MoCo: Modality-Aware Decoder-Side Token Compression for Video–Audio OmniLLMs

Anonymous authors
Paper under double-blind review

Abstract

Long audio–video inputs make inference costly for omnimodal large language models (OmniLLMs), motivating token compression that preserves question-relevant evidence. Question-conditioned attention offers a natural signal for selecting which tokens to retain. However, we find that shallow question-conditioned attention favors audio even for visual questions, causing direct cross-modal ranking to allocate less capacity to visual evidence. Allocating audio and video budgets separately improves accuracy without changing token scores or within-modality rankings. Based on these findings, we propose MoCo, a training-free framework for decoder-side token compression that separates modality-budget allocation from question-conditioned token selection. It assigns audio and video shares of a joint token budget, then uses question-conditioned attention to rank tokens within each modality. Within a single prefill pass, MoCo prunes audio early to reduce computation and delays video selection to exploit more informative attention scores at later shallow layers. Across four OmniLLMs, MoCo achieves higher average performance than OmniZip at all three evaluated retention ratios. MoCo preserves 98.7% and 97.3% of full-token performance on Qwen2.5-Omni-7B and 3B, respectively, while retaining only 25% of multimodal tokens. At matched FLOPs, MoCo preserves up to 3.1 percentage points more of full-token performance than OmniZip on Qwen2.5-Omni-7B.

1 Introduction

Omnimodal large language models (OmniLLMs) unify audio and video understanding, enabling a single model to interpret visual content, acoustic events, and their relationships (Xu et al., 2025a; Ye et al., 2025; Fu et al., 2024; Ge et al., 2025). This capability requires processing a combined sequence of audio and video tokens through a shared decoder. As input duration increases, the resulting sequence imposes growing computational and memory costs, making efficient inference increasingly challenging. Token compression aims to reduce this burden while preserving the multimodal evidence needed for accurate understanding.

Recent OmniLLM compressors reduce audio and video tokens before they enter the shared decoder (Tao et al., 2025b; Ding et al., 2026; Li & Huang, 2026). These methods exploit within-modality redundancy and cross-modal guidance. This reduces computation throughout the decoder, but makes retention decisions without access to decoder attention from question tokens. Question-guided visual token selection provides a natural starting point for using this decoder signal (Zhang et al., 2025b). Joint audio–video compression must decide both which tokens to retain and how to divide the budget between modalities. Using attention for both assumes that cross-modal score differences reflect the evidence required by the question. Since pruning removes tokens before later layers can refine their scores, we argue: **How can we use question-conditioned attention to compress audio–video inputs without allowing modality bias to dictate what is retained?**

Our analysis shows that directly ranking audio and video tokens by shallow attention can yield a suboptimal modality split. Extending OmniZip's observation of higher audio attention (Tao et al., 2025b), we find that shallow question-conditioned attention favors audio even for visual questions (Fig. 1). Compared with shared Top-K selection, separate modality budgets improve accuracy using the same token scores and within-modality rankings (Fig. 2). This controlled comparison supports separating modality-budget allocation from within-modality token ranking. Video selection benefits

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.

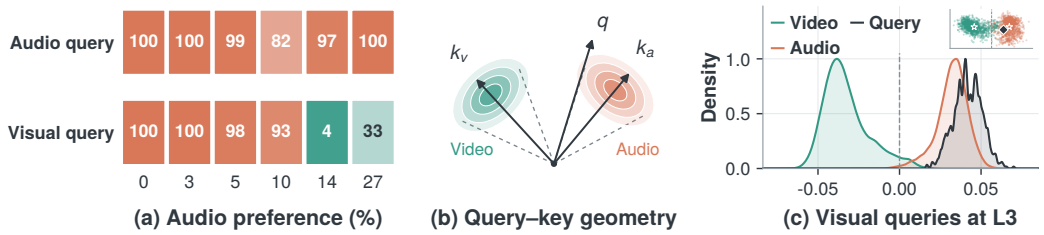


Figure 1: **Shallow attention favors audio even for visual questions. (a)** Percentage of 100 paired DailyOmni clips with higher per-token attention to audio than video. Both question types favor audio at shallow layers. **(b)** Schematic: text query q aligns more closely with the audio-key centroid k_a than the video-key centroid k_v. **(c)** At L3, all 645 query positions from 30 video-essential prompts have positive cosine gaps, indicating closer audio-centroid alignment.

from later shallow readouts, whereas audio selection is comparatively stable across early layers, motivating separate scoring layers for the two modalities (Fig. 3).

These findings motivate separating modality-budget allocation from question-conditioned token selection, so that shallow audio preference does not crowd out visual evidence. We propose MoCo, a training-free framework for decoder-side token compression that implements this design through audio–video budget allocation, modality-specific scoring, and staged token pruning. MoCo assigns each modality a share of the joint token budget, then uses question-conditioned attention to select tokens within each modality's budget. Audio tokens are pruned at an earlier decoder layer to reduce subsequent computation, while video candidates remain available for scoring and pruning at a later shallow layer, reflecting their greater sensitivity to the scoring layer. Attention normalization and a lightweight temporal-coverage bonus refine selection within the allocated token budgets. The entire process runs in a single prefill pass, without retraining or an additional forward pass.

Our contributions are threefold:

- **A failure mode of cross-modal attention ranking.** We identify a shallow audio preference that persists across question types and biases the audio–video split among retained tokens under direct top-K selection. Controlling modality token budgets improves accuracy without changing token scores or within-modality rankings.
- **Joint budgeting and staged selection.** We propose a training-free framework for decoder-side token compression that allocates audio and video token budgets separately from question-conditioned token selection, scoring and pruning the two modalities at different early decoder layers within a single prefill pass.
- **Strong accuracy preservation at reduced token budgets.** MoCo preserves 98.7% and 97.3% of full-token performance on Qwen2.5-Omni-7B and 3B, respectively, while retaining only 25% of multimodal tokens. On Qwen2.5-Omni-7B, it achieves higher average performance than OmniZip at five of six matched FLOPs budgets.

2 Related Work

Omni-Modal Large Language Models. Progress in vision–language (Li et al., 2023; Lin et al., 2024; Li et al., 2024a) and audio–language modeling (Deshmukh et al., 2023; Tang et al., 2024; Chu et al., 2024) has enabled language models to interpret increasingly diverse inputs. Audio–video models combine visual and acoustic evidence for joint understanding (Cheng et al., 2024; Tang et al., 2025; Ge et al., 2025). Recent omnimodal architectures integrate visual and acoustic inputs into unified language models (Fu et al., 2024; Li et al., 2024b; Ye et al., 2025; Xu et al., 2025a;b). Processing both streams produces long token sequences that increase attention computation and memory use. Token compression aims to reduce this cost while retaining the visual and auditory evidence needed to answer the question.

