$

where $Z_v^t \in \mathbb{R}^{h \times w \times D}$ denotes the spatial video token grid at temporal index t. This video token grid is explicitly aligned with its corresponding subset of κ audio tokens $Z_a^{(t-1)\kappa+1}, \ldots, Z_a^{t\kappa} \in \mathbb{R}^D$. Subsequently, Z_v is flattened into a 1D sequence $Z_v \in \mathbb{R}^{N_v hw \times D}$ and fed into the LLM alongside Z_a for unified understanding and reasoning. The objective of token compression is to reduce this sequence length while maintaining the inference accuracy of the Full-token baseline.

3.2 The JEC-Omni Framework

This subsection presents JEC-Omni, a token compression method that preserves the joint entropy of the retained tokens by systematically eliminating intra- and cross-modal redundancies. Specifically, Section 3.2.1 details the MAST-P module, which performs Motion-Aware Spatio-Temporal Token Pruning on video tokens. Subsequently, Section 3.2.2 introduces the SICM-P module, which executes Synergistic Intra- and Cross-Modal Token Pruning on audio tokens.

3.2.1 Video Token Compression via MAST-P

According to the Eq. 1, maximizing the overall joint entropy $H(V, A)$ necessitates preserving the marginal entropy $H(V)$ of the visual modality. We analyze the structural distribution of video tokens by drawing insights from the classic H.264 standard Wiegand et al. (2003). In practical video compression, sequences are partitioned into groups to leverage two fundamental paradigms: intra-frame prediction exploits the high spatial similarity of adjacent blocks within a single frame, while inter-frame prediction tracks the displacement of matching blocks across consecutive frames to capture their relative motion.

Guided by this classical paradigm, we categorize video token redundancies into two main types: (1) *spatial redundancy* within a single frame, caused by the high semantic similarity of adjacent regions, and (2) *temporal redundancy* across consecutive frames, arising from the localized semantic overlap of moving objects. Consequently, we propose the Motion-Aware Spatio-Temporal Token Pruning (MAST-P) module to comprehensively eliminate both forms of redundancy.

Frame Grouping. Mimicking the group-based structure in H.264 coding, we first partition the video token sequence Z_v into consecutive, non-overlapping temporal groups to facilitate redundancy extraction. For long-context video understanding, the frame sampling rate is typically sparse (e.g., 1–2 frames per second), meaning that adjacent token grids already exhibit noticeable semantic shifts. Therefore, we set the group size to two visual temporal indices. Formally, the m-th group $\mathcal{G}_m$ contains the following paired video and audio tokens:

$$\mathcal{G}_m = \left(\{ Z_v^{2m-1}, Z_v^{2m} \}, \{ Z_a^{(2m-2)\kappa+1}, \ldots, Z_a^{2m\kappa} \} \right), \tag{5}$$

where $m \in \{1, 2, \ldots, N_v/2\}$.

Intra-Frame Redundancy Estimation. For the first video token grid in each group, denoted as the anchor frame Z_v^{2m-1}, we perform spatial redundancy estimation. Let $z_v^{2m-1}[i, j] \in \mathbb{R}^D$ represent a token at the spatial coordinate (i, j) within this $h \times w$ grid. Let $\text{sim}(a, b) = \frac{a \cdot b^\top}{\|a\| \|b\|}$ denote the

cosine similarity function. We define the spatial redundancy metric $\rho_s(z_v^{2m-1}[i,j])$ as the average similarity between $z_v^{2m-1}[i,j]$ and its local spatial neighborhood $\mathcal{N}(i,j)$ of size $\omega_s \times \omega_s$:

$$\rho_s(z_v^{2m-1}[i,j]) = \frac{1}{|\mathcal{N}(i,j)|} \sum_{(x,y)\in\mathcal{N}(i,j)} \text{sim}\left(z_v^{2m-1}[i,j], z_v^{2m-1}[x,y]\right), \tag{6}$$

where a higher ρ_s indicates that the token is surrounded by homogeneous features (e.g., background) and thus carries higher spatial redundancy.

