DASH: Dynamic Audio-Driven Semantic Chunking for Efficient Omnimodal Token Compression

Bingzhou Li[1,2] and Tao Huang[1⋆]

[1] Shanghai Jiao Tong University, Shanghai, China
[2] Tongji University, Shanghai, China

Abstract. Omnimodal large language models (OmniLLMs) jointly process audio and visual streams, but the resulting long multimodal token sequences make inference prohibitively expensive. Existing compression methods typically rely on fixed window partitioning and attention-based pruning, which overlook the piecewise semantic structure of audio-visual signals and become fragile under aggressive token reduction. We propose Dynamic Audio-driven Semantic cHunking (DASH), a training-free framework that aligns token compression with semantic structure. DASH treats audio embeddings as a semantic anchor and detects boundary candidates via cosine-similarity discontinuities, inducing dynamic, variable-length segments that approximate the underlying piecewise-coherent organization of the sequence. These boundaries are projected onto video tokens as a soft temporally co-registered segmentation prior. Within each segment, token retention is determined by a tri-signal importance estimator that fuses structural boundary cues, representational distinctiveness, and attention-based salience, mitigating the sparsity bias of attention-only selection. This structure-aware allocation preserves transition-critical tokens while reducing redundant regions. Extensive experiments on AVUT, VideoMME, and WorldSense demonstrate that DASH maintains competitive or superior accuracy while achieving higher compression ratios compared to prior methods. Code is available at: `https://github.com/laychou666/DASH`.

Keywords: Omnimodal Large Language Models · Token Compression · Audio-Visual Processing

1 Introduction

Recent multimodal and omnimodal large language models [7, 8, 20, 28, 30, 34] extend video-language systems [4, 13, 16, 18, 40, 42] toward unified processing of visual, audio, and textual signals. By integrating speech, scene dynamics, and textual reasoning, these models enable richer multimodal understanding across tasks such as video question answering and real-world reasoning. However, this

⋆ Correspondence to: Tao Huang (`t.huang@sjtu.edu.cn`).

capability comes at a significant computational cost: a single video clip can produce tens of thousands of multimodal tokens, and the quadratic complexity of self-attention quickly makes inference memory- and latency-bound.

Token compression [1, 2, 23, 33, 35, 37] has become a practical solution to mitigate this bottleneck. By pruning or merging tokens before they enter the language model, prior work reduces sequence length without retraining large models. Yet despite steady progress, existing approaches share a critical assumption: multimodal tokens are treated as flat, uniformly structured sequences. Compression decisions are typically made within fixed windows [1, 3, 32] and guided primarily by attention scores. This positional partitioning neglects the piecewise-coherent structure of audio-visual sequences, where semantic transitions correspond to distributional shifts in embedding space.

The core challenge lies in the intrinsic structure of audio-visual signals. Unlike static images, audio and video streams evolve over time with highly **non-uniform** semantic density. Speech contains natural boundaries (*e.g.*, pauses, topic transitions, and speaker shifts) that often provide useful cues for meaningful visual changes. These transitions delineate coherent semantic units that should ideally be preserved as atomic structures during compression. When tokens are grouped uniformly and selected via sparse attention alone, these structural transitions are fragmented or discarded, leading to disproportionate information loss under aggressive pruning. In other words, current compression strategies optimize local redundancy while overlooking global semantic organization. From this perspective, token compression can be viewed as allocating limited representational capacity across a temporally structured and semantically segmented signal.

In this work, we argue that effective omnimodal compression should be guided by semantic structure rather than positional regularity. In particular, the audio stream provides a natural structural prior. Compared to visual tokens, audio embeddings exhibit sharper distributional discontinuities at semantic transitions (e.g., pauses and topic shifts), making them a more reliable boundary signal. Because audio and video inputs are temporally co-registered, these audio transitions provide useful cues for organizing video tokens without requiring exact audio-visual boundary equality. Audio therefore acts as a semantic anchor that can guide both segmentation and importance allocation across modalities.

Building on this perspective, we propose **DASH (Dynamic Audio-driven Semantic cHunking)**, a training-free framework for structure-aware omnimodal token compression (Fig. 1). DASH models the multimodal sequence as a set of dynamically inferred semantic segments instead of imposing fixed windows. Specifically, it detects boundary candidates from cosine-similarity drops between adjacent audio embeddings, treating sharp representational changes as proxies for semantic discontinuities. These boundaries induce variable-length chunks that respect content rhythm and approximate the underlying piecewise structure of the sequence.

To introduce a cross-modal structural prior, detected audio boundaries are projected onto video token indices via temporal ratio mapping, yielding audio-

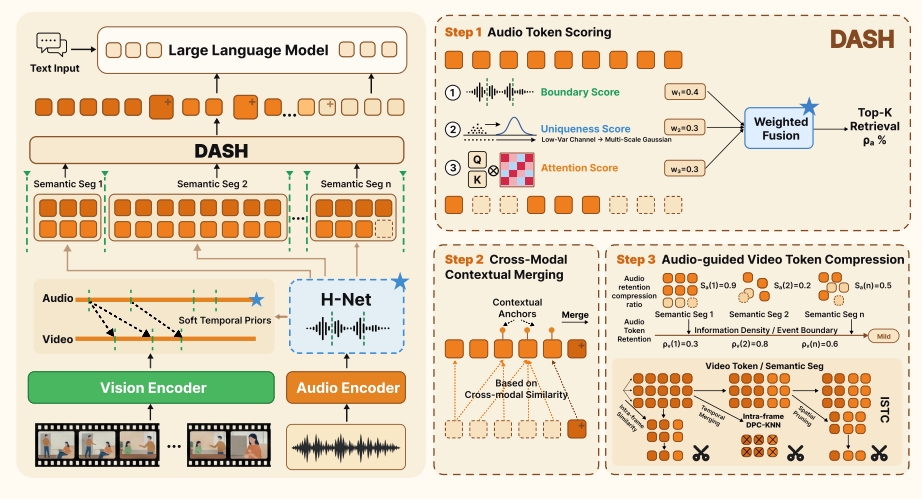

Fig. 1: Overview of DASH. Left: Overall DASH pipeline, where audio/video inputs are encoded, audio semantic boundaries are detected, and projected as soft temporal priors to form dynamic audio-guided segments before compression. **Right**: DASH module details, including tri-signal audio token scoring, contextual merging, and audio-guided video token compression with mild boundary-neighborhood retention.