Token Pruning for OmniLLMs. Multimodal token compression reduces sequence length through attention-guided pruning (Chen et al., 2024; Zhang et al., 2025b; Lin et al., 2025; Xing et al., 2025),

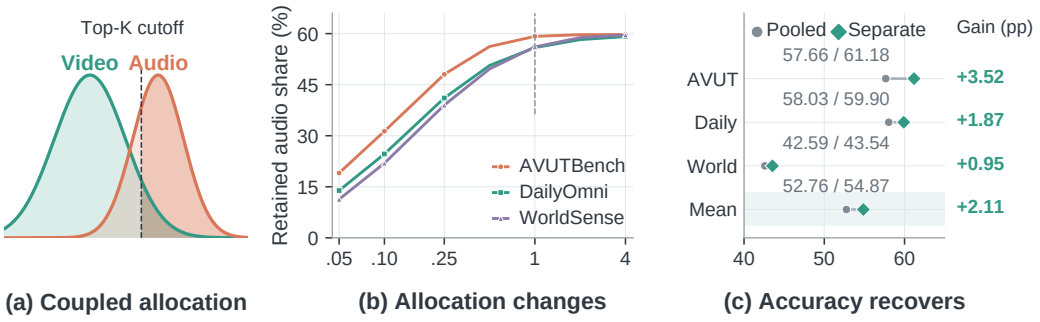


Figure 2: **Capacity calibration improves accuracy without changing local rankings. (a,b)** Shared-Top-K converts cross-modal score scale into retained capacity: rescaling audio scores changes the modality split while preserving both within-modality rankings. Horizontal axes show shared attention score in (a) and the audio-score multiplier in (b). **(c)** Separate audio/video budgets improve mean accuracy by 2.11 percentage points with the same scores, total budget, and pruning step after L5; the horizontal axis shows accuracy (%). The allocation settings are given in Section 5.2. Both diagnostics use Qwen2.5-Omni-7B at 25% retention.

representative selection and merging (Shang et al., 2025; Yang et al., 2025a; Wen et al., 2025; Zhang et al., 2025a), and spatiotemporal or instruction-guided reduction (Shen et al., 2025; Fan et al., 2026; Tao et al., 2025a; Liu et al., 2025). For joint audio–video inputs, cross-modal compression methods use encoder attention, learned cross-modal selectors, or audio-derived temporal structure to reduce tokens before the shared decoder (Tao et al., 2025b; Ding et al., 2026; Li & Huang, 2026). Related work also explores adaptive KV-cache allocation and cross-modal calibration (Jiang et al., 2025). Our method separates audio–video budget allocation from question-conditioned token selection, reading attention at modality-specific depths within the shared decoder.

Modality competition has been studied in joint training, motivating methods that balance optimization and strengthen modality representations (Huang et al., 2022; Peng et al., 2022; He et al., 2026; Li et al., 2026). Related analyses examine information flow across decoder layers (Zhang et al., 2025c; Song et al., 2026; Lu et al., 2026), modality preferences across prompts and depths (Chen et al., 2025; Selvakumar et al., 2026), and token removal after information transfer (Suharitdamrong et al., 2026). Audio-dominant attention has also motivated audio-guided compression (Tao et al., 2025b). We examine whether this audio preference tracks the evidence required by the question and how it affects token retention. Our analysis motivates separating modality-budget allocation from within-modality ranking and reading audio and video attention at different decoder depths.

3 Making Decoder Attention Useful for Compression

Question-conditioned decoder attention can identify relevant tokens, but joint audio–video compression must also decide how many tokens to retain from each modality. We examine whether cross-modal attention scores provide a reliable basis for this budget decision and how the choice of scoring layer affects selection. Direct shared-Top-K serves as a diagnostic rule that uses the same scores for both ranking and allocation.

One attention score, two decisions. Shared-Top-K gives the same score two roles: ordering tokens *within a modality* and determining *how many tokens each modality retains* (Fig. 2a). A decomposition of attention (Vaswani et al., 2017) makes the distinction explicit. For a given query and head, let $k_i^{(m)} = c_m + \epsilon_i^{(m)}$, where $c_m = \frac{1}{|m|}\sum_{i\in m} k_i^{(m)}$ is the per-clip modality centroid and $\epsilon_i^{(m)}$ is the within-modality residual. With the attention scaling absorbed into q, the log attention mass on modality m decomposes as

$$\log \text{mass}_m(q) = \underbrace{q^\top c_m}_{\text{centroid alignment}} + \underbrace{\log \sum_{i\in m} \exp\left(q^\top \epsilon_i^{(m)}\right)}_{\text{residual aggregate}} - \log Z, \tag{1}$$

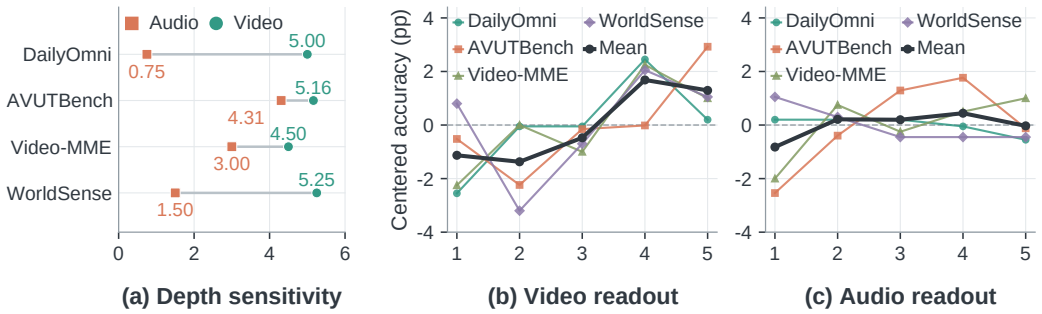


Figure 3: **Video selection is more sensitive to scoring depth than audio selection. (a)** Accuracy ranges (max–min, pp) across L1–L5. **(b)** Later video readouts improve mean accuracy. **(c)** Audio readouts show smaller variation. In (b,c), the x-axis is scoring layer; profiles average over the other modality's layers and are centered at their own mean. Thick curves denote benchmark means. Qwen2.5-Omni-7B, 25% retention, 30% audio budget share.

where $Z = \sum_{m'} \sum_{j \in m'} \exp(q^\top k_j^{(m')})$ is the shared attention normalizer over all token groups, including text. For audio-video comparison, Z cancels:

$$\log \frac{\text{mass}_a(q)}{\text{mass}_v(q)} = q^\top (c_a - c_v) + \left[\log \sum_{i \in a} \exp(q^\top \epsilon_i^{(a)}) - \log \sum_{j \in v} \exp(q^\top \epsilon_j^{(v)}) \right]. \tag{2}$$

A modality-common logit offset leaves within-modality ordering unchanged while altering cross-modal attention mass. This distinguishes relative token ranking from modality-budget allocation at the level of a single query and head. Observation 2 tests the corresponding allocation decision using the aggregated scores employed for compression.

Observation 1: shallow attention favors audio even under visual questions. If shallow cross-modal scores tracked the evidence needed to answer a question, visual questions would be expected to shift attention toward video. On 100 paired clips, however, shallow attention favors audio per token under both audio and visual questions; visual questions shift attention toward video only at greater depths (Fig. 1). At L3, all 645 examined question tokens from 30 video-essential prompts also satisfy $\cos(q, c_a) > \cos(q, c_v)$. This directional alignment is consistent with the observed preference, although cosine similarity alone does not determine attention mass. Question conditioning therefore does not ensure that shallow cross-modal score magnitudes reflect the modality required for answering.

Audio-content and token-distance controls show that the preference depends on both the input and token position (Appendix C).

Under shared-Top-K, this persistent audio preference can reduce the number of retained video tokens even when the question requires visual evidence. We next test whether controlling the audio–video split improves accuracy with the token scores unchanged.

Observation 2: separate modality budgets improve selection with the same scores. Multiplying all audio scores by a positive scalar preserves both within-modality rankings but changes the audio share retained by shared-Top-K (Fig. 2a,b). This diagnostic shows that retained composition can change without changing which tokens rank highest within either modality. To test whether composition matters for accuracy, we compare raw shared-Top-K with separate audio and video budgets, holding the raw scores, candidate tokens, total budget, and pruning step unchanged. Both variants use the same modality-specific readouts and a single pruning step, without attention normalization or a coverage bonus; the configuration is specified in Section 5.2.

