O-MARC: OMNI MEMORY-AUGMENTED COMPRESSION DISTILLATION FOR EFFICIENT VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Omnimodal large language models can reason over audio and video together, but their long input sequences make inference costly. Evaluating this capability is also difficult: existing benchmarks do not specifically isolate audiovisual association in noisy user-generated videos. We introduce UGC-AVQA, a dataset of 1,000 public videos and 4,816 question-answer pairs. An audio removal test selects benchmark questions that a strong model cannot answer from video alone. To reduce inference cost, we propose OMAC, a training-free compression method that preserves visual memory and temporally aligned audio anchors. We then introduce O-MARC, a compression distillation framework that trains models to reason with these compressed inputs. On Qwen2.5-Omni-3B, O-MARC reaches an average score of 45.8 across four benchmarks, compared with 44.1 for full-token inference and 41.0 for OmniZip. OMAC reduces inference latency by 34.6% (a 1.53$\times$ speedup) and GPU memory use by 34.7% relative to full-token inference.

1 INTRODUCTION

Omnimodal large language models jointly process audio, video, and language (Cheng et al., 2024; Xu et al., 2025; Liu et al., 2026). Compared with vision-language models that primarily use visual inputs (Wang et al., 2024; Liu et al., 2023), they can combine visible actions with speech, ambient sounds, and other acoustic cues. This matters for real-world videos, where neither modality alone may explain an event. The computational cost is substantial, however. Audio and video both produce many tokens, and processing them together increases memory use and inference latency, especially for long videos (Shao et al., 2025a;b; Shen et al., 2026).

Token compression reduces this cost, but audiovisual inputs pose a specific difficulty: the two modalities differ in redundancy and temporal structure. Video often contains repeated spatial and temporal content (Wu et al., 2025; Song et al., 2024), whereas a brief sound can carry decisive evidence. The value of a token also depends on the other modality. A sound may need a visible source to be interpreted, and an action may remain ambiguous without its accompanying audio. Compressing the two streams independently can therefore discard evidence needed to connect them (Tao et al., 2025; Ding et al., 2026).

Evaluation presents a related problem. Existing audiovisual benchmarks cover broad omnimodal understanding, long-video reasoning, and temporal alignment, but do not specifically isolate audiovisual association in noisy user-generated videos (Li et al., 2025; Hong et al., 2025; Zhou et al., 2025). VideoMME (Fu et al., 2025), TempCompass (Liu et al., 2024), and LongVideoBench (Wu et al., 2024) are widely used to evaluate video understanding, often with models that rely primarily on visual input. We introduce UGC-AVQA to evaluate questions designed to require both acoustic and visual evidence in public short videos. During benchmark construction, we remove the audio and retain a question only if a strong model fails to answer it. This check reduces the number of questions that can be solved from visual information alone.

For efficient inference, we propose Omni Memory-Augmented Compression (OMAC), a plug-in method inspired by video memory compression, particularly MARC (Wu et al., 2025). OMAC retains compact visual memory and audio anchors, then uses the distribution of visual memory over time to allocate the audio token budget. It requires no changes to the backbone model or additional training.

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.

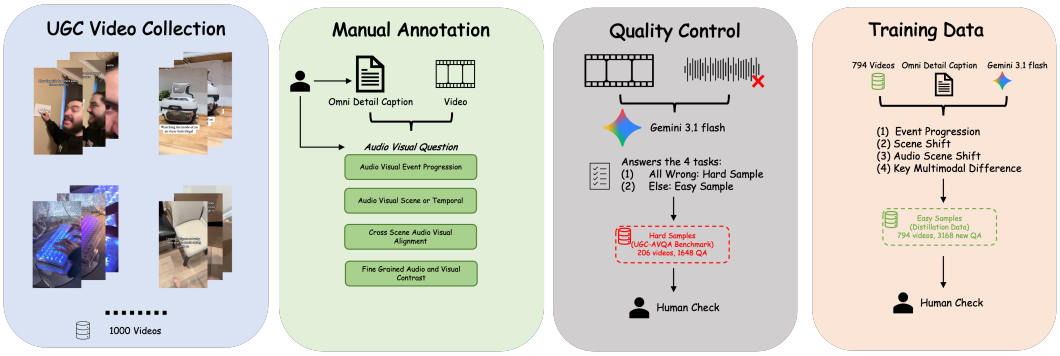


Figure 1: **UGC-AVQA construction.** We collect public videos, annotate detailed captions and audiovisual questions, use an audio removal test to select benchmark samples, and review a subset of benchmark and generated training questions with trained annotators.