guided video segments without additional learning. Within each segment, token selection becomes a *multi-signal importance estimation* problem. We combine (i) boundary probability as a structural prior, (ii) multi-scale Gaussian uniqueness capturing representational distinctiveness, and (iii) normalized attention reflecting model-perceived salience. Their fusion yields a smoother importance landscape (Fig. 2) and alleviates the heavy-tailed sparsity of attention-only selection.

DASH introduces no learnable parameters and operates as a lightweight plug-in between modality encoders and the LLM. By reallocating compression capacity according to semantic density and preserving transition-critical tokens, DASH maintains narrative continuity and cross-modal coherence under aggressive pruning. Experiments on AVUT [36], VideoMME [6], and WorldSense [10] with Qwen2.5-Omni (7B and 3B) show that DASH maintains competitive accuracy at 25% token retention, matching or exceeding prior methods operating at 35%, while achieving up to 3.8× prefill speedup and 1.7× end-to-end latency reduction.

Our contributions are threefold.

– We formulate omnimodal token compression as a structure-aware segmentation problem, showing that compression should follow semantic boundaries rather than fixed positional windows.

- We propose DASH, a training-free, audio-anchored framework that detects semantic discontinuities in embedding space and uses them as soft temporal priors for video segmentation and density-aware compression.
- We demonstrate on three audio-video benchmarks that structure-aware compression enables stable performance at substantially lower token retention (25%) than prior methods (35%), achieving significant speedups with negligible overhead.

2 Related Work

2.1 Token Compression for Multimodal LLMs

Token compression reduces input tokens to alleviate the computational bottleneck of LLM inference, exploiting redundancy in multimodal inputs. For *image* inputs, FastV [2] performs training-free inference acceleration by monitoring attention patterns at a designated LLM layer and pruning uninformative visual tokens during prefill. LLaVA-PruMerge [23] selects representative tokens via adaptive clustering and absorbs the rest through weighted averaging. ToMe [1] introduces a training-free bipartite matching algorithm that progressively merges similar token pairs across ViT layers. Other methods use attention-based selection [35], progressive layer-wise reduction [33, 37], or learned projectors [14]. For *video* inputs, DyCoke [31] proposes a two-stage pipeline separating temporal and spatial redundancy, while others apply spatiotemporal merging [24, 26] or attention-based pruning [11, 27] (see [25] for a survey). However, these methods operate within single modalities and ignore cross-modal correlations. OmniZip [32] pioneers *omnimodal* compression using audio attention to guide video pruning with interleaved spatio-temporal compression. We improve upon OmniZip with dynamic semantic chunking, audio-guided temporal priors for video segmentation, and tri-signal fusion scoring.

2.2 Dynamic Chunking and Boundary Detection

Traditional sequence processing relies on fixed-size segmentation: subword tokenizers (*e.g.*, BPE, WordPiece) split text at predetermined granularity, and token compression methods uniformly group frames into equal-sized windows. This uniform strategy ignores semantic structure—fragmenting coherent units across arbitrary boundaries and failing to adapt to information density. Dynamic chunking addresses this by adapting segment lengths to semantic structure, yielding variable-length chunks that respect natural boundaries. MEGABYTE [38] models byte sequences with fixed-size patches in a multiscale architecture, while Late Chunking [9] segments text for contextual retrieval. H-Net [12] learns dynamic chunking boundaries end-to-end using cosine-similarity-based routing between adjacent representations. MMHNet [29] further applies hierarchical routing and dynamic chunking to long-form video-to-audio generation. In video and audio, boundary detection has been studied for shot/scene segmentation [21, 22], action localization [39], and token pruning [15, 17], but these typically require

task-specific supervision or operate on raw signals. Inspired by H-Net's cosine-similarity routing for dynamic chunking, we use adjacent-representation similarity as a training-free boundary cue for audio tokens in OmniLLMs, and extend it to audio-guided video segmentation by projecting audio boundaries onto video indices as soft temporal priors.

3 Method

We present DASH, a training-free framework that operationalizes the principle of structure-aware compression through three key innovations. As illustrated in Fig. 1, given an input video with audio, DASH first detects semantic boundaries in the audio stream via cosine-similarity-based boundary detection (Sec. 3.2). It then projects these boundaries onto video token indices through temporal ratio mapping to form dynamic audio-guided segments (Sec. 3.3), providing a soft cross-modal structural prior without learned parameters. Finally, it scores tokens within each segment via tri-signal fusion for selective retention (Sec. 3.4). The retained tokens are passed to the LLM for inference.

3.1 Problem Formulation

Given an OmniLLM with audio encoder $\mathcal{E}_a$ and video encoder $\mathcal{E}_v$, we obtain audio token sequence $\mathbf{A} = \{a_t\}_{t=1}^{N_a} \in \mathbb{R}^{N_a \times D}$ and video token sequence $\mathbf{V} = \{v_t\}_{t=1}^{N_v} \in \mathbb{R}^{N_v \times D}$, where D is the embedding dimension. For video, we have $N_v = F \times K$ tokens, where F is the number of frames and K is the number of tokens per frame.