Across 6,103 aligned examples from three benchmarks, separate audio and video budgets raise mean accuracy from 52.76% to 54.87%, with improvements on all three (Fig. 2c). Both variants use identical scores, candidate tokens, total budgets, and pruning steps, isolating the contribution of budget allocation. These results support separating modality-budget allocation from within-modality token ranking.

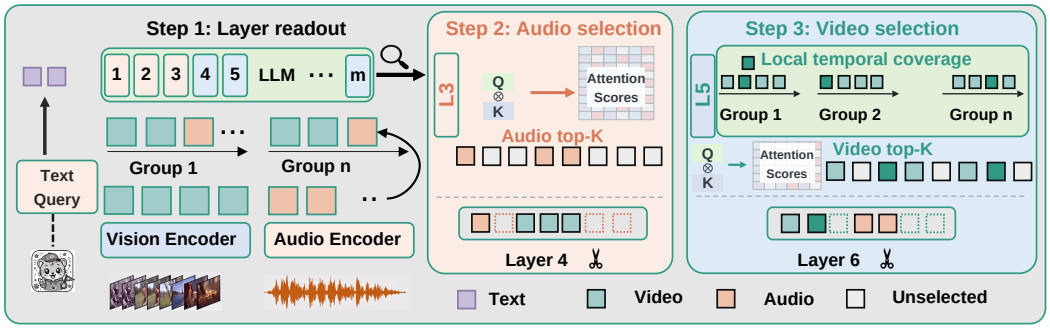


Figure 4: **MoCo pipeline.** Audio tokens are scored and pruned at ℓ_a. The decoder continues with retained audio, all video candidates, and text until ℓ_v, where video tokens are scored and pruned. The remaining layers process the retained audio and video tokens together with all text tokens in their original order, within a single prefill pass.

Observation 3: audio and video differ in readout-depth sensitivity. Separating budget allocation from ranking leaves a second decision: at which decoder layer should each modality be scored? With the audio–video allocation held unchanged and total retention at 25%, we vary audio and video scoring layers over L1–L5 on four benchmarks (Fig. 3; Appendix D.2). Averaging over the other modality's scoring layers gives a marginal accuracy profile for each modality. The mean profile range is 4.98 percentage points for video versus 2.39 for audio, with greater video sensitivity on all four benchmarks.

At this operating point, video selection benefits more from delaying the attention readout, while audio selection is less sensitive to scoring depth. These profiles motivate separate scoring layers; Section 5.2 evaluates the resulting readout and pruning configuration.

From diagnostics to compression design. These findings motivate three steps: split the joint audio–video budget, read question-conditioned attention at modality-specific layers, and select tokens within each modality's budget. MoCo sets the budget split before prefill, prunes audio after the earlier readout, and retains video candidates until the later readout (Sections 4.1–4.2). Attention normalization and temporal coverage refine selection within the assigned budgets.

4 Method: MoCo

MoCo compresses audio–video tokens through modality-specific budgets and staged token selection (Fig. 4). Motivated by the shallow audio preference observed in Section 3, MoCo allocates separate budgets to audio and video, then selects tokens within each modality using question-conditioned attention. Audio tokens are pruned at an earlier shallow layer, while video candidates remain available for a later readout. This schedule reduces computation early while exploiting the benefit of later video scoring. All text tokens are preserved, and both stages run in a single prefill pass without retraining.

4.1 Audio–Video Budget Allocation

MoCo controls the retained audio–video composition through explicit modality budgets. Let I_a and I_v be the audio and video token sets entering the decoder, with $n_a = |I_a|$ and $n_v = |I_v|$. Given a retention ratio $\rho \in [0, 1]$, we first determine the total multimodal budget:

$$K_{\text{mm}} = \min\{n_a + n_v, \text{round}(\rho(n_a + n_v))\}. \tag{3}$$

The audio share $\alpha \in [0, 1]$ then allocates this budget between modalities:

$$\bar{K}_a = \text{round}(\alpha K_{\text{mm}}), \qquad \bar{K}_v = K_{\text{mm}} - \bar{K}_a, \tag{4}$$

Here, round denotes nearest-integer rounding.

When a modality has fewer tokens than its nominal budget, unused slots are transferred to the other modality:

$$K_a = \min\{n_a, \bar{K}_a + \max(0, \bar{K}_v - n_v)\}, \qquad K_v = K_{\mathrm{mm}} - K_a. \tag{5}$$

This yields $K_a + K_v = K_{\mathrm{mm}}$ with $K_a \le n_a$ and $K_v \le n_v$. Both budgets are computed from the original sequence before prefill and carried through the two pruning stages, so removing audio tokens leaves the assigned video capacity intact. Question-conditioned decoder attention then selects tokens within each modality's budget.

4.2 Question-Conditioned Token Selection

Token selection must balance the information available for scoring against the cost of retaining candidates in the decoder. The depth profiles in Section 3 show that video selection benefits from later shallow readouts, whereas audio selection is comparatively stable across early layers. Selecting both modalities early forgoes the benefit of later video readouts, while delaying both retains audio candidates through additional decoder layers. Motivated by this asymmetry, MoCo scores and prunes audio at layer ℓ_a, then scores and prunes video at layer ℓ_v, with $\ell_a < \ell_v$. The full input is available through ℓ_a; thereafter, audio pruning reduces the sequence processed by subsequent layers while all video candidates remain available until ℓ_v.

At each scoring layer, we use attention from the question tokens to rank candidates in the target modality. Let $\boldsymbol{Q}$ denote the question tokens used as probes and $A_h^{(\ell)}(q,i)$ the attention probability from question token q to candidate i at layer ℓ and head h, with $\boldsymbol{H}$ heads. Directly averaging these probabilities weights question–head rows according to their total attention mass on the modality. To aggregate their relative preferences among candidates, we normalize each row within the target modality before averaging. This reduces the influence of modality-level mass differences across rows. For $m \in \{a, v\}$, the resulting score is

$$\widetilde{A}_{m,h}^{(\ell)}(q,i) = \frac{A_h^{(\ell)}(q,i)}{\sum_{j \in I_m} A_h^{(\ell)}(q,j) + \varepsilon}, \qquad s_i^{(m)} = \frac{1}{|Q|H} \sum_{q \in Q} \sum_{h=1}^{H} \widetilde{A}_{m,h}^{(\ell_m)}(q,i). \tag{6}$$

Here ε is a numerical stabilizer. The score $s_i^{(m)}$ measures the relative attention received by token i within its modality across question tokens and heads. Normalization changes the contribution of each row to this ranking; it does not change K_a or K_v.

We apply this scoring rule as the forward pass reaches each selected layer. After layer ℓ_a, we retain the K_a highest-scoring audio tokens:

$$S_a \in \operatorname*{arg\,max}_{S \subseteq I_a,\ |S| = K_a} \sum_{i \in S} s_i^{(a)}. \tag{7}$$

The other audio tokens are removed. The same forward pass then continues with retained audio S_a, all video candidates I_v, and all text tokens. At layer ℓ_v, Eq. (6) is evaluated for video in this continuing forward pass. The later video readout therefore reflects the decoder context after audio pruning, with retained audio and all text tokens still present.