We also introduce O-MARC, a compression distillation framework that trains omnimodal models to reason with these compressed inputs. OMAC improves on direct token compression at comparable inference cost across 3B, 7B, and 30B models. O-MARC further improves the 3B model and exceeds the overall average of the training-based baseline when applied to the 7B model.

Our contributions are threefold: UGC-AVQA, a dataset for evaluating audiovisual association in user-generated videos; OMAC, a training-free method for compressing audiovisual inputs; and O-MARC, a distillation framework for learning to reason with compressed multimodal memory.

2 Related Work

2.1 Omnimodal Large Language Models

Multimodal language models have expanded from image understanding (Liu et al., 2023) to video understanding (Li et al., 2024a; Bai et al., 2025) and joint processing of audio and visual inputs (Xu et al., 2025; Cui et al., 2026). Visual input alone does not capture speech, sound events, or the acoustic context of an action. Omnimodal models incorporate these signals during training and inference, allowing them to use audiovisual cues that vision-only models can miss (Fu et al., 2024; Ge et al., 2025; Li et al., 2024b). Systems such as GPT-4o (Hurst et al., 2024), Qwen-Omni (Xu et al., 2025), and Gemini (Comanici et al., 2025) illustrate this shift toward joint audiovisual understanding.

2.2 Token Compression for Omnimodal Models

Long audiovisual token sequences make inference expensive. Compression methods exploit redundancy in multimodal inputs to reduce this cost (Song et al., 2024; Wu et al., 2025; Bolya et al., 2022; Chen et al., 2024; Shang et al., 2025). Most work has focused on visual or textual inputs; joint audiovisual compression must also preserve the evidence linking the two modalities. OmniZip (Tao et al., 2025) uses audio to guide dynamic token compression, while OmniSIFT (Ding et al., 2026) uses modality-asymmetric compression with additional training. These approaches differ in their accuracy, inference cost, and training requirements. Our work examines both training-free memory compression and distillation under compressed inputs, with an emphasis on retaining related acoustic and visual evidence.

3 UGC-AVQA Dataset

We introduce **UGC-AVQA**, a dataset for evaluating fine-grained audiovisual association in user-generated short videos. It combines human-annotated captions and questions with model-assisted training annotations. Trained annotators review 30% of both retained benchmark candidates and generated training examples, correcting them where needed. Questions ask models to relate sounds

Dataset	Videos	QA pairs	Focus
WorldSense (Hong et al., 2025)	1,662	3,172	Real-world omnimodal reasoning
OmniVideoBench (Li et al., 2025)	628	1,000	Long-video reasoning
Daily-Omni (Zhou et al., 2025)	684	1,197	Temporal alignment
UGC-AVQA	1,000	4,816	Audiovisual association in UGC

Table 1: **Dataset comparison.** UGC-AVQA focuses on audiovisual association in public user-generated videos, complementing broader omnimodal and long-video benchmarks.

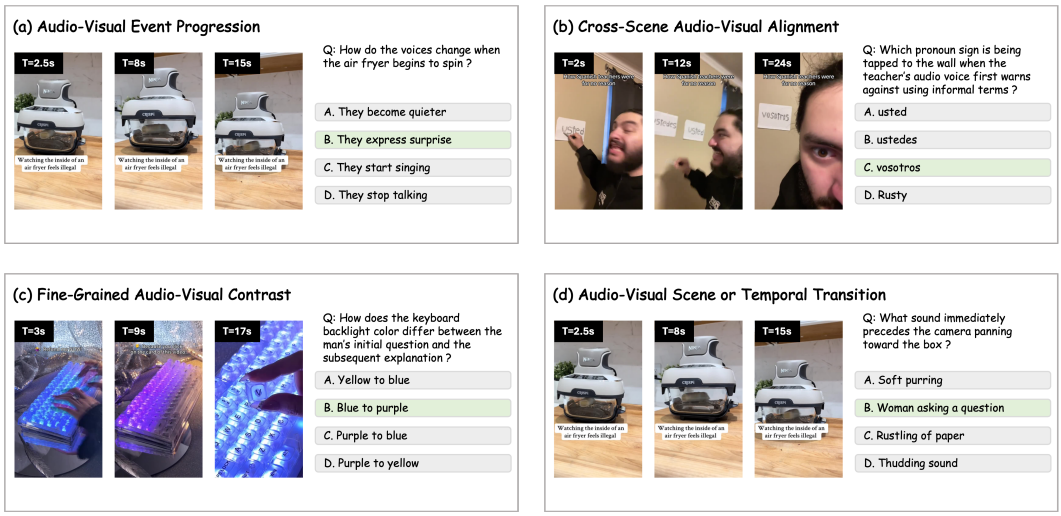


Figure 2: **UGC-AVQA examples.** The four panels illustrate (a) a vocal reaction associated with a visible event, (b) a spoken cue aligned with an on-screen sign, (c) a visual change between two stages of speech, and (d) an acoustic event preceding camera movement. Correct options are highlighted in green.

to visible events, scenes, and temporal changes. Camera motion, background speech, environmental sounds, and abrupt transitions make this setting challenging for compression.