Static grouping in prior work. Existing methods (*e.g.*, OmniZip [32]) divide video into fixed groups of 4 frames ($G_v = 4K$ tokens) and audio into fixed groups of 50 tokens ($G_a = 50$), regardless of semantic content. This rigid partitioning treats information-dense segments (rapid dialogue, complex actions) and information-sparse ones (silence, static scenes) identically.

Our goal. We aim to replace static grouping with dynamic semantic chunking that adapts to content structure, and replace single-signal token selection with multi-signal fusion scoring that captures structural, content, and model-perceived importance jointly.

3.2 Dynamic Semantic Chunking (Segment-Level)

Rather than imposing fixed boundaries every G_a tokens, we let the audio content itself determine where segmentation should occur. This design choice directly addresses the core limitation of prior work: *positional partitioning treats information-dense and information-sparse segments identically.* Our approach is motivated by a key property of audio encoder embeddings: within a semantically coherent segment—such as a continuous sentence or a single topic—adjacent tokens encode similar contextual information, producing high cosine similarity. Conversely, at semantic transitions—sentence pauses, topic shifts, or speaker

changes—the embedding space shifts abruptly, causing a sharp similarity drop. This property provides a training-free signal for boundary detection that is both reliable and computationally efficient. Inspired by H-Net's [12] cosine-similarity routing for dynamic chunking, we adapt adjacent-representation similarity to audio tokens as a training-free boundary cue.

Boundary probability computation. For the audio token sequence $\mathbf{A} = \{a_t\}_{t=1}^{N_a}$, we compute the boundary probability at position t as:

$$\text{sim}_t = \frac{\langle a_{t-1}, a_t \rangle}{\|a_{t-1}\| \cdot \|a_t\|}, \tag{1}$$

$$p_t^{\text{boundary}} = \text{clip}\left(\frac{1 - \text{sim}_t}{2}, 0, 1\right), \tag{2}$$

where $p_t^{\text{boundary}} \in [0, 1]$ is high when adjacent tokens are dissimilar (semantic discontinuity) and low when they are similar (semantic continuity). The first token is always treated as a boundary ($p_1^{\text{boundary}} = 1$).

Boundary detection with minimum chunk constraint. A boundary is detected at position t when the cosine similarity drops below threshold τ_a (default 0.4) *and* the distance from the last boundary exceeds a minimum chunk size $C_{\min}$ (default 30 tokens, ≈1 second of audio):

$$m_t^{\text{boundary}} = \mathbb{I}(\text{sim}_t < \tau_a) \cdot \mathbb{I}(t - t_{\text{last}} \geq C_{\min}), \tag{3}$$

where t_{last} is the position of the most recent boundary. The minimum chunk constraint is essential: without it, noise-induced similarity fluctuations would produce excessively short chunks (2–3 tokens) that are too small for meaningful importance comparison. The default $C_{\min} = 30$ corresponds to approximately 1 second of audio, ensuring each chunk spans at least one prosodic unit while still capturing sentence-level pauses and topic shifts. This produces a set of audio boundary positions $\mathcal{B}_a = \{b_0^a, b_1^a, \ldots, b_S^a\}$.

The resulting variable-length chunks ensure that tokens within each chunk share coherent semantics for fairer importance comparison, and naturally adapt to information density.

3.3 Audio-Guided Video Segmentation (Cross-Modal Level)

A key observation in audio-guided video understanding is that *speech is a strong semantic carrier*: audio boundaries (sentence pauses, topic transitions) often provide useful cues for visual semantic transitions (scene changes, action shifts). We do not assume exact audio-visual boundary coincidence. Instead, since audio and video tokens are temporally co-registered in the OmniLLM's time-window structure, temporal position provides a practical proxy for transferring audio-derived structure to video tokens. We therefore use audio boundaries as a soft temporal prior for video segmentation. This replaces rigid fixed-step grouping with content-aware segments without adding learned parameters.

Temporal ratio mapping. Given audio boundaries $\mathcal{B}_a$ detected on N_a audio tokens, we project them onto video token indices via linear scaling:

$$b_i^v = \left\lfloor b_i^a \cdot \frac{N_v}{N_a} \right\rfloor, \quad i = 0, 1, \ldots, S, \tag{4}$$

where N_v is the total number of video tokens. The projected boundaries $\mathcal{B}_v = \{b_0^v, b_1^v, \ldots, b_S^v\}$ are deduplicated, sorted, and clamped to $[0, N_v]$, with $b_0^v = 0$ and $b_S^v = N_v$ enforced. The resulting segments are audio-guided rather than hard visual cuts: they encourage video compression to follow audio-derived temporal structure, while the final video token mask is still computed from visual embeddings within each segment.

Strength-based boundary refinement. The projected boundaries $\mathcal{B}_v$ may produce video segments shorter than the minimum required by interleaved spatial-temporal compression ($2K$ tokens, i.e., two frames). To resolve this, we apply a greedy refinement: each inner boundary b_i^v is assigned a strength $p_{b_i^a}^{\text{boundary}}$ from its corresponding audio position. We sort inner boundaries by strength in descending order and greedily insert them into the final set $\mathcal{B}_v^*$, initialized as $\{0, N_v\}$. A boundary is accepted only if both adjacent segments exceed $2K$ tokens:

$$b_i^v \in \mathcal{B}_v^* \iff (b_i^v - b_{\text{left}}) \geq 2K \ \wedge \ (b_{\text{right}} - b_i^v) \geq 2K, \tag{5}$$

where b_{left} and b_{right} are the nearest existing boundaries in $\mathcal{B}_v^*$. By prioritizing the strongest audio-indicated boundaries (highest p^{boundary}) rather than arbitrarily dropping boundaries, this greedy strategy preserves the most reliable temporal priors while ensuring all segments are large enough for the interleaved spatial-temporal compression that follows. DASH compresses each segment independently from the original encoder embeddings; compressed outputs from earlier segments are not used to determine later boundaries or later token features.