Within the video budget, we also account for the temporal distribution of selected tokens. Our concentration diagnostic shows that score-only selection is more temporally concentrated for video than for audio (Appendix D.1). We therefore refine video selection with a diminishing bonus for regions with fewer selected tokens, encouraging temporal spread without imposing an equal quota on every region. Let $c(j)$ denote the temporal partition of video token j and let $r_g = |\{u \in S_v : c(u) = g\}|$ count the tokens already selected from partition g. Starting from $S_v = \emptyset$, each step selects

$$j^\star = \arg \max_{j \in I_v \setminus S_v} \left[s_j^{(v)} + \sqrt{\frac{\lambda_c}{1 + r_{c(j)}}} \right], \tag{8}$$

Table 1: **Comparison on Qwen2.5-Omni-7B at matched token retention.** The **best** result among token-pruning methods within each retained-ratio block (25%/35%/45%) is in bold, and the second-best is underlined. FastV† is evaluated on H100 due to OOM on Ada6000 under the long multimodal context; its average excludes Video-MME, denoted by "-".

METHOD	SETTINGS		AVUT	DailyOmni	WorldSense	Video-MME	NORM. AVG.
	RETAINED RATIO	FLOPs RATIO	AVG.	AVG.	AVG.	AVG.	
Qwen2.5-Omni-7B							
Full Tokens	100%	100%	64.5	62.1	46.8	66.0	100%
Random	55%	49.2%	61.0	59.2	43.6	65.4	95.5%
FastV†	50%	54.1%	58.4	59.8	44.3	-	93.8%
OmniZip	45%	39.2%	63.07	59.82	**44.61**	66.3	97.5%
MoCo (Ours)	45%	50.0%	**64.72**	**63.32**	44.00	**66.9**	**99.4%**
OmniZip	35%	29.6%	61.05	59.40	43.98	65.7	96.0%
MoCo (Ours)	35%	42.2%	**64.63**	**62.32**	**45.05**	**66.2**	**99.3%**
OmniZip	25%	20.6%	59.10	57.39	38.18	**66.5**	91.6%
MoCo (Ours)	25%	34.8%	**63.67**	**62.07**	**44.80**	66.2	**98.7%**

adds $j^\star$ to S_v, and updates the corresponding count. The bonus falls as a region receives more tokens, while the attention score continues to guide the choice of candidates. Selection continues until K_v tokens are retained, combining question-conditioned relevance with temporal coverage within the allocated video budget. The weight λ_c controls the bonus; $\lambda_c = 0$ recovers video top-K selection. Temporal partitions and parameter settings are specified in Appendix B and Section 5.

After the second pruning stage, the decoder continues with $S_a \cup S_v$ and all text tokens in their original sequence order. The remaining layers therefore process exactly K_{mm} multimodal tokens.

5 Experiment Results

Benchmarks and baselines. We evaluate Qwen2.5-Omni-7B (Xu et al., 2025a) on AVUT-Bench (Yang et al., 2025b), DailyOmni (Zhou et al., 2025), WorldSense (Hong et al., 2025), and Video-MME (Fu et al., 2025), covering audio-grounded, audio–video joint, real-world omnimodal, and video-centric benchmarks. We compare MoCo with random token retention, FastV (Chen et al., 2024) for decoder-attention-based token pruning, and OmniZip (Tao et al., 2025b) for audio-guided omnimodal compression. We define the retention ratio as the number of retained multimodal tokens divided by that in the uncompressed decoder input. Norm. Avg. is the mean of per-benchmark accuracy ratios to the corresponding full-token model, expressed as a percentage. We additionally evaluate Qwen3-Omni-30B-A3B, Qwen2.5-Omni-3B, and OmniVinci-9B (Ye et al., 2025).

Implementation details. Qwen2.5-Omni-7B is evaluated on the full benchmark test sets using LMMs-Eval (Zhang et al., 2024). We evaluate Qwen2.5-Omni-7B/3B and OmniVinci-9B on NVIDIA RTX 6000 Ada GPUs with 48 GB memory. For the main comparisons, we retain 25%, 35%, or 45% of multimodal tokens through decoder-prefill pruning while preserving all text tokens. We use greedy decoding with FlashAttention-2 (Dao, 2023) enabled.

Hyperparameter configuration. We configure the audio budget share α, readout layers (ℓ_a, ℓ_v), and temporal coverage weight λ_c on a development split disjoint from evaluation. For each backbone, we reuse the resulting configuration across benchmarks, question types, and retention ratios.

We use $(\ell_a, \ell_v, \alpha, \lambda_c) = (3, 5, 0.30, 0.20)$ for Qwen2.5-Omni-7B and reuse the same configuration on Qwen3-Omni-30B-A3B. Additional implementation and evaluation details are provided in Appendix B, with allocation, readout-depth, and selection ablations in Appendices D.3, D.2, and D.1, respectively.

5.1 Main Results

On Qwen2.5-Omni-7B, MoCo retains only 25% of multimodal tokens while preserving 98.7% of full-token performance, compared with 91.6% for OmniZip (Table 1). MoCo outperforms OmniZip in 10 of 12 benchmark–retention comparisons. At 25% retention, accuracy gains over OmniZip are

4.57, 4.68, and 6.62 percentage points on AVUTBench, DailyOmni, and WorldSense, respectively. Video-MME accuracy is 66.2%, compared with 66.0% for the full-token model. On Qwen3-Omni-30B-A3B (Xu et al., 2025b), MoCo uses the same configuration as Qwen2.5-Omni-7B and achieves higher average performance than OmniZip at all three retention ratios (Table 2). At 25% retention, MoCo preserves 95.00% of full-token performance, compared with 92.87% for OmniZip, and improves accuracy on all three benchmarks.

Table 2: **Comparison on Qwen3-Omni-30B-A3B.** Accuracy on three benchmarks at matched token retention. The **best** result within each retained-ratio block is in bold. OmniZip† denotes the Qwen3 budget adaptation in Appendix B. Norm. Avg. reports the mean percentage of full-token performance preserved across the three benchmarks.

METHOD	RETAIN	AVUT	DailyOmni	WorldSense	NORM. AVG.
Full Tokens	100%	75.20	71.68	54.38	100.00
OmniZip†	45%	72.03	70.18	**53.81**	97.55
MoCo (Ours)	45%	**73.47**	**71.18**	52.77	**98.01**
OmniZip†	35%	71.68	68.84	**52.81**	96.15
MoCo (Ours)	35%	**73.59**	**70.26**	52.33	**97.37**
OmniZip†	25%	68.51	67.34	50.88	92.87
MoCo (Ours)	25%	**70.99**	**69.42**	**50.98**	**95.00**

MoCo also outperforms FastVID (Shen et al., 2025) and FlashVID (Fan et al., 2026) across all three benchmarks at both retention ratios (Table 3). At 25% and 35% total audio–video retention, MoCo preserves 4.3 and 3.5 percentage points more of full-token performance than the stronger temporal compression baseline, respectively.

Table 3: **Comparison with temporal compression methods at matched total audio–video retention on Qwen2.5-Omni-7B.** Norm. Avg. reports the mean percentage of full-token performance preserved across the three benchmarks. The **best** result within each retention block is bold.

RETAIN	METHOD	AVUT	DailyOmni	WorldSense	NORM. AVG.
25%	FastVID	61.30	57.14	42.91	92.9
	FlashVID	62.14	57.39	43.38	93.8
	MoCo (Ours)	**63.67**	**62.07**	**44.80**	**98.1**
35%	FastVID	63.24	58.39	44.01	95.4
	FlashVID	63.88	57.89	43.60	95.1
	MoCo (Ours)	**64.63**	**62.32**	**45.05**	**98.9**

Additional matched-retention comparisons with DASH (Li & Huang, 2026) are reported in Table 8. Results on Qwen2.5-Omni-3B and OmniVinci-9B are provided in Appendix A.