Table 1 compares UGC-AVQA with recent audiovisual benchmarks. WorldSense covers broad real-world omnimodal understanding, OmniVideoBench emphasizes long-video reasoning, and Daily-Omni includes temporal alignment. UGC-AVQA focuses on audiovisual dependence through event progression, scene or temporal transitions, cross-scene alignment, and fine-grained contrast. Appendix A.3 compares the question categories in detail.

3.1 Collection, Annotation, and Data Release

We collect 1,000 videos from publicly accessible short-video platforms. We release the original video URLs, source metadata, detailed captions, and QA annotations, without redistributing raw video files. This preserves provenance and keeps access tied to the original platforms. The dataset and annotations are intended for audiovisual reasoning research, subject to the access conditions of the source platforms.

Each video receives a detailed multimodal caption covering visible entities, scene transitions, actions, acoustic events, and temporal relations. Human annotators then write multiple-choice questions in four categories: **Audio-Visual Event Progression**, **Audio-Visual Scene or Temporal Transition**, **Cross-Scene Audio-Visual Alignment**, and **Fine-Grained Audio-Visual Contrast**. Section 3.2 presents representative cases.

3.2 Representative Examples

Figure 2 illustrates how the four categories connect acoustic and visual evidence. The relevant information may be a change in vocal tone, a word spoken at a particular moment, or a visual difference between intervals identified through speech.

Event progression. In panel (a), the question asks how the voices change when the air fryer begins to spin. The annotated answer is that they express surprise. The visible action identifies the event of interest, while the accompanying voices provide the reaction. Answering requires relating the acoustic change to that event, rather than simply recognizing the appliance.

Cross-scene alignment. Panel (b) asks which pronoun sign is being attached to the wall when the teacher first warns against using informal terms. The answer, *vosotros*, requires matching the spoken warning to the corresponding visual moment. Recognizing the signs in individual frames does not by itself establish which one coincides with the warning.

Fine-grained contrast. In panel (c), the model must compare the keyboard backlight during the speaker's initial question and subsequent explanation. The annotated change is from blue to purple. Speech identifies the two intervals to compare, and the video supplies the color information. Retaining a single representative frame can lose the contrast even if the keyboard remains clearly visible.

Scene or temporal transition. Panel (d) asks which sound immediately precedes the camera panning toward a box. The annotated answer is a woman asking a question. This case tests whether the model preserves the order of an acoustic event and a visual transition, including evidence just before the camera movement.

These cases illustrate why audiovisual association matters for compression. A method may preserve salient frames but discard a brief vocal reaction, or retain an audio anchor without the visual context needed to interpret it. Comparing speech-defined intervals also requires retaining multiple visual states. UGC-AVQA tests whether these cues and their temporal relations remain available within a reduced token budget.

3.3 Difficulty-Guided Split Construction

To reduce reliance on visual information alone, we remove the audio track from each candidate sample and evaluate Gemini-3.1-Flash on the video-only input (Figure 1). A sample enters the benchmark only if the model fails under this condition. Trained annotators then review 30% of the retained candidates and correct them where needed.

The benchmark split contains 206 videos and 1,648 QA pairs. The remaining 794 videos form the training split. Gemini-assisted annotation produces 3,168 training QA pairs in four corresponding categories: **Event Progression**, **Scene Shift**, **Audio-Scene Match**, and **Key Multimodal Difference**. Trained annotators review and, where necessary, correct 30% of these generated pairs. This split assigns easier examples to training and reserves the harder, audio-dependent candidates for evaluation.

3.4 Evaluation Protocol

All questions use a multiple-choice format. We report overall accuracy and accuracy within each category. By filtering questions through an audio removal test, UGC-AVQA targets the preservation of acoustic evidence and its visual and temporal context under compression.