3.4 Tri-Signal Fusion Token Scoring (Token-Level)

Existing methods select audio tokens solely by attention scores from the audio encoder. However, as shown in Fig. 2 (a), attention distributions are extremely sparse—importance concentrates on a handful of tokens while the vast majority receive near-zero scores. This sparsity is problematic for compression: a single signal cannot capture the full spectrum of token importance. Structurally critical tokens at semantic boundaries and content-distinctive tokens carrying unique information are discarded if they happen to fall outside the attention spotlight. The consequence is that aggressive compression (e.g., 25% retention) based on attention alone destroys narrative continuity by removing transition anchors. To address this fundamental limitation, we propose a tri-signal fusion mechanism that combines three complementary importance signals, each capturing a different aspect of token importance: structural criticality, content distinctiveness, and model-perceived salience.

Signal 1: Boundary probability ($w_b = 0.4$). Tokens at semantic boundaries are content transition points and should be prioritized. We normalize the

Fig. 2: Token scoring and selection comparison. (a) Attention score alone is extremely sparse, concentrating on a few tokens. (b) Tri-signal combined score produces a balanced importance landscape. (c) Selection comparison: green = rescued by fusion (631), blue = selected by both (866), pink = attention-only tokens replaced (631). Fusion rescues ∼42% of retained tokens that attention alone would discard.

boundary probability from Eq. (2)

$$s_t^{\text{bnd}} = \frac{p_t^{\text{boundary}}}{\max_j p_j^{\text{boundary}} + \epsilon}. \tag{6}$$

Signal 2: Probabilistic density-based uniqueness ($w_u = 0.3$). Inspired by the density-peak clustering (DPC) principle used in VidCom [5, 19], we measure the distinctiveness of each token. Unlike vanilla DPC which uses a hard distance cutoff for density estimation [5], we propose a multi-scale Gaussian kernel to capture token importance across various feature granularities, providing a continuous and robust density approximation.

To enhance robustness, we first perform **low-variance channel selection**. Given token features $\mathbf{A} \in \mathbb{R}^{N_a \times D}$, we compute per-channel variance $\sigma_d^2 = \text{Var}([\mathbf{A}]_{:,d})$ and retain the bottom-$\lfloor D/2 \rfloor$ channels by variance, yielding $\tilde{\mathbf{A}} \in \mathbb{R}^{N_a \times D/2}$. This preprocessing step ensures that our uniqueness score is computed based on stable semantic dimensions rather than noisy, high-variance channels that may contain transient artifacts.

We then compute the ℓ_2-normalized features $\hat{\mathbf{A}} = \text{normalize}(\tilde{\mathbf{A}})$ and the global center $\mathbf{c} = \frac{1}{N_a} \sum_t \hat{a}_t$. To approximate the local density ρ defined in DPC-KNN [5] in a continuous space, we define the multi-scale Gaussian similarity:

$$g_t = \sum_{\alpha \in \mathcal{A}} \exp\left(-\frac{\|\hat{a}_t - \mathbf{c}\|^2}{2\alpha} \right), \tag{7}$$

where $\mathcal{A} = \{0.125, 0.25, 0.5, 1.0, 2.0\}$ are multi-scale bandwidth parameters that enable the kernel to capture both fine-grained and coarse-grained feature variations. The uniqueness score is:

$$s_t^{\text{uniq}} = 1 - \frac{g_t}{\max_j g_j + \epsilon}. \tag{8}$$

Signal 3: Attention score ($w_a = 0.3$). We use the attention scores from the audio encoder, normalized to $[0, 1]$:

$$s_t^{\text{attn}} = \frac{\text{attn}_t}{\max_j \text{attn}_j + \epsilon}. \tag{9}$$

Fusion and selection. The final importance score is the weighted sum:

$$s_t = w_b \cdot s_t^{\text{bnd}} + w_u \cdot s_t^{\text{uniq}} + w_a \cdot s_t^{\text{attn}}, \tag{10}$$

where $w_b = 0.4$, $w_u = 0.3$, $w_a = 0.3$. We select the top-N_{keep} tokens by s_t as the retained audio tokens, where $N_{\text{keep}} = \lfloor (1 - \rho_a) \cdot N_a \rfloor$.

The three signals capture orthogonal importance dimensions: boundary probability measures *structural* importance at transition points, density-based uniqueness measures *content* distinctiveness, and attention reflects *model-perceived* importance. A boundary token may receive low attention but high boundary probability; fusion ensures such tokens are not overlooked. We set $w_b = 0.4$ slightly higher because structural boundaries are critical under aggressive compression. When boundary information is unavailable (*e.g.*, very short sequences), the method falls back to attention-only selection.

3.5 Boundary-Aware Video Compression

Given the audio-guided video segments, we perform adaptive compression within each segment. The key idea is that not all segments deserve equal compression: segments associated with information-dense audio (rapid speech, complex narration) should retain more video tokens to preserve the rich visual context, while segments associated with silence or ambient noise can be compressed more aggressively.