5.2 Ablation Studies

Separate modality budgets outperform shared Top-K allocation on all three benchmarks, raising mean accuracy from 52.76 to 54.87 (+2.11 percentage points; Table 4a). Both variants use identical raw attention scores, a 25% total retention budget, and a single pruning step at L5, preserving within-modality rankings. MoCo's staged modality-specific readout and pruning improve mean accuracy from 54.38 to 56.85 (+2.47 percentage points) over single-layer readout (Table 4b). Additional scoring and temporal-coverage ablations are provided in Appendix D.1.

5.3 Accuracy–Compute Trade-off

At matched FLOPs, MoCo achieves higher average performance than OmniZip at five of six budgets and higher accuracy in 14 of 18 benchmark comparisons, with one tie (Table 5). At 35% of full-model FLOPs, MoCo preserves 98.1% of full-token performance, compared with 95.0% for OmniZip, while retaining 25% rather than 40.43% of multimodal tokens. Gains in relative performance range from

Table 4: **Capacity allocation and modality-specific readout on Qwen2.5-Omni-7B at 25% retention.** (a) Both variants use identical raw scores and a single L5 mask, with attention normalization and temporal coverage disabled. (b) Single-layer readout is compared with staged A3/V5 readout and pruning under the same modality budgets. Comparisons are made within each block. Avg. is mean accuracy across the three benchmarks.

CONFIGURATION	AVUT	DailyOmni	WorldSense	AVG.
(a) Capacity allocation: same scores and single L5 mask				
Raw shared-Top-K allocation	57.66	58.03	42.59	52.76
Separate modality budgets	**61.18**	**59.90**	**43.54**	**54.87**
(b) Readout configuration				
Single-layer readout	57.47	60.87	**44.80**	54.38
MoCo: modality-specific readout	**63.67**	**62.07**	**44.80**	**56.85**

0.3 to 3.1 percentage points across the five winning budgets; at the lowest budget, MoCo preserves 91.8% of full-token performance, compared with 92.3% for OmniZip.

Table 5: **Accuracy at matched FLOPs on Qwen2.5-Omni-7B.** Accuracy is measured by model evaluation, and decoder FLOPs are computed analytically relative to the full-token model. Norm. Avg. reports the mean percentage of full-token performance preserved across the three benchmarks. The **best** result within each matched FLOPs pair is in bold.

FLOPs	METHOD	RETAIN	AVUT	DailyOmni	WorldSense	NORM. AVG.
100%	Full Tokens	100.00%	64.50	62.10	46.80	100.0
24%	OmniZip	29.19%	**60.01**	**57.89**	42.37	**92.3**
	MoCo	10.00%	59.10	57.56	**42.62**	91.8
28%	OmniZip	32.91%	61.24	59.23	43.00	94.1
	MoCo	15.00%	**61.79**	**59.90**	**43.85**	**95.3**
31%	OmniZip	36.65%	62.16	59.73	43.85	95.4
	MoCo	20.00%	**63.01**	**60.48**	**43.98**	**96.4**
35%	OmniZip	40.43%	61.47	59.48	43.98	95.0
	MoCo	25.00%	**63.67**	**62.07**	**44.80**	**98.1**
42%	OmniZip	48.08%	62.68	59.90	**45.05**	96.6
	MoCo	35.00%	**64.63**	**62.32**	**45.05**	**98.9**
50%	OmniZip	55.84%	64.34	61.74	**45.08**	98.5
	MoCo	45.00%	**64.72**	**63.32**	44.00	**98.8**

6 LIMITATIONS

MoCo targets audio–video OmniLLMs that process both modalities within a shared decoder. Extending modality-budget allocation and modality-specific readout to other fusion architectures and additional modalities remains an open direction.

7 CONCLUSION

We proposed MoCo, a training-free framework for decoder-side token compression in audio–video OmniLLMs. MoCo separates cross-modal budget allocation from question-conditioned token selection and prunes each modality at its own shallow readout layer. On Qwen2.5-Omni-7B, MoCo retains 25% of multimodal tokens while preserving 98.7% of full-token performance, compared with 91.6% for OmniZip. The same configuration also achieves higher average performance than OmniZip across all evaluated retention ratios on Qwen3-Omni-30B-A3B. These results establish modality-aware allocation and readout as effective components of omnimodal token compression.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, including drafting, language polishing, and figure preparation. The authors take responsibility for the final content of this paper.

REFERENCES

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 *Computer Vision – ECCV 2024*, pp. 19–35. Springer, 2024. doi: 10.1007/978-3-031-73004-7_2. URL https://doi.org/10.1007/978-3-031-73004-7_2.

Tianle Chen, Chaitanya Chakka, Arjun Reddy Akula, Xavier Thomas, and Deepti Ghadiyaram. Some modalities are more equal than others: Decoding and architecting multimodal integration in MLLMs. *arXiv preprint arXiv:2511.22826*, 2025. doi: 10.48550/arXiv.2511.22826. URL https://arxiv.org/abs/2511.22826.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing spatial-temporal modeling and audio understanding in video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. URL https://arxiv.org/abs/2406.07476.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. URL https://arxiv.org/abs/2407.10759.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 18090–18108, 2023.

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. *arXiv preprint arXiv:2602.04804*, 2026. doi: 10.48550/arXiv.2602.04804. URL https://arxiv.org/abs/2602.04804.

Ziyang Fan, Keyu Chen, Ruilong Xing, Yulin Li, Li Jiang, and Zhuotao Tian. FlashVID: Efficient video large language models via training-free tree-based spatiotemporal token merging. In *The Fourteenth International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2602.08024. URL https://arxiv.org/abs/2602.08024.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. 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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

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.

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30183–30193, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/He_Enhance-then-Balance_Modality_Collaboration_for_Robust_Multimodal_Sentiment_Analysis_CVPR_2026_paper.html.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating real-world omnimodal understanding for multimodal LLMs, 2025. URL https://arxiv.org/abs/2502.04326.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259, 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

Zhonghua Jiang, Kui Chen, Kunxi Li, Keting Yin, Yiyun Zhou, Zhaode Wang, Chengfei Lv, and Shengyu Zhang. AccKV: Towards efficient audio-video llms inference via adaptive-focusing and cross-calibration kv cache optimization. *arXiv preprint arXiv:2511.11106*, 2025. URL https://arxiv.org/abs/2511.11106.

Bingzhou Li and Tao Huang. DASH: Dynamic audio-driven semantic chunking for efficient omnimodal token compression. *arXiv preprint arXiv:2603.15685*, 2026. doi: 10.48550/arXiv.2603.15685. URL https://arxiv.org/abs/2603.15685.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. URL https://arxiv.org/abs/2408.03326.

Hengzhuang Li, Xinsong Zhang, Qiming Peng, Bin Luo, Han Hu, Dengyang Jiang, Han-Jia Ye, Teng Zhang, and Hai Jin. Unleashing the intrinsic visual representation capability of multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1771–1786, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Unleashing_the_Intrinsic_Visual_Representation_Capability_of_Multimodal_Large_Language_CVPR_2026_paper.html.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. URL https://arxiv.org/abs/2305.06355.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-Omni technical report. *arXiv preprint arXiv:2410.08565*, 2024b. URL https://arxiv.org/abs/2410.08565.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024. URL https://aclanthology.org/2024.emnlp-main.342/.