4 Omni Memory-Augmented Compression

Omni Memory-Augmented Compression (OMAC) compresses audiovisual inputs without additional training. Let $\mathcal{A} = \{a_i\}_{i=1}^{N_a}$ denote audio tokens, $\mathcal{V} = \{v_{t,p}\}$ denote video tokens, and $\mathbf{q} \in \mathbb{R}^d$ denote the query embedding. Here, $t \in \{1, \ldots, T\}$ indexes frames and $p \in \{1, \ldots, P\}$ indexes visual positions. OMAC produces audio memory $\mathcal{M}_a$ and visual memory $\mathcal{M}_v$, preserving their original temporal order. It selects frames relevant to the query, retains distinctive visual tokens and frame summaries, and merges nearby audio evidence into anchors. The retained visual memory then guides the allocation of the audio budget over time. Audio and video compression are controlled by ρ_a and ρ_v, respectively; other local selection and merging hyperparameters are fixed across experiments.

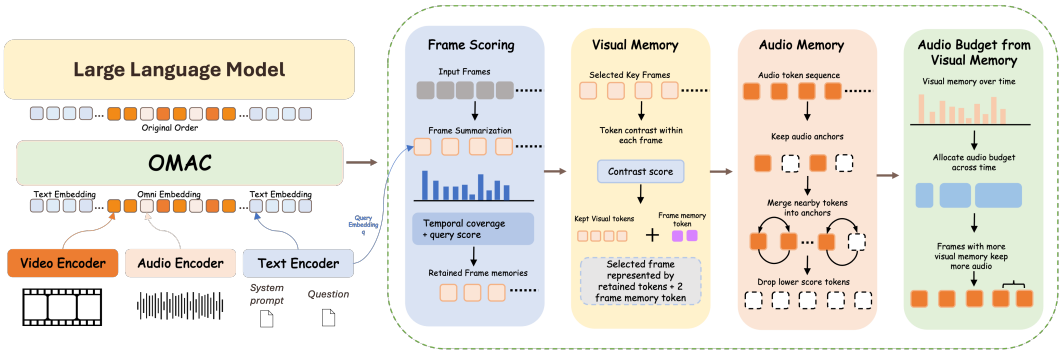


Figure 3: **OMAC compression.** OMAC retains informative visual and acoustic cues, constructs compact frame memory tokens, and assigns more audio capacity to intervals that retain more visual memory.

4.1 Frame Scoring and Visual Memory

We first average the token embeddings within each frame:

$$\bar{\mathbf{v}}_t = \frac{1}{P}\sum_{p=1}^{P}\mathbf{v}_{t,p}, \tag{1}$$

where $\mathbf{v}_{t,p}$ is the embedding at position p in frame t. We score each frame by the cosine similarity between this summary and the query:

$$s_t = \cos(\bar{\mathbf{v}}_t, \mathbf{q}). \tag{2}$$

OMAC selects high-scoring frames subject to temporal coverage, with the number of retained frames controlled by ρ_v. This balances query relevance with coverage across the video.

Within each selected frame, we compute a contrast score for every token:

$$\alpha_{t,p} = 1 - \cos(\mathbf{v}_{t,p}, \mathbf{c}_t), \quad \mathbf{c}_t = \frac{1}{P}\sum_{p=1}^{P}\mathbf{v}_{t,p}. \tag{3}$$

Here, $\mathbf{c}_t$ is the frame centroid. Tokens that differ from the average frame content receive higher scores, favoring distinctive local evidence over repeated background content. After normalizing $\alpha_{t,p}$, we retain the highest-scoring tokens as explicit visual memory.

We also construct a frame memory token by weighted pooling:

$$\mathbf{z}_t = \sum_{p\in\mathcal{K}_t}\frac{\exp(\hat{\alpha}_{t,p})}{\sum_{j\in\mathcal{K}_t}\exp(\hat{\alpha}_{t,j})}\mathbf{v}_{t,p}, \tag{4}$$

where $\mathcal{K}_t$ contains the retained tokens in frame t and $\hat{\alpha}_{t,p}$ is the normalized contrast score. We insert $\mathbf{z}_t$ into one available token position in that frame, preserving the original token order. Visual memory therefore has two levels: frame summaries determine which moments to retain, while selected local tokens and a pooled memory token represent each retained frame.