Audio-guided adaptive compression ratio. We use the audio retention rate as a proxy for segment-level information density. For each video segment s associated with audio interval $[b_{s-1}^a, b_s^a)$, we compute the audio retention rate $\bar{m}_a^{(s)}$ (fraction of audio tokens retained in that segment by tri-signal fusion). The video compression ratio is then adapted:

$$\rho_v^{(s)} = \rho_v + \lambda_r(0.5 - \bar{m}_a^{(s)}), \quad \rho_v^{(s)} \in [0.1, 0.95], \tag{11}$$

where ρ_v is the base video compression ratio and $\lambda_r = 0.1$ controls adaptation strength. Segments with high audio retention (semantically important) receive lower compression; segments with low audio retention receive higher compression.

Boundary-neighborhood retention. For frames near projected boundary positions, we mildly increase the retention ratio:

$$r_f^{\text{boundary}} = r_s + (1 - r_s) \cdot 0.3 \cdot p_f^{\text{boundary}}, \tag{12}$$

where $r_s = 1 - \rho_v^{(s)}$ is the base retention ratio. This is a conservative safeguard, not a hard assumption that the projected audio boundary is an exact visual cut. It preserves transition-neighboring visual context while leaving the final video token choice to visual-feature-based pruning.

Interleaved spatial-temporal pruning. Within each segment, we adopt the interleaved spatial-temporal compression (ISTC) strategy [32] with the adaptive retention ratio: **Even frames**: Spatial pruning via DPC-KNN [5] that removes tokens with high local density (spatially redundant). **Odd frames**: Temporal pruning that removes tokens most similar to the previous frame (temporally redundant).

4 Experiments

4.1 Evaluation Setups and Implementation Details

Benchmarks. Following OmniZip [32], we evaluate on established audio-video understanding benchmarks: AVUT [36], VideoMME [6], and WorldSense [10]. AVUT is an audio-centric video understanding benchmark focusing on six tasks: event localization (EL), object matching (OM), OCR matching (OR), information extraction (IE), content counting (CC), and character matching (CM). VideoMME is widely used for video-understanding evaluations where including audio can improve accuracy. WorldSense assesses models' ability to understand audio and video jointly across eight domains.

Comparison methods. Given the absence of token pruning methods specifically designed for the omnimodal setting, we follow OmniZip and select representative prior methods from single-modal domains for adaptation: FastV [2] performs training-free inference-time pruning by utilizing the attention score matrix of the L-th layer; DyCoke [31] applies its TTM module to both video and audio tokens; Random pruning serves as a control group. We also compare directly with OmniZip [32]. For fair comparison, we reproduce results for OmniZip and DASH, while results for Random, FastV, and DyCoke are taken from OmniZip [32].

Implementation details. We implement DASH on Qwen2.5-Omni (7B and 3B) models [34] using NVIDIA H20 (96GB) GPUs. We use the overall FLOPs ratio as the metric to ensure fair comparison across methods. For video input, we cap the maximum number of frames at 768 for VideoMME and 128 for other datasets. Each time window contains 50 audio tokens and 288 video tokens. For DASH-specific hyperparameters, we set $\tau_a = 0.4$, $C_{\min} = 30$, tri-signal weights $w_b = 0.4$, $w_u = 0.3$, $w_a = 0.3$, and channel selection ratio 0.5. The reported retention ratios (e.g., 25%, 35%) represent target compression levels; in practice, actual per-sample retention may deviate from these targets due to

Table 1: Comparison of different methods on omnimodal (audio & video) QA benchmarks. Norm. Avg. is the mean of per-benchmark normalized scores, where the baseline accuracy is 100%. "-" indicates OOM error. †FastV Norm. Avg. is computed from AVUT only. Best result among token pruning methods is in **bold**, second best is underlined.

Method	Retained Ratio	FLOPs Ratio	AVUT EL	OR	OM	IE	CC	CM	Avg.	VideoMME wo	Norm. Avg.
Qwen2.5-Omni-7B											
Full Tokens	100%	100%	38.2	67.8	59.6	85.6	44.1	66.7	64.5	66.0	100%
Random	40%	34%	31.7	58.5	53.3	74.9	**43.2**	59.0	56.9	65.0	93.4%
FastV	35%	42%	24.1	60.7	54.3	81.6	40.7	58.3	57.8	-	89.6%†
DyCoke (V&A)	35%	29%	32.9	62.1	54.9	74.5	39.0	58.3	57.4	65.2	93.9%
OmniZip	35%	29%	33.3	**67.2**	54.6	84.7	38.4	61.4	60.6	66.0	97.0%
DASH (ours)	35%	29%	32.0	62.1	**58.9**	85.5	40.4	**64.2**	**61.5**	**66.7**	**98.2%**
DASH (ours)	25%	20%	**36.0**	64.7	56.0	**86.3**	36.5	60.8	60.9	66.0	97.2%
Qwen2.5-Omni-3B											
Full Tokens	100%	100%	32.9	65.3	58.4	85.0	44.1	62.6	62.2	62.6	100%
Random	40%	31%	28.2	60.8	54.9	73.1	42.3	61.6	57.5	60.6	94.6%
FastV	35%	37%	24.2	60.8	54.3	81.6	40.7	58.3	57.7	-	92.8%†
DyCoke (V&A)	35%	26%	**32.9**	62.1	54.9	74.5	38.9	58.3	57.4	61.0	94.9%
OmniZip	35%	26%	28.2	58.6	57.8	80.9	41.5	**62.1**	58.7	61.9	96.6%
DASH (ours)	35%	26%	**32.9**	61.7	55.6	**83.4**	**42.4**	60.0	59.9	**62.6**	**98.2%**
DASH (ours)	25%	18%	29.0	**65.0**	**59.5**	81.1	39.1	58.4	58.8	61.7	96.6%

variations in dataset characteristics and individual sample properties (e.g., audio-video token distribution, content complexity). For all experiments, we leverage FlashAttention to reduce memory usage. Evaluation uses deterministic decoding and rule-based multiple-choice parsing.