Yueqian Lin, Yuzhe Fu, Jingyang Zhang, Yudong Liu, Jianyi Zhang, Jingwei Sun, Hai Li, and Yiran Chen. SpeechPrune: Context-aware token pruning for speech information retrieval. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 2025. doi: 10.1109/ICME59968.2025.11209113. URL https://ieeexplore.ieee.org/document/11209113.

Zhihang Liu, Chen-Wei Xie, Pandeng Li, Liming Zhao, Longxiang Tang, Yun Zheng, Chuanbin Liu, and Hongtao Xie. Hybrid-level instruction injection for video token compression in multi-modal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.16036.

Haolang Lu, Bolun Chu, WeiYe Fu, Guoshun Nan, Junning Liu, Minghui Pan, Qiankun Li, Yi Yu, Hua Wang, and Kun Wang. Reallocating attention across layers to reduce multimodal hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4157–4167, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_Reallocating_Attention_Across_Layers_to_Reduce_Multimodal_Hallucination_CVPR_2026_paper.html.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Peng_Balanced_Multimodal_Learning_via_On-the-Fly_Gradient_Modulation_CVPR_2022_paper.html.

Ramaneswaran Selvakumar, Kaousheik Jayakumar, S Sakshi, Sreyan Ghosh, Ruohan Gao, and Dinesh Manocha. Do audio-visual large language models really see and hear? *arXiv preprint arXiv:2604.02605*, 2026. URL https://arxiv.org/abs/2604.02605.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Shang_LLaVA-PruMerge_Adaptive_Token_Reduction_for_Efficient_Large_Multimodal_Models_ICCV_2025_paper.html.

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. *Advances in Neural Information Processing Systems*, 2025. doi: 10.48550/arXiv.2503.11187. URL https://arxiv.org/abs/2503.11187.

Shezheng Song, Shasha Li, Shan Zhao, Xiaopeng Li, Qian Wan, Chengyu Wang, Tianwei Yan, Jun Ma, and Jie Yu. Where does vision meet language? understanding and refining visual fusion in MLLMs via contrastive attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10051–10060, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Song_Where_Does_Vision_Meet_Language_Understanding_and_Refining_Visual_Fusion_CVPR_2026_paper.html.

Wish Suharitdamrong, Muhammad Awais, Xiatian Zhu, and Sara Atito. From senses to decisions: The information flow of auditory and visual perception in multimodal llms. *arXiv preprint arXiv:2606.10147*, 2026. URL https://arxiv.org/abs/2606.10147.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/pdf/2310.13289.

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. URL https://arxiv.org/abs/2506.15220.

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, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Tao_DyCoke_Dynamic_Compression_of_Tokens_for_Fast_Video_Large_Language_CVPR_2025_paper.html.

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. *arXiv preprint arXiv:2511.14582*, 2025b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9961–9980, 2025. doi: 10.18653/v1/2025.emnlp-main.505. URL https://aclanthology.org/2025.emnlp-main.505/.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Conical visual concentration for efficient large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14593–14603, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xing_Conical_Visual_Concentration_for_Efficient_Large_Vision-Language_Models_CVPR_2025_paper.html.

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. doi: 10.48550/arXiv.2503.20215. 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. URL https://arxiv.org/abs/2509.17765.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_VisionZip_Longer_is_Better_but_Not_Necessary_in_Vision_Language_CVPR_2025_paper.html.

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. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.333. URL https://aclanthology.org/2025.emnlp-main.333/.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*, 2025.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Reality check on the evaluation of large multimodal models, 2024. URL https://arxiv.org/abs/2407.12772.

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

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. PMLR, 2025b. URL https://proceedings.mlr.press/v267/zhang25s.html.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19781–19791, 2025c. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_Cross-modal_Information_Flow_in_Multimodal_Large_Language_Models_CVPR_2025_paper.html.

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. URL https://arxiv.org/abs/2505.17862.

APPENDIX

A ADDITIONAL EXPERIMENTAL RESULTS

A.1 MOCO ON QWEN2.5-OMNI-3B

On Qwen2.5-Omni-3B, MoCo preserves 97.3% of full-token performance at 25% retention and achieves higher average performance than OmniZip at all three retention ratios (Table 6). It improves 11 of 12 benchmark–ratio comparisons, including all four benchmarks at 25% retention.

Table 6: **Comparison on Qwen2.5-Omni-3B.** The **best** result among token-pruning methods within each retained-ratio block is in bold, and the second-best is underlined. FastV† is evaluated on H100 due to OOM on Ada6000 under the long multimodal context; its average excludes Video-MME, denoted by "-".

METHOD	SETTINGS		AVUT	DailyOmni	WorldSense	Video-MME	NORM. AVG.
	RETAINED RATIO	FLOPs RATIO	AVG.	AVG.	AVG.	AVG.	
Qwen2.5-Omni-3B							
Full Tokens	100%	100%	62.2	60.2	46.4	62.6	100%
Random	55%	46.1%	58.7	57.4	42.8	61.1	94.9%
FastV†	50%	49.2%	55.9	58.0	44.4	-	94.0%
OmniZip	45%	36.1%	61.30	**59.73**	45.20	62.80	98.9%
MoCo (Ours)	45%	48.5%	**62.13**	59.15	**46.45**	**63.21**	**99.8%**
OmniZip	35%	26.8%	60.10	58.06	44.30	62.70	97.2%
MoCo (Ours)	35%	41.0%	**60.52**	**58.15**	**45.55**	**63.37**	**98.3%**
OmniZip	25%	18.3%	58.90	56.47	43.40	62.60	95.5%
MoCo (Ours)	25%	34.1%	**59.74**	**58.40**	**44.24**	**63.01**	**97.3%**

A.2 MOCO ON OMNIVINCI-9B

On OmniVinci-9B, MoCo achieves the highest average performance at all three retention ratios, exceeding the strongest baseline by 2.45, 0.47, and 1.02 percentage points, respectively (Table 7). MoCo leads on DailyOmni at every ratio, while FastV remains stronger on AVUTBench.

Table 7: **Comparison on OmniVinci-9B at matched token retention.** The **best** result among token-pruning methods within each retained-ratio block is in bold.

RETAIN	METHOD	AVUT	DailyOmni	WorldSense	NORM. AVG.
100%	Full Tokens	64.36	66.50	48.23	100.00
25%	FastV	**61.31**	61.29	42.56	91.89
	OmniZip	54.02	58.61	**45.74**	88.97
	DASH	48.39	64.25	44.25	87.85
	MoCo	54.02	**69.32**	**45.74**	**94.34**
35%	FastV	**62.25**	61.29	45.54	94.44
	OmniZip	56.37	62.56	46.24	92.51
	DASH	54.02	67.06	46.74	93.90
	MoCo	53.56	**68.19**	**47.73**	**94.91**
45%	FastV	**62.25**	63.55	**47.53**	96.94
	OmniZip	61.54	63.12	45.74	95.12
	DASH	56.37	67.63	45.25	94.37
	MoCo	56.84	**71.57**	47.24	**97.96**

A.3 COMPARISON WITH DASH

MoCo achieves higher accuracy than DASH in five of the six matched-retention comparisons (Table 8).

Table 8: Comparison with DASH (Li & Huang, 2026) under matched retention. Bold indicates the higher accuracy; Δ is MoCo minus DASH in percentage points.