4.2 Audio Memory Budget Adjustment

For audio, OMAC retains high-scoring tokens as anchors and merges nearby discarded tokens into them. Let ℓ_i denote the importance score of audio token a_i, and let $\mathcal{G}(j)$ contain the tokens assigned to anchor a_j. The updated anchor feature is

$$\tilde{\mathbf{a}}_j = \frac{\mathbf{a}_j + \sum_{i\in\mathcal{G}(j)}\omega_{ij}\mathbf{a}_i}{1 + \sum_{i\in\mathcal{G}(j)}\omega_{ij}}, \tag{5}$$

where $\mathbf{a}_j$ is the original anchor feature and ω_{ij} is the merge weight for token a_i. The number of anchors is controlled primarily by ρ_a. Each updated anchor summarizes nearby acoustic evidence,

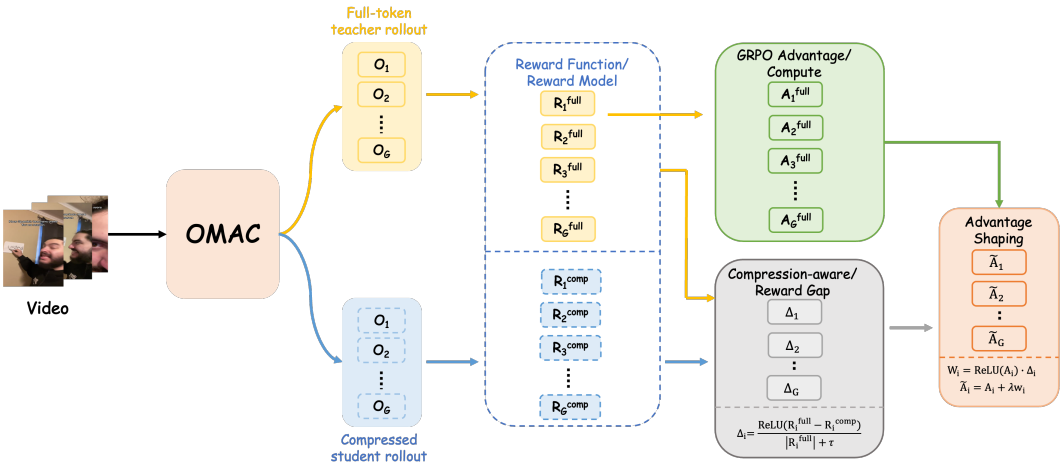


Figure 4: **O-MARC training.** Full-token and compressed rollouts are sampled from the current policy. Their reward gap shapes the GRPO advantage to train the policy on compressed inputs.

including speech cues, sound events, and local temporal context. OMAC also groups audio tokens into aligned temporal segments to allocate memory across the video.

The retained visual memory provides a temporal signal for this allocation. Let $\pi(i)$ map audio token a_i to its aligned frame, and let w_t be a visual importance weight derived from the amount of memory retained for frame t, including explicit tokens and frame memory tokens. For a total audio budget B_a and n_t audio tokens aligned to frame t, the target budget is

$$b_t = B_a \cdot \frac{n_t w_t}{\sum_{\tau=1}^{T} n_\tau w_\tau}. \tag{6}$$

This allocation preserves the total audio budget. It assigns more acoustic memory to intervals with more retained visual evidence and compresses other intervals more strongly, helping preserve related evidence across the two modalities.

5 Omni Memory-Augmented Compression Distillation

O-MARC combines OMAC with compression self-distillation through GRPO (Feng et al., 2026; Guo et al., 2025; Wu et al., 2026). As shown in Figure 4, each query q is evaluated through two branches of the current policy π_θ: a full-token teacher branch and a student branch with memory compression enabled. Both branches generate rollouts online from the current policy, so training remains on-policy.

Let R_i^{full} and R_i^{comp} be the scalar rewards of the i-th teacher and student rollouts. We measure the reward drop under compression as

$$\Delta_i = \frac{\mathrm{ReLU}(R_i^{\text{full}} - R_i^{\text{comp}})}{|R_i^{\text{full}}| + \tau}, \tag{7}$$

where $\mathrm{ReLU}(\cdot)$ retains only positive drops and τ is a small constant for numerical stability. A larger Δ_i indicates a greater loss of reward under compression for the same query.

We use this gap to reshape the GRPO advantage. Let A_i be the original advantage of rollout i. The distillation weight is

$$w_i = \mathrm{ReLU}(A_i) \cdot \Delta_i, \tag{8}$$

and the shaped advantage is

$$\tilde{A}_i = A_i + \lambda w_i, \tag{9}$$

where λ controls the strength of the adjustment. The added weight is positive only when the GRPO advantage is positive and the compressed branch receives a lower reward than the full-token branch.