4.2 Main Results

We evaluate on Qwen2.5-Omni at two parameter scales (7B and 3B). Following OmniZip, we report baselines at 35% token retention. DASH is evaluated at both 35% and 25% retention to demonstrate its ability to maintain accuracy under more aggressive compression. For VideoMME, we use LMMs-Eval [41] for evaluation. Following OmniZip, results in Tab. 1 are normalized with the baseline model's accuracy set to 100%.

Comparison with state-of-the-art methods. Tab. 1 compares DASH with existing methods. The key finding is that **structure-aware compression enables stable performance at substantially lower token retention**: DASH at 25% retention achieves 60.9% AVUT average on the 7B model, competitive with OmniZip at 35% (60.6%), despite using only 20% FLOPs. This 10-percentage-point reduction in token retention with maintained accuracy directly validates our core claim that semantic structure, not positional regularity, should guide compression decisions. On VideoMME, DASH at 25% matches the full-

Table 2: Comparison of different methods on the WorldSense benchmark. FLOPs calculation considers only multimodal tokens from audio and video inputs. Best result among token pruning methods is in **bold**, second best is underlined.

Method	Retained Ratio	FLOPs (T)	Tech & Science	Culture & Politics	Daily Life	Film & TV	Performance	Games	Sports	Music	Avg.
Qwen2.5-Omni-7B											
Full Tokens	100%	73.2	52.4	50.1	48.5	44.6	43.8	41.6	41.6	47.3	46.8
Random	55%	35.5	47.1	47.0	44.4	41.2	40.0	40.1	40.1	46.3	43.6
FastV	50%	39.3	48.8	47.4	44.2	**44.1**	41.2	38.3	40.0	**46.6**	44.3
DyCoke (V&A)	50%	31.9	48.4	**49.9**	46.7	41.4	39.9	**40.8**	40.2	46.5	44.6
OmniZip	35%	21.4	48.8	49.5	**47.1**	40.6	40.4	**40.8**	40.7	45.6	44.7
DASH (ours)	25%	14.9	**50.0**	49.2	47.0	39.3	**41.6**	39.9	**41.6**	45.8	**44.9**
Qwen2.5-Omni-3B											
Full Tokens	100%	37.4	51.5	50.8	45.0	45.4	43.8	42.5	44.2	46.1	46.4
Random	55%	17.0	48.2	46.3	40.7	41.4	38.6	40.0	41.8	43.4	42.8
FastV	50%	18.2	50.0	**50.5**	44.1	43.0	40.5	41.6	41.8	42.1	44.4
DyCoke (V&A)	50%	15.1	48.1	48.5	42.3	43.3	39.7	**43.4**	42.1	43.0	44.0
OmniZip	35%	9.9	48.4	49.2	41.6	**44.3**	**41.2**	40.8	**43.3**	**43.6**	44.1
DASH (ours)	25%	6.7	**50.2**	47.9	**44.7**	43.3	**41.2**	42.5	40.9	**43.6**	**44.6**

token baseline (66.0%), confirming that content-aware chunking preserves information that fixed-size grouping destroys.

Methods ignoring cross-modal structure (Random, FastV) degrade significantly, confirming that temporal window structure is critical for audio-video understanding. The Norm. Avg. column highlights the overall advantage: DASH at 35% achieves 98.2% on both model scales, surpassing OmniZip (97.0% on 7B, 96.6% on 3B). The consistency across the 7B and 3B variants suggests that the gains are robust across model scales within the same OmniLLM family.

Tab. 2 presents per-domain results on WorldSense. DASH at 25% retention achieves 44.9% average accuracy on the 7B model, matching or surpassing OmniZip's 44.7% at 35% retention while consuming only 14.9T FLOPs versus 21.4T—a 30% computational saving with improved accuracy. On the 3B model, DASH (44.6%) likewise outperforms OmniZip (44.1%) with 32% fewer FLOPs. Across both scales, DASH leads in domains such as Tech & Science and Sports, while remaining competitive in others, suggesting that dynamic semantic chunking generalizes well to diverse audio-video content types.

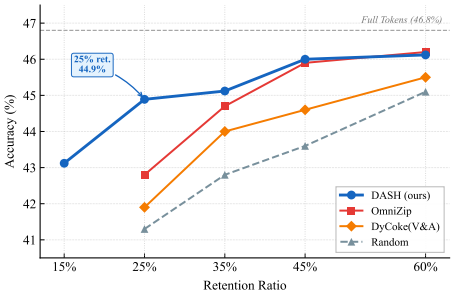

Fig. 3: Accuracy vs. retention ratio on WorldSense (Qwen2.5-Omni-7B).

In Fig. 3, we further illustrate the accuracy-retention trade-off of different methods across retention ratios from 15% to 60%. DASH maintains a clear advantage at low-to-mid retention, with its 25% point (44.9%) already exceeding

Table 3: Inference efficiency on WorldSense.