Inter-Frame Redundancy Estimation. For the second token grid in each group, denoted as the non-anchor frame Z_v^{2m}, we execute motion-aware temporal redundancy estimation. Objects or regions frequently displace across adjacent frames due to motion. To accommodate this spatial shift, the temporal redundancy metric $\rho_m(z_v^{2m}[i,j])$ for a token $z_v^{2m}[i,j] \in Z_v^{2m}$ is defined as its maximum similarity to a local spatial search window $\mathcal{W}(i,j)$ of size $\omega_m \times \omega_m$ in the preceding anchor frame Z_v^{2m-1}:

$$\rho_m(z_v^{2m}[i,j]) = \max_{(x,y)\in\mathcal{W}(i,j)} \text{sim}\left(z_v^{2m}[i,j], z_v^{2m-1}[x,y]\right), \tag{7}$$

where a higher ρ_m indicates that the visual content can be closely matched within the preceding frame despite local motion, and thus carries higher temporal redundancy.

Global Sorting and Pruning. Finally, we aggregate the estimated redundancy metrics to perform token-level pruning. Specifically, the spatial metrics ρ_s of all tokens in the anchor frames $\{Z_v^{2m-1}\}$ and the temporal metrics ρ_m of all tokens in the non-anchor frames $\{Z_v^{2m}\}$ are globally ranked within their respective categories. To maximize the retained visual marginal entropy $H(V)$, we preserve only the top $r_v\%$ tokens exhibiting the lowest redundancy scores in each category.

3.2.2 Audio Token Compression via SICM-P

According to the information-theoretic decomposition in Eq. 1, maximizing the joint entropy $H(V,A)$ further necessitates preserving the audio marginal entropy $H(A)$ while minimizing the cross-modal mutual information $I(V;A)$. Driven by this dual objective, the SICM-P module eliminates intra-modal temporal redundancies and cross-modal semantic overlaps within the audio token sequence to retain maximal acoustic information under a specified token budget.

Intra-Modal Redundancy Estimation. Audio signals intrinsically exhibit high temporal continuity, resulting in consecutive tokens sharing similar acoustic semantics (e.g., sustained background noise). To quantify this, we define a 1D temporal neighborhood $\mathcal{T}(i)$ of size ω_a centered at the i-th audio token $z_a[i]$. Using the similarity function $\text{sim}(\cdot,\cdot)$ defined previously, we compute the intra-modal redundancy metric $\rho_{intra}(z_a[i])$ as the average similarity within this local temporal window:

$$\rho_{intra}(z_a[i]) = \frac{1}{|\mathcal{T}(i)|} \sum_{j\in\mathcal{T}(i)} \text{sim}\left(z_a[i], z_a[j]\right). \tag{8}$$

Cross-Modal Redundancy Estimation. To minimize the mutual information $I(V;A)$, we must penalize audio tokens whose semantic information is already adequately conveyed by the co-occurring visual frames. For an audio token $z_a[i]$ belonging to the m-th temporal group $\mathcal{G}_m$, let $\tilde{\mathcal{V}}_m$ denote the subset of video tokens within $\mathcal{G}_m$ that survived the MAST-P pruning phase. The cross-modal redundancy metric $\rho_{cross}(z_a[i])$ is defined as its average similarity to all retained visual tokens in $\tilde{\mathcal{V}}_m$:

$$\rho_{cross}(z_a[i]) = \frac{1}{|\tilde{\mathcal{V}}_m|} \sum_{v\in\tilde{\mathcal{V}}_m} \text{sim}\left(z_a[i], v\right). \tag{9}$$

Synergistic Metric Fusion. To unify these two dimensions of redundancy, we independently map ρ_{intra} and ρ_{cross} to a normalized range of $[0,1]$ via Min-Max normalization, yielding $\hat{\rho}_{intra}$ and $\hat{\rho}_{cross}$. The comprehensive redundancy score $\rho_a(z_a[i])$ is formulated as a weighted sum of both metrics, governed by a balancing hyperparameter $\alpha \in [0,1]$:

$$\rho_a(z_a[i]) = (1-\alpha)\cdot\hat{\rho}_{intra}(z_a[i]) \quad + \alpha\cdot\hat{\rho}_{cross}(z_a[i]). \tag{10}$$

Here, α controls the trade-off between retaining acoustic uniqueness and minimizing visual-acoustic semantic overlap.