Methods	Rank	Avg.	AVEP	AVST	CSAVA	FGAVC
			UGC-AVQA			
Proprietary Models (API)						
Gemini-3.1-pro	1	69.1	68.7	71.4	68.5	68.0
Gemini-3.1-flash-lite	2	67.1	68.5	68.0	64.3	67.5
Open-source Models						
Gemma-4-E4B-it	4	59.7	58.5	61.7	55.6	63.1
Gemma-4-E2B-it	9	45.6	46.1	45.2	42.0	49.0
Gemma-3n-E4B-it	5	54.6	58.0	57.0	47.6	55.6
Gemma-3n-E2B-it	7	53.0	55.8	55.6	46.4	54.4
Qwen2.5-Omni-7B	6	54.2	52.7	54.4	51.2	58.3
Qwen2.5-Omni-3B	8	49.0	44.7	52.4	48.3	50.5
Qwen3-Omni-30B-A3B	3	65.0	65.3	66.8	60.9	67.0

Table 2: **UGC-AVQA results.** Average and category accuracy for proprietary and open-source omnimodal models. AVEP, AVST, CSAVA, and FGAVC refer to event progression, scene or temporal transition, cross-scene alignment, and fine-grained contrast, respectively.

We denote the clipped policy ratio by

$$\rho_i(\theta) = \mathrm{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\mathrm{old}}}(o_i \mid q)},\ 1-\epsilon,\ 1+\epsilon\right), \tag{10}$$

where $\pi_{\theta_{\mathrm{old}}}$ is the rollout policy used to generate the current batch and ϵ is the clipping threshold.

Replacing A_i with $\tilde{A}_i$ gives the compression-aware objective

$$\mathcal{L}_{\text{C-GRPO}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\rho_i(\theta)\tilde{A}_i - \beta\,\mathrm{KL}\big(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big)\right]. \tag{11}$$

Here, G is the rollout group size, π_{ref} is the reference policy, and β weights the KL term. The modification to the underlying GRPO procedure is the shaped advantage. The full-token branch supplies a reward-based training signal, using advantage shaping to transfer information to the compressed branch without matching token-level probabilities.

6 Experiments

6.1 Experimental Settings

Inference settings. We sample video at 1 frame per second (FPS), retain at most 32 frames per video, and set the maximum frame size to 50,174 pixels. Audio is sampled at 16 kHz. Unless stated otherwise, compression retains 30% of the tokens. We evaluate the resulting inputs on downstream QA benchmarks and report the mean of two runs.

Training settings. Our primary model is Qwen2.5-Omni-3B, trained on four NVIDIA RTX PRO 6000 GPUs with 96 GB of memory each. Training uses the same video preprocessing as evaluation: 1 FPS, at most 32 frames, and a maximum frame size of 50,174 pixels. We train for 1 epoch with a batch size of 1 per device, a learning rate of 1×10^{-5}, and four sampled generations per prompt. The 3B model is our main setting because of its lower inference cost. We also evaluate Qwen2.5-Omni-7B and Qwen3-Omni-30B-A3B to examine behavior at larger scales. Additional details are given in Appendix A.1.

6.2 UGC-AVQA Benchmark Results

Table 2 reports UGC-AVQA results across proprietary and open-source models. Gemini-3.1-pro and Gemini-3.1-flash-lite achieve the highest average accuracies, at 69.1 and 67.1. Qwen3-Omni-30B-A3B leads the open-source models at 65.0, while Qwen2.5-Omni-3B reaches 49.0. The gap shows that the benchmark remains challenging for smaller omnimodal models.

Method	Daily-Omni	UGC-AVQA					OmniVideo Bench	WorldSense	Average
	Acc.	AVEP	AVST	CSAVA	FGAVC	Acc.	Score	Acc.	Score
Qwen3-Omni-30B-A3B									
Full tokens	69.9	65.3	66.8	60.9	67.0	65.0	39.4	52.5	56.7
OmniZip	61.1	51.5	56.6	54.4	53.4	53.9	31.7	47.4	48.5
OMAC	**63.2**	57.3	59.0	53.9	57.5	**56.9**	**35.5**	**48.0**	**50.9**
Qwen2.5-Omni-7B									
Full tokens	56.3	52.7	54.4	51.2	58.3	54.2	34.6	43.6	47.2
OmniZip	51.8	51.5	53.6	50.0	55.8	52.7	30.0	39.6	43.5
OMAC	53.6	52.4	54.9	49.3	56.3	53.2	30.9	42.4	45.0
OmniSIFT (100k+)	59.0	52.9	55.8	55.1	61.7	56.4	34.0	**44.4**	48.5
O-MARC (3k+)	54.6	56.3	64.1	51.2	65.3	59.2	31.5	43.4	47.2
O-MARC (100k+)	**60.4**	65.7	68.2	58.3	66.3	**64.6**	**35.2**	44.0	**51.1**
Qwen2.5-Omni-3B									
Full tokens	53.8	44.7	52.4	48.3	50.5	49.0	31.6	42.2	44.1
OmniZip	47.7	44.7	49.0	44.9	49.8	47.1	29.3	39.7	41.0
OMAC	49.9	47.6	52.4	47.1	49.5	49.2	30.8	41.2	42.8
O-MARC (3k+)	**52.3**	52.4	61.4	55.8	61.9	**57.9**	**31.0**	**42.1**	**45.8**