BENCHMARK	RETAIN	DASH	MoCo	Δ
AVUTBench	25%	60.90	**63.67**	+2.77
AVUTBench	35%	61.50	**64.63**	+3.13
WorldSense	25%	**44.90**	44.80	-0.10
DailyOmni	25%	59.32	**62.07**	+2.75
DailyOmni	35%	60.65	**62.32**	+1.67
DailyOmni	45%	61.49	**63.32**	+1.83

B Implementation and Evaluation Details

We select hyperparameters on approximately 100 randomly sampled development clips. We group video tokens by native temporal boundaries when available and into equal-size token bins otherwise. For OmniVinci-9B, MoCo reads audio and video scores at layers 5 and 7, respectively, with $\alpha = 0.30$; FastV prunes at layer 7.

Video-MME uses no subtitles. FastVID and FlashVID retain all text and audio tokens and allocate the remaining audio–video token budget to video.

Qwen3-Omni protocol. We use the native Qwen3-Omni processor at 2 fps with at most 128 frames and a per-frame pixel cap of $128(2p)^2$ for vision patch size p. Both methods share the question/options prompt and use greedy decoding with a 16-token output limit (unparseable answers count as incorrect). We adapt OmniZip† to enforce the target total audio–video token budget while preserving its scoring and audio-merging rules.

Computational cost. We report estimated decoder FLOPs relative to full-token inference as $[L_p + (L - L_p)\rho(1 - \eta(1 - \rho))]/L$, where L is decoder depth and η is the attention share of decoder computation. Input-pruning methods use $L_p = 0$; MoCo uses $L_p = \max(\ell_a, \ell_v)$, counting full-context computation through the later readout. We use $(L, \eta) = (28, 0.235)$ for Qwen2.5-Omni-7B and $(36, 0.36)$ for 3B.

C Attention Diagnostics

We examine audio-content and token-position effects on 30 paired clips using Qwen2.5-Omni-7B, averaging attention over heads within each layer. Replacing real audio with silence or scale-matched white noise shifts mid-layer attention from audio-favoring to video-favoring, while mismatched real audio produces an intermediate shift. After quantile-binned token-distance matching, attention is approximately balanced at L7 and L10, whereas L14 and L19 retain an audio-favoring residual (Fig. 5).

D Ablation and Sensitivity Analysis

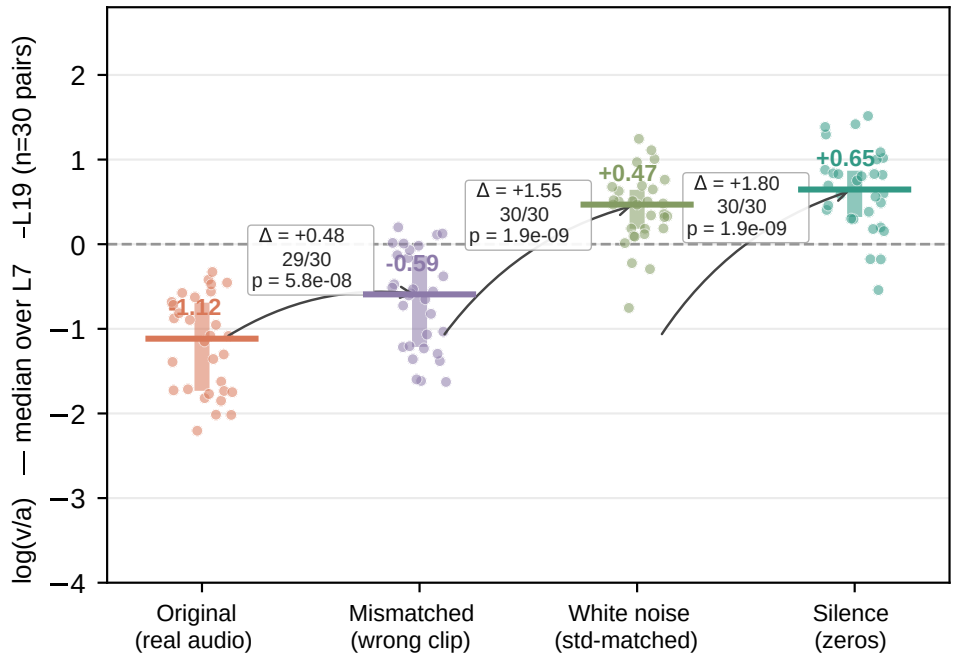
D.1 Component Ablations

For the concentration diagnostic, retained tokens are binned into native 2-second blocks. At $\rho = 0.25$, score-only selection gives audio and video Gini coefficients of 0.06 and 0.33, respectively (lower is more uniform), motivating video-local coverage.

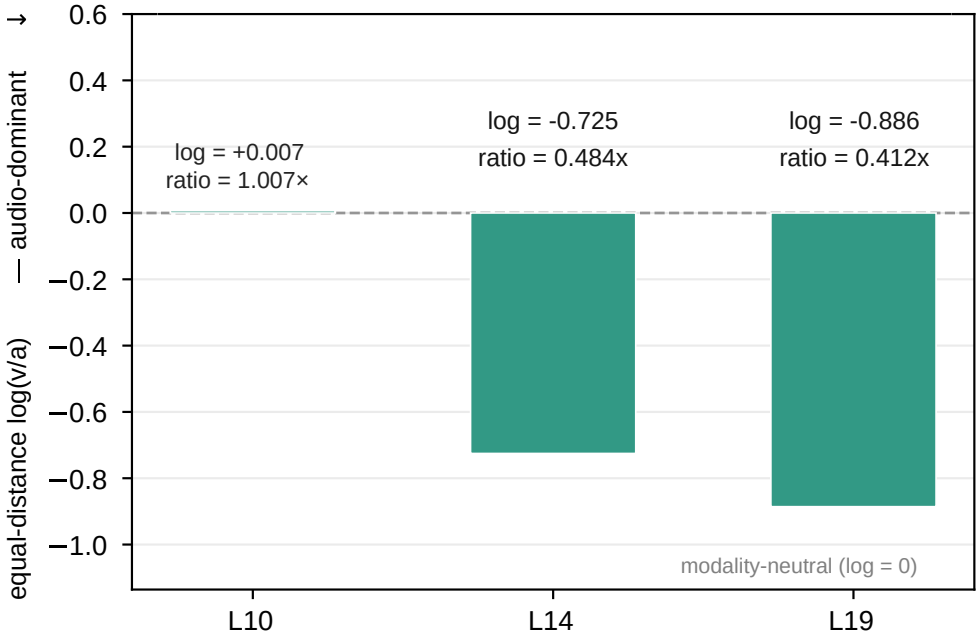
Without temporal coverage, normalizing each attention row within its modality before aggregation raises mean accuracy from 56.29 to 56.54 (+0.25 points). Adding temporal coverage to normalized scoring raises the average to 56.85 (+0.31 points; Table 9).

D.2 Scoring Layer Sensitivity

Figure 6 presents the complete scoring-layer grids underlying Fig. 3 in the main paper.



(a) Audio-content interventions



(b) Distance-matched attention

Figure 5: **Audio-content and token-distance controls on 30 paired clips. (a)** Changes in mid-layer $\log(v/a)$ after replacing audio with silence, white noise, or mismatched audio. **(b)** Audio–video attention comparisons at matched token distances to the query at L10, L14, and L19. Panel (a) displays two-sided sign-test p-values.

Table 9: **Within-modality selection ablations.** Qwen2.5-Omni-7B at 25% retention. Avg. is mean accuracy across three benchmarks.