Global Sorting and Pruning. Finally, we aggregate the comprehensive redundancy scores ρ_a to perform token-level pruning. All audio tokens are globally ranked in ascending order of their scores, preserving only the top $r_a\%$ tokens exhibiting the lowest ρ_a values.

4 EXPERIMENTS

Method	Settings		AVUT							VideoMME	ShortVid	Overall Avg.	Rel. Perf.
	Retained Ratio	FLOPs Ratio	EL	OR	OM	IE	CC	CM	Avg.	wo	Avg. Score		
Qwen2.5-Omni-7B													
Full Tokens	100.0%	100.0%	38.2	67.8	59.6	85.6	44.1	67.7	64.5	66.0	70.5	67.0	100.0%
Random	40.0%	34.0%	31.7	58.5	53.3	74.9	**43.2**	59.0	56.9	65.0	67.7	63.2	94.3%
DyCoke (CVPR2025)	35.0%	29.0%	32.9	62.1	54.9	74.5	39.0	58.3	57.4	65.2	68.0	63.5	94.8%
OmniZip (CVPR2026)	35.0%	29.0%	34.1	67.5	54.6	83.7	42.4	61.2	61.0	66.1	69.0	65.4	97.6%
DASH (Arxiv2026)	35.0%	29.0%	32.0	62.1	**58.9**	85.5	40.4	**64.2**	61.5	**66.7**	68.4	65.5	97.8%
JEC-Omni (Ours)	35.0%	29.0%	**37.1**	**68.5**	53.8	**86.2**	40.7	63.5	**62.3**	66.2	**69.1**	**65.9**	**98.4%**
DASH (Arxiv2026)	25.0%	20.0%	36.0	64.7	**56.0**	**86.3**	36.5	60.8	60.9	**66.0**	**67.8**	**64.9**	**96.9%**
JEC-Omni (Ours)	25.0%	20.0%	**36.5**	**66.2**	53.8	84.4	**39.8**	**61.9**	**61.1**	65.4	67.6	64.7	96.6%
Qwen2.5-Omni-3B													
Full Tokens	100.0%	100.0%	32.9	65.3	58.4	85.0	44.1	62.6	62.2	62.6	69.4	64.7	100.0%
Random	40.0%	31.0%	28.2	60.8	54.9	73.1	42.3	61.6	57.5	60.6	67.0	61.7	95.4%
DyCoke (CVPR2025)	35.0%	26.0%	**32.9**	62.1	54.9	74.5	38.9	58.3	57.4	61.0	67.5	62.0	95.8%
OmniZip (CVPR2026)	35.0%	26.0%	28.8	**63.1**	**58.2**	84.0	**42.4**	60.4	60.1	**62.7**	68.0	63.6	98.3%
DASH (Arxiv2026)	35.0%	26.0%	**32.9**	61.7	55.6	83.4	**42.4**	60.0	59.9	62.6	68.1	63.5	98.1%
JEC-Omni (Ours)	35.0%	26.0%	32.4	60.8	56.6	**85.6**	40.7	**62.6**	**60.8**	62.5	**68.5**	**63.9**	**98.8%**
DASH (Arxiv2026)	25.0%	18.0%	29.0	**65.0**	**59.5**	81.1	**39.1**	58.4	58.8	**61.7**	67.0	62.5	96.6%
JEC-Omni (Ours)	25.0%	18.0%	**30.0**	61.4	55.6	**84.7**	39.0	**60.0**	**59.5**	**61.7**	**67.1**	**62.8**	**97.1%**

Table 1: Performance comparison of different methods on Audio-Video QA benchmarks across various token retention budgets (35%-40% and 25%) for each model. The best result among token pruning methods for each metric is highlighted in bold, and the second-best is underlined.

4.1 EVALUATION SETUPS AND IMPLEMENTATION DETAILS

Benchmarks. We employ four audio-video understanding benchmarks (AVUT Yang et al. (2025), VideoMME Fu et al. (2025), ShortVid-Bench Ge et al. (2025), and WorldSense Hong et al. (2025)) to evaluate model performance before and after token compression. Specifically, AVUT focuses on six distinct tasks: event localization (EL), object matching (OM), OCR matching (OR), information extraction (IE), content counting (CC), and character matching (CM). WorldSense assesses joint audio-video understanding capabilities across diverse scenarios, such as Tech & Science, Culture & Politics, Daily Life, Film & TV Performance, Games, Sports, and Music. VideoMME serves as a classic video understanding benchmark. Finally, ShortVid-Bench evaluates holistic comprehension of real-world short videos.