Table 3: **Main compression results.** All methods use 32 frames. Full-token inference retains 100% of tokens; all compression methods retain 30%. Parentheses give the training data scale. Daily-Omni, UGC-AVQA, and WorldSense report accuracy; OmniVideoBench reports average score. The final column averages these four benchmark scores. Bold marks the best compressed result for each backbone on these scores. UGC-AVQA category results are also shown.

Cross-Scene Audio-Visual Alignment (CSAVA) is the lowest-scoring category for several open-source models, including Qwen3-Omni-30B-A3B at 60.9 and Gemma-3n-E4B-it at 47.6. These results motivate preserving acoustic cues together with their visual and temporal context during compression.

6.3 Main Results of OMAC and O-MARC

At 30% token retention, OMAC improves the average over OmniZip (Tao et al., 2025) by 2.4, 1.5, and 1.8 points on the 30B, 7B, and 3B backbones, respectively (Table 3). On the 30B model, the gains reach 3.0 points on UGC-AVQA and 3.8 on OmniVideoBench. Memory compression preserves useful evidence at a fixed budget, though its average remains below full-token inference.

With 3k training examples, O-MARC raises the 3B average to 45.8, exceeding full-token inference by 1.7 points. UGC-AVQA accuracy reaches 57.9, up from 49.2 with OMAC, while the other benchmarks remain close to their full-token baselines. On the 7B model, 3k examples yield 47.2, matching the full-token average; the 100k+ setting reaches 51.1.

OmniSIFT (Ding et al., 2026) uses 107k training examples; our comparable setting combines 104k supervised fine-tuning examples with 3k UGC examples for reinforcement learning. On the same 7B backbone, O-MARC averages 51.1 versus 48.5, a 2.6-point gain. It improves UGC-AVQA from 56.4 to 64.6 and OmniVideoBench from 34.0 to 35.2. It also leads on Daily-Omni (60.4 versus 59.0), while trailing on WorldSense (44.0 versus 44.4). The strongest gain is on the audiovisual association tasks targeted by UGC-AVQA training.

6.4 Ablation Studies

Compression ratio. At 55% token removal, OMAC reaches 49.5 on UGC-AVQA, above full-token inference at 49.0 and OmniZip at 48.6 (Table 4). At 70% removal, OMAC retains 49.2 while OmniZip falls to 47.1, widening the gap under stronger compression.

Video size. OMAC improves on OmniZip by 1.5 points with either 64 frames (41.7 versus 40.2) or 32 frames (41.2 versus 39.7). Table 5 shows that the gain persists when both frame count and maximum frame size decrease.

Method	Compression Ratio	Avg.
Full tokens	0%	49.0
OmniZip	55%	48.6
OmniZip	70%	47.1
OMAC	55%	49.5
OMAC	70%	49.2

Table 4: **Compression ratio on UGC-AVQA.** The ratio denotes the percentage of tokens removed.

Method	Frames	Size	Avg.
Full tokens	64	$128 \times 28 \times 28$	43.6
Full tokens	32	$64 \times 28 \times 28$	42.2
OmniZip	64	$128 \times 28 \times 28$	40.2
OmniZip	32	$64 \times 28 \times 28$	39.7
OMAC	64	$128 \times 28 \times 28$	41.7
OMAC	32	$64 \times 28 \times 28$	41.2

Table 5: **Video size on WorldSense.** Accuracy at different frame counts and maximum frame sizes.

Method	Time (s)	Memory (GB)
Full tokens	3.81	24.2
OmniZip	2.47	15.7
OMAC	2.49	15.8

Table 6: **Inference cost on WorldSense.** Latency and GPU memory use with 64 frames and a maximum frame size of $128 \times 28 \times 28$.

Method	Guidance	Daily-Omni
Full tokens	NA	53.8
OmniZip	A-V	47.7
OMAC (V-A only)	V-A	48.2
OMAC	V-A-VA	49.9

Table 7: **Cross-modal guidance.** All methods use 32 frames. A-V and V-A denote audio-guided visual and visual-guided audio compression. V-A-VA uses retained visual memory to adjust the audio budget.