SCORING	COVERAGE	AVUT	DailyOmni	WorldSense	AVG.
Raw attention	Off	63.50	60.99	44.39	56.29
Modality-normalized	Off	63.33	61.82	44.47	56.54
Modality-normalized	On	**63.67**	**62.07**	**44.80**	**56.85**

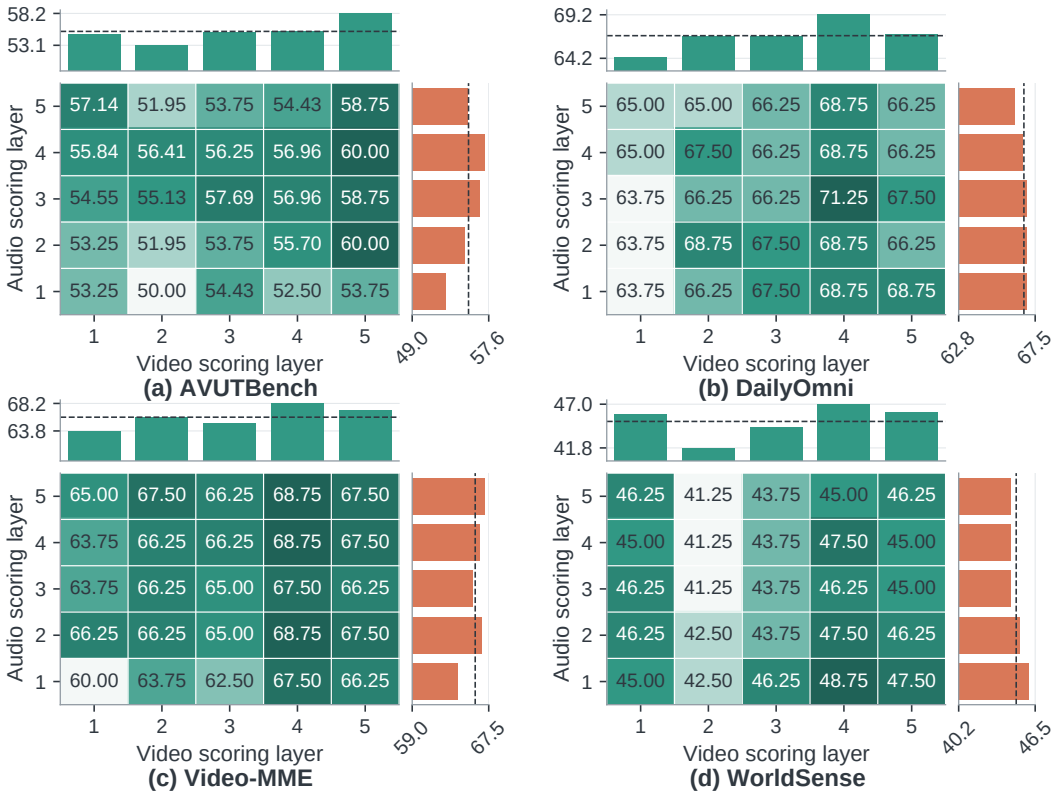


Figure 6: **Modality-specific scoring-layer sensitivity.** Qwen2.5-Omni-7B, prescribed share $\alpha = 0.30$, $\rho = 0.25$, last-four question-token readout, and 100 examples per cell. Rows vary the audio scoring layer and columns vary the video scoring layer (1-based); cells show accuracy (%). Teal top bars and coral right bars show the video- and audio-layer marginal means, respectively; dashed lines mark the corresponding overall means. These grids underlie Fig. 3.

Table 10: **Sensitivity to the audio budget share.**

SETTING	AVUT	DailyOmni	WorldSense	Video-MME	NORM. AVG.
Full tokens	65.83	69.50	44.00	77.00	100.00
$\alpha = 0.10$	63.32	65.00	**44.50**	74.00	96.74
$\alpha = 0.20$	64.82	67.00	44.00	74.50	97.91
$\alpha = 0.30$ (default)	66.50	**68.50**	43.50	74.50	**98.80**
$\alpha = 0.40$	**67.00**	68.00	43.00	**75.00**	98.69
$\alpha = 0.50$	64.00	66.50	42.50	74.00	96.40

D.3 Audio Budget Sensitivity

Table 10 evaluates sensitivity to the retained-audio share across five settings on a 200-example held-out split at 25% retention. The default share $\alpha = 0.30$ achieves the highest average performance, with $\alpha = 0.40$ only 0.11 percentage points lower.

E VISUALIZATION

Figure 7 shows a selected DailyOmni success using Qwen3-Omni-30B-A3B-Instruct. All methods receive the same input and prompt; compressed methods retain 25% of the original audio–video tokens. FastV uses our audio–video adaptation; OmniZip follows Appendix B. Yellow overlays indicate retained visual-token support, with circles and insets identifying the same region across methods. Audio bars show the original RMS envelope; opacity indicates retained-token fractions in uniform index bins and provides approximate temporal localization. OmniZip representatives may contain merged audio information.

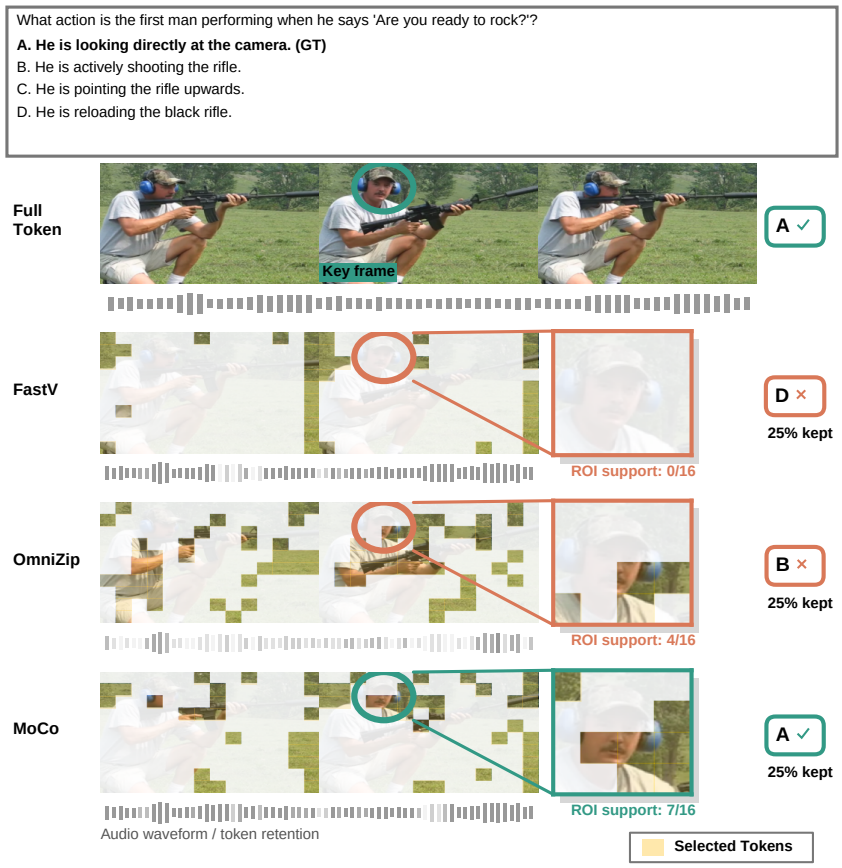


Figure 7: **Recognizing a camera-facing pose during a spoken question.** DailyOmni example 522. Full Token and MoCo answer A correctly; FastV answers D and OmniZip answers B. In the highlighted frame, MoCo, OmniZip, and FastV retain 7, 4, and 0 of the 16 token-cell centers inside the rectangular head-region inset, respectively.