Evaluation Metrics. Beyond the standard accuracy metrics of the aforementioned benchmarks, we assess the inference efficiency after token compression using two practical metrics: average GPU memory usage and Time-to-First-Token (TTFT) latency. Both metrics provide an end-to-end efficiency evaluation spanning from the encoder to the LLM, aligning with the requirements of real-world deployment scenarios, rather than measuring only the LLM backbone as in Tao et al. (2025b).

Compared Methods. We evaluate our framework against OmniZip Tao et al. (2025b) and DASH Li & Huang (2026), two state-of-the-art (SOTA) token compression methods specifically designed for OmniLLMs. Additionally, we adapt DyCoke Tao et al. (2025a), a classic visual token compression method originally developed for VideoLLMs, to the OmniLLM architecture for comparison. Specifically, we employ its first-stage TTM module to process both video and audio tokens.

Hyperparameters. For the main results, we set the search window sizes to $\omega_s = 3$, $\omega_m = 3$, and $\omega_a = 23$, and the balancing factor to $\alpha = 0.4$. However, for Qwen3-Omni, which generates fewer audio tokens per unit time, we adjust its audio temporal window to $\omega_a = 11$. For an overall token retention budget of 35%, we set the video and audio retention ratios to $r_v = 30\%$ and $r_a = 64\%$, respectively. For a stricter 25% budget, we configure $r_v = 20\%$ and $r_a = 54\%$.

4.2 MAIN RESULTS

Comparison with State-of-the-Art Methods. Table 1 shows our method achieves remarkable performance retention across Qwen2.5-Omni 3B and 7B models. Specifically, it preserves over 98.4% of average downstream accuracy under a 35% token retention budget, maintaining approximately 97% even at a more aggressive 25% budget. Compared to state-of-the-art baselines, our approach significantly outperforms OmniZip and surpasses DASH in most configurations. Furthermore, Table 2 demonstrates its superior robustness on the WorldSense benchmark, achieving near-lossless performance retention at a 35% budget.

Results on Advanced Models. To evaluate our framework's adaptability, we extend JEC-Omni to Qwen3-Omni. In practical implementation, since this model leverages multi-scale features based on the Deepstack method Meng et al. (2024), the pruning decisions derived from the token sequence must be consistently mapped and applied across these multi-scale representations. As presented in Table 2, our method robustly preserves 98.0% of the average accuracy even under an aggressive 25% token budget. This consistent effectiveness across diverse models strongly validates our approach's versatility.