(a) Qwen2.5-Omni-7B

Method	Mem. ↓	Prefill ↓	Acc. ↑	Latency ↓
Full	35G	1.0×	46.8	1.0×
FastV	OOM	–	–	–
DyCoke	31G	1.6×	44.6	1.2×
OmniZip	25G	3.4×	44.7	1.4×
Ours	26G	3.5×	44.9	1.7×

(b) Qwen2.5-Omni-3B

Method	Mem. ↓	Prefill ↓	Acc. ↑	Latency ↓
Full	25G	1.0×	46.4	1.0×
FastV	45G	1.2×	44.4	1.1×
DyCoke	20G	1.5×	44.0	1.2×
OmniZip	16G	3.3×	44.1	1.3×
Ours	16G	3.8×	44.6	1.4×

OmniZip at 35% (44.7%) and DyCoke at 45% (44.6%). On WorldSense, DASH at 25% retention reaches the accuracy level that prior methods achieve only around 35–45% retention. As retention increases, all methods converge toward the Full Tokens ceiling, indicating that the benefit of content-aware compression is most pronounced under aggressive pruning—precisely the regime where efficiency gains matter most for practical deployment.

4.3 Efficiency Analysis

Tab. 3 reports inference efficiency on WorldSense. At 25% retention, DASH achieves 3.5× prefill speedup on the 7B model and 3.8× on the 3B model, surpassing OmniZip at 35% (3.4× and 3.3×) on both scales. End-to-end latency is also reduced to 1.7× (7B) and 1.4× (3B), while maintaining competitive accuracy. Critically, the overhead of boundary detection and tri-signal scoring is negligible (<40ms), as these operations involve only cosine similarity and element-wise computations on the already-extracted token embeddings. This confirms that structure-aware compression adds minimal computational cost while delivering substantial efficiency gains—a favorable trade-off for practical deployment.

4.4 Ablation Studies

All ablation experiments are conducted on Qwen2.5-Omni-3B at 25% retention ratio on the WorldSense benchmark.

Component ablation and sensitivity analysis. Tab. 4 progressively adds each DASH component. Starting from the baseline with static grouping and attention-only selection, we first add tri-signal fusion alone (Static+TSF), then dynamic chunking with attention-only selection (DSC+ADVS), and finally the full DASH combining all three components. Fig. 4 shows the sensitivity of w_b.

From Tab. 4, both Static+TSF and DSC+ADVS independently improve over OmniZip at 35% (44.1%→44.4%), confirming that tri-signal fusion and dynamic chunking provide complementary gains of comparable magnitude; combining them in Full DASH further lifts accuracy to 44.6%. This additive improvement pattern validates our design principle: structure-aware segmentation and multi-signal importance estimation address orthogonal limitations of prior work.

Boundary detection algorithm comparison. We evaluate four similarity metrics for audio boundary detection (Tab. 5).

Configuration	Acc. (%)
OmniZip (35% ret.)	44.1
Static + TSF	44.4
DSC+ADVS (attn-only)	44.4
Full DASH	**44.6**

Table 4: Ablation on Qwen2.5-Omni-3B (WorldSense, 25% ret.). OmniZip at 35% retention is the reference baseline.

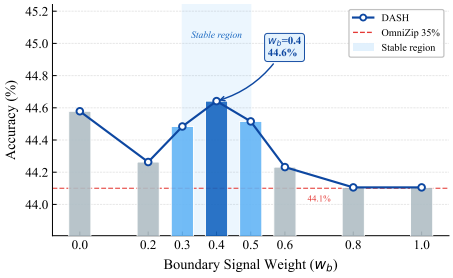

Fig. 4: Sensitivity of w_b on WorldSense (3B, 25% retention).

Cosine similarity achieves the highest accuracy (44.6%), while dot product (43.6%), change rate (44.0%), and random boundaries (43.8%) perform worse. The gap between cosine and dot product confirms the importance of scale invariance in handling feature magnitude variations across audio segments. The consistent advantage over random boundaries indicates that the boundary detection module captures meaningful semantic transitions in audio-visual sequences.

Similarity Method	Acc. (%)
Random	43.8
Dot Product	43.6
Change Rate	44.0
Cosine	**44.6**

Table 5: Boundary detection algorithm comparison on WorldSense (3B, 25% ret.).

From Fig. 4, performance is stable for $w_b \in [0.3, 0.5]$, peaking at w_b=0.4, and degrades for $w_b > 0.5$. The degradation beyond 0.5 suggests that over-emphasizing boundary probability at the expense of content distinctiveness and attention leads to suboptimal selection, validating our default weights.

5 Conclusion

We presented DASH, a training-free framework for structure-aware token compression in omnimodal large language models. Rather than treating multimodal tokens as uniformly structured sequences, DASH aligns compression with the intrinsic semantic organization of audio-visual signals. By using audio embeddings as a semantic anchor, DASH detects boundary candidates that approximate the piecewise structure of multimodal sequences and uses them as soft temporal priors for video compression. This design preserves transition-critical information while reducing redundant regions, enabling aggressive compression without disrupting cross-modal coherence. Experiments on AVUT, VideoMME, and WorldSense show that DASH maintains competitive accuracy at substantially lower token retention while significantly improving inference efficiency. These results highlight the importance of aligning compression with semantic structure for efficient omnimodal reasoning.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62506235.