Method	Retained Ratio	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.	Rel. Perf.
Qwen2.5-Omni-7B											
Full Tokens	100.0%	52.4	50.1	48.5	44.6	43.8	41.6	41.6	47.3	46.8	100.0%
Random	55.0%	47.1	47.0	44.4	41.2	40.0	40.1	40.1	46.3	43.6	93.2%
DyCoke (CVPR2025)	50.0%	48.4	49.9	46.7	41.4	39.9	40.8	40.2	46.5	44.6	95.3%
OmniZip (CVPR2026)	35.0%	48.3	49.5	**47.6**	42.5	40.1	40.2	42.3	46.3	45.3	96.8%
DASH (Arxiv2026)	35.0%	**50.8**	52.1	46.2	40.9	40.8	38.6	**43.5**	**47.3**	45.6	97.4%
JEC-Omni (Ours)	35.0%	**50.8**	**54.0**	46.0	**43.0**	**44.9**	**43.3**	43.3	**47.3**	**46.7**	**99.8%**
DASH (Arxiv2026)	25.0%	50.0	49.2	**47.0**	39.3	41.6	39.9	**41.6**	**45.8**	44.9	95.9%
JEC-Omni (Ours)	25.0%	**50.2**	**52.1**	45.9	**42.0**	**44.9**	**43.8**	40.9	44.3	**45.6**	**97.4%**
Qwen2.5-Omni-3B											
Full Tokens	100.0%	51.5	50.8	45.0	45.4	43.8	42.5	44.2	46.1	46.4	100.0%
Random	55.0%	48.2	46.3	40.7	41.4	38.6	40.0	41.8	43.4	42.8	92.2%
DyCoke (CVPR2025)	50.0%	48.1	48.5	42.3	43.3	39.7	43.4	42.1	43.0	44.0	94.8%
OmniZip (CVPR2026)	35.0%	48.8	48.9	41.8	**46.4**	39.8	42.5	**42.6**	43.1	44.3	95.5%
DASH (Arxiv2026)	35.0%	50.8	**50.2**	**45.7**	43.0	41.9	40.8	**42.6**	43.8	45.3	97.6%
JEC-Omni (Ours)	35.0%	**51.2**	**50.2**	**45.7**	**46.4**	**42.7**	**43.8**	42.3	**45.8**	**46.3**	**99.6%**
DASH (Arxiv2026)	25.0%	**50.2**	47.9	44.7	43.3	41.2	42.5	40.9	43.6	44.6	96.1%
JEC-Omni (Ours)	25.0%	49.4	**49.8**	**45.1**	**45.1**	**42.3**	**44.2**	**41.6**	**44.1**	**45.3**	**97.6%**
Qwen3-Omni-30B-A3B											
Full Tokens	100.0%	56.3	60.5	54.7	54.9	56.2	48.9	49.5	54.2	54.5	100.0%
JEC-Omni (Ours)	35.0%	57.1	58.6	55.0	54.6	56.6	46.4	47.2	54.2	54.0	99.1%
JEC-Omni (Ours)	25.0%	55.9	57.6	53.8	54.4	55.1	48.1	47.0	54.7	53.4	98.0%

Table 2: Performance comparison of different methods on the WorldSense benchmark across various token retention budgets (35%-55% and 25%) for each Qwen2.5-Omni model. The best result among token pruning methods for each metric is highlighted in bold, and the second-best is underlined.

4.3 ABLATION STUDY

Ablation on MAST-P. The Motion-Aware Spatio-Temporal Token Pruning (MAST-P) module is the core of our video token compression framework. Table 3 illustrates the accuracy variations on the WorldSense benchmark as temporal and motion-aware redundancy mechanisms are progressively integrated. Specifically, the *Spatial* baseline performs only intra-frame redundancy estimation. The *Spatial+Temporal* variant introduces inter-frame redundancy but lacks a motion search window for capturing object displacements (i.e., $\omega_m = 1$). *Spatial+Temporal+Motion* represents our complete approach. Results demonstrate that the motion-aware mechanism more thoroughly exploits video token redundancy, yielding superior accuracy retention under a fixed token budget.

Comparison with Alternative Video Compression Methods. Table 4 further validates our design by replacing MAST-P with alternative video compression methods. For OmniZip, we adopt its dedicated ISTC Block. Inherently, DyCoke and OmniZip are limited to standard spatio-temporal pruning and lack motion awareness. Consequently, MAST-P significantly outperforms all baselines. This demonstrates that our method better aligns with the underlying redundancy distribution of video tokens, achieving more precise redundancy elimination.

Ablation on SICM-P. The core of the Similarity-Induced Cross-Modal Pruning (SICM-P) module lies in simultaneously exploiting both intra-modal and cross-modal redundancies for audio to-

Settings			Retain Ratio	
Spatial	Temporal	Motion	25%	35%
✓			44.5	44.3
✓	✓		44.8	45.6
✓	✓	✓	**45.3**	**46.3**

Table 3: Ablation study of MAST-P on WorldSense using Qwen2.5-Omni-3B. The best results are in bold, and our proposed approach is shaded in gray.

Settings		Qwen2.5-Omni	
Video	Audio	3B	7B
Random	SICM-P (Ours)	45.7	45.7
DyCoke		44.9	45.3
OmniZip		45.7	46.2
MAST-P (Ours)		**46.3**	**46.7**

Table 4: Comparison of various video token compression methods on the WorldSense benchmark under a 35% token retention budget. Best results are in bold.

Settings		Qwen2.5-Omni	
Intra-modal	Cross-modal	3B	7B
✓		46.0	46.5
	✓	46.1	46.3
✓	✓	**46.3**	**46.7**

Table 5: Ablation study of SICM-P on WorldSense under a 35% token retention budget. The best results are in bold, and our approach is shaded in gray.

Settings		Qwen2.5-Omni	
Video	Audio	3B	7B
MAST-P (Ours)	Random	43.9	44.3
	DyCoke	44.4	44.8
	OmniZip	44.7	45.2
	SICM-P (Ours)	**45.3**	**45.6**

Table 6: Comparison of various audio token compression methods on the WorldSense benchmark under a 25% token retention budget. Best results are in bold.

ken compression. Table 5 evaluates the individual contribution of each mechanism by comparing our complete design against variants that exclusively eliminate either intra-modal or cross-modal redundancy. The results indicate that jointly addressing both forms of redundancy yields the highest accuracy retention, demonstrating their strong complementarity in preserving essential audio tokens.

Comparison with Alternative Audio Compression Methods. Table 6 further validates our design by replacing SICM-P with alternative audio compression baselines. Inherently, OmniZip relies primarily on intra-modal redundancy for audio compression. Although it computes cross-modal similarity, it retains these highly similar tokens, failing to actually prune cross-modal redundancy. Consequently, SICM-P significantly outperforms all baselines. This demonstrates that our method better aligns with the underlying redundancy distribution of audio tokens, achieving more precise redundancy elimination.

4.4 Hyperparameter Sensitivity

As illustrated in Figure 3, our sensitivity analyses on the WorldSense benchmark reveal a consistent underlying trend: JEC-Omni achieves optimal performance when employing strictly bounded spatio-temporal scopes ($\omega_s = 3, \omega_m = 3, \omega_a = 23$) alongside a balanced multi-modal fusion ($\alpha = 0.4$). Across both MAST-P and SICM-P modules, excessively small search windows fail to capture local motion or temporal continuity, while overly large scopes inevitably introduce irrelevant background noise and semantic mismatches. Furthermore, the optimal balance factor firmly validates our synergistic design, confirming that fusing intra- and cross-modal metrics is strictly superior to evaluating redundancies in isolation.

5 Efficiency Analyses

Figure 4 reports the average end-to-end GPU memory footprint and Time to First Token (TTFT) per sample on the AVUT dataset. The results indicate that our approach maintains a slight efficiency advantage over OmniZip and DASH. We attribute this to our attention-logit-free redundancy formulation. By avoiding the extraction of attention logits, our method ensures seamless compatibility with FlashAttention Dao (2024), allowing all attention layers to achieve highly efficient inference.

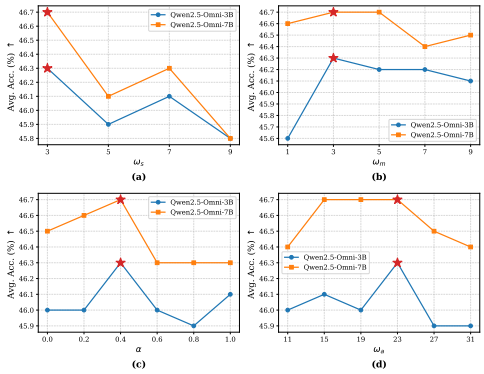

Figure 3: Hyperparameter sensitivity analysis on the WorldSense benchmark under a 35% token retention budget. Red stars denote the default parameter configurations adopted in our main evaluation.

Method	GPU Mem. (GB)↓	TTFT (ms)↓
	Qwen2.5-Omni-7B	
Full Tokens	24.87 (1.00×)	2471.09 (1.00×)
DyCoke	22.26 (1.12×)	1256.72 (1.97×)
OmniZip	22.51 (1.10×)	1271.52 (1.94×)
DASH	22.40 (1.11×)	1234.68 (2.00×)
Ours	**22.03 (1.13×)**	**1198.03 (2.06×)**
	Qwen2.5-Omni-3B	
Full Tokens	14.89 (1.00×)	1698.89 (1.00×)
DyCoke	12.46 (1.20×)	1042.29 (1.63×)
OmniZip	12.75 (1.17×)	1054.60 (1.61×)
DASH	12.69 (1.17×)	1063.18 (1.60×)
Ours	**12.31 (1.21×)**	**1031.70 (1.65×)**

Figure 4: Efficiency evaluation on the AVUT dataset. We compare average GPU memory and Time to First Token (TTFT). The values in parentheses denote the speedup/reduction ratio compared to Full Tokens.