References

1. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your ViT but faster. In: International Conference on Learning Representations (ICLR) (2023)
2. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: 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 (ECCV). pp. 19–35 (2024)
3. Chen, X., Tao, K., Shao, K., Wang, H.: StreamingTOM: Streaming token compression for efficient video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24675–24685 (2026)
4. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24185–24198 (2024)
5. Du, M., Ding, S., Jia, H.: Study on density peaks clustering based on k-nearest neighbors and principal component analysis. Knowledge-Based Systems **99**, 135–145 (2016)
6. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: 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 (CVPR). pp. 24108–24118 (2025)
7. Fu, C., Lin, H., Long, Z., Shen, Y., Dai, Y., Zhao, M., Zhang, Y.F., Dong, S., Li, Y., Wang, X., Cao, H., Yin, D., Ma, L., Zheng, X., Ji, R., Wu, Y., He, R., Shan, C., Sun, X.: VITA: Towards open-source interactive omni multimodal LLM. arXiv preprint arXiv:2408.05211 (2024)
8. Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
9. Günther, M., Mohr, I., Williams, D.J., Wang, B., Xiao, H.: Late chunking: Contextual chunk embeddings using long-context embedding models. arXiv preprint arXiv:2409.04701 (2024)
10. Hong, J., Yan, S., Cai, J., Jiang, X., Hu, Y., Xie, W.: WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs. In: International Conference on Learning Representations (ICLR) (2026)
11. Huang, X., Zhou, H., Han, K.: PruneVid: Visual token pruning for efficient video large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19959–19973 (2025)

12. Hwang, S., Wang, B., Gu, A.: Dynamic chunking for end-to-end hierarchical sequence modeling. arXiv preprint arXiv:2507.07955 (2025)
13. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: VideoChat: Chat-centric video understanding. Science China Information Sciences **68**(10), 200102 (2025)
14. Li, W., Yuan, Y., Liu, J., Tang, D., Wang, S., Qin, J., Zhu, J., Zhang, L.: TokenPacker: Efficient visual projector for multimodal LLM. International Journal of Computer Vision **133**(10), 6794–6812 (2025)
15. Li, Y., Wu, Y., Li, J., Liu, S.: Accelerating transducers through adjacent token merging. In: Proc. Interspeech 2023. pp. 1379–1383 (2023)
16. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-LLaVA: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5971–5984 (2024)
17. Lin, Y., Fu, Y., Zhang, J., Liu, Y., Zhang, J., Sun, J., Li, H., Chen, Y.: Speech-Prune: Context-aware token pruning for speech information retrieval. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2025)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916 (2023)
19. Liu, X., Wang, Y., Ma, J., Zhang, L.: Video compression commander: Plug-and-play inference acceleration for video large language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 1910–1924 (2025)
20. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
21. Rao, A., Wang, J., Xu, L., Jiang, X., Huang, Q., Zhou, B., Lin, D.: A unified framework for shot type classification based on subject centric lens. In: European Conference on Computer Vision (ECCV). pp. 17–34 (2020)
22. Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multi-modal movie scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10146–10155 (2020)
23. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22857–22867 (2025)
24. Shao, K., Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: HoliTom: Holistic token merging for fast video large language models. arXiv preprint arXiv:2505.21334 (2025)
25. Shao, K., Tao, K., Zhang, K., Feng, S., Cai, M., Shang, Y., You, H., Qin, C., Sui, Y., Wang, H.: A survey of token compression for efficient multimodal large language models. Transactions on Machine Learning Research (2026)
26. Shen, L., Gong, G., He, T., Zhang, Y., Liu, P., Zhao, S., Ding, G.: FastVID: Dynamic density pruning for fast video large language models. In: Advances in Neural Information Processing Systems. vol. 38 (2025)
27. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., Liu, Z., Xu, H., Kim, H.J., Soran, B., Krishnamoorthi, R., Elhoseiny, M., Chandra, V.: LongVU: Spatiotemporal adaptive compression for long video-language understanding. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 54582–54599. PMLR (2025)

28. Shu, F., Zhang, L., Jiang, H., Xie, C.: Audio-visual LLM for video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 4305–4314 (2025)
29. Simon, C., Ishii, M., Wang, W.Y., Saito, K., Hayakawa, A., Shim, D., Zhong, Z., Cui, S., Shibuya, T., Takahashi, S., Mitsufuji, Y.: Echoes over time: Unlocking length generalization in video-to-audio generation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15840–15849 (2026)
30. Tang, C., Li, Y., Yang, Y., Zhuang, J., Sun, G., Li, W., Ma, Z., Zhang, C.: video-SALMONN 2: Caption-enhanced audio-visual large language models. arXiv preprint arXiv:2506.15220 (2025)
31. Tao, K., Qin, C., You, H., Sui, Y., Wang, H.: DyCoke: Dynamic compression of tokens for fast video large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18992–19001 (2025)
32. Tao, K., Shao, K., Yu, B., Wang, W., Liu, J., Wang, H.: 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 (CVPR). pp. 17682–17692 (2026)
33. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., Lin, D.: Conical visual concentration for efficient large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14593–14603 (2025)
34. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., Lin, J.: Qwen2.5-Omni technical report. arXiv preprint arXiv:2503.20215 (2025)
35. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: VisionZip: Longer is better but not necessary in vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19792–19802 (2025)
36. Yang, Y., Zhuang, J., Sun, G., Tang, C., Li, Y., Li, P., Jiang, Y., Li, W., Ma, Z., Zhang, C.: Audio-centric video understanding benchmark without text shortcut. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 6569–6587 (2025)
37. Ye, W., Wu, Q., Lin, W., Zhou, Y.: Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 22128–22136 (2025)
38. Yu, L., Simig, D., Flaherty, C., Aghajanyan, A., Zettlemoyer, L., Lewis, M.: MEGABYTE: Predicting million-byte sequences with multiscale transformers. Advances in Neural Information Processing Systems **36**, 78808–78823 (2023)
39. Zhang, C., Wu, J., Li, Y.: ActionFormer: Localizing moments of actions with transformers. In: European Conference on Computer Vision (ECCV). pp. 492–510 (2022)
40. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553 (2023)
41. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., Liu, Z.: LMMs-Eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 881–916 (2025)

42. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: LLaVA-Video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)