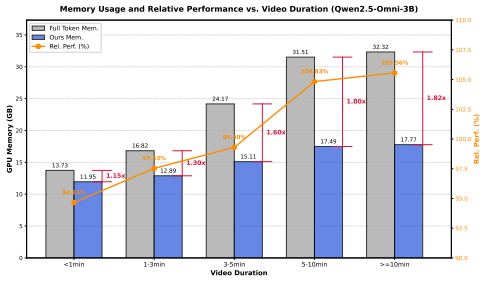

Figure 5: End-to-end GPU memory consumption before and after token compression for videos of varying durations. The results are evaluated on the WorldSense under a 25% token retention budget.

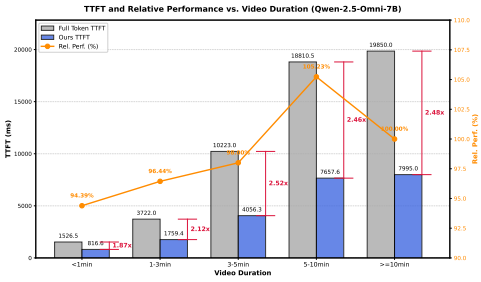

Figure 6: End-to-end Time to First Token (TTFT) before and after token compression for videos of varying durations. The results are evaluated on the WorldSense under a 25% token retention budget.

Figures 5 and 6 illustrate the variations in end-to-end GPU memory consumption and inference latency across videos of different durations before and after token compression. The results demonstrate that our framework achieves more pronounced efficiency gains as the video duration extends, delivering up to a 1.82× reduction in memory footprint and a 2.48× speedup in Time to First Token (TTFT). Crucially, on longer video sequences, our method also exhibits superior accuracy retention, even outperforming the baseline.

6 Conclusions

In this paper, we introduced JEC-Omni, a novel training-free token compression framework tailored for efficient OmniLLMs. By grounding our approach in joint entropy optimization, this method eliminates both intra-modal and cross-modal redundancies through the proposed MAST-P and SICM-P modules. Extensive evaluations demonstrate that JEC-Omni significantly reduces memory footprints and TTFT latency, while preserving 97% of the original model accuracy under 25% token budget.

AI use statement

Generative AI assisted with language editing, including text polishing and translation. The authors checked every generated item and take full responsibility.

Ethics statement

This work uses only publicly available and widely recognized multimodal benchmarks for evaluation. It involves no human subjects, sensitive data, or ethical concerns.

Reproducibility statement

We provide the source code as Supplementary Materials to facilitate reproduction, and all key hyperparameters are specified in the "Evaluation Setups and Implementation Details" section.

References

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. *arXiv preprint arXiv:2510.18269*, 2025.

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.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations*, volume 2024, pp. 35549–35562, 2024.

Yuchen Deng, Zidang Cai, Hai-Tao Zheng, Jie Wang, Feidiao Yang, and Yuxing Han. Omnirefine: Alignment-aware cooperative compression for efficient omnimodal large language models, 2026. URL https://arxiv.org/abs/2605.12056.

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

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

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.

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

Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

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

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

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Wanshun Su, Yang Shi, Feihu Liu, Ziwen Yu, Yan Min, Zhuoran Zhang, Qixun Wang, Haotian Wang, Shixuan Liu, Yuanxing Zhang, Peng Wu, Chengfu Huo, and Liang Ding. Omnipack: Unified token compression for efficient omni-modal large language models, 2026. URL `https://arxiv.org/abs/2608.03812`.

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 Computer Vision and Pattern Recognition Conference*, pp. 18992–19001, 2025a.

Keda Tao, Kele Shao, Bohan Yu, Weiqiang Wang, Huan Wang, et al. Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. *arXiv preprint arXiv:2511.14582*, 2025b.

Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, et al. Longcat-flash technical report. *arXiv preprint arXiv:2509.01322*, 2025.

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

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7): 560–576, 2003.

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, 2025a. URL https://arxiv.org/abs/2503.20215.

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, 2026. URL https://arxiv.org/abs/2605.18041.

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