Efficiency. OMAC reduces latency from 3.81 s to 2.49 s and GPU memory use from 24.2 GB to 15.8 GB, saving 34.6% and 34.7% (Table 6). Its higher accuracy comes at nearly the same cost as OmniZip, which uses 2.47 s and 15.7 GB.

Guidance stage. Audio-guided visual compression reaches 47.7 on Daily-Omni, and direct visual guidance of audio compression reaches 48.2 (Table 7). OMAC reaches 49.9 by first constructing memory, then adjusting the audio budget, supporting guidance based on retained visual evidence.

7 Conclusion

We presented UGC-AVQA, OMAC, and O-MARC to study efficient audiovisual reasoning. UGC-AVQA focuses on questions that connect acoustic cues with visible events, using audio removal to retain questions the filtering model cannot answer from video alone. OMAC compresses these inputs into visual memory and audio anchors, then allocates acoustic capacity according to the retained visual evidence. O-MARC trains models to use this representation through reward-based self-distillation between full-token and compressed rollouts from the same policy.

OMAC improves on OmniZip across three model scales. In the WorldSense efficiency test, it reduces latency and GPU memory use by about 35% relative to full-token inference. With 30% of the tokens retained, O-MARC raises the 3B average to 45.8, above the full-token baseline of 44.1. At a comparable training data scale, the 7B model also exceeds OmniSIFT by 2.6 points on average, with its largest gain on UGC-AVQA. These results suggest that efficient omnimodal reasoning depends both on preserving related acoustic and visual evidence and on teaching the model to reason from the compressed representation.

AI Use Statement

For manuscript preparation, generative AI tools were used only to polish the language and improve readability. The authors retain full responsibility for the scientific content, conclusions, and final wording of the manuscript.

REFERENCES

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

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

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

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

Yue Ding, Yiyan Ji, Jungang Li, Xuyang Liu, Xinlong Chen, Junfei Wu, Bozhou Li, Bohan Zeng, Yang Shi, Yushuo Guan, et al. Omnisift: Modality-asymmetric token compression for efficient omni-modal large language models. *arXiv preprint arXiv:2602.04804*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

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, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24108–24118, 2025.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

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

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, Jiayi Ji, et al. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. *Advances in Neural Information Processing Systems*, 38:142289–142324, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

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.

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *arXiv preprint arXiv:2505.21334*, 2025a.

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

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. Fastvid: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2026.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

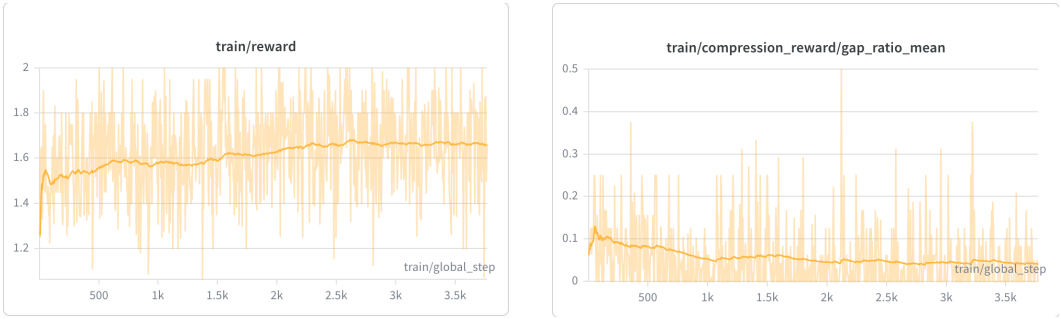
Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Peiran Wu, Zhuorui Yu, Yunze Liu, Chi-Hao Wu, Enmin Zhou, and Junxiao Shen. Marc: Memory-augmented rl token compression for efficient video understanding. *arXiv preprint arXiv:2510.07915*, 2025.

Peiran Wu, Yunze Liu, Miao Liu, and Junxiao Shen. St-think: How multimodal large language models reason about 4d worlds from ego-centric videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5174–5183, 2026.

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*, 2025.

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



(a) Training reward.

(b) Compression reward gap ratio.

Figure 5: **O-MARC training dynamics.** Total reward increases over training, while the reward gap between compressed and full-token rollouts decreases.

A Additional Details

A.1 Training Details

GRPO training uses four sampled rollouts per prompt, a maximum completion length of 1,024 tokens, and a maximum context length of 32,768 tokens. We train for 1 epoch with a learning rate of 1×10^{-5} and a batch size of 1 per device, using bfloat16 and DeepSpeed ZeRO-2. The reward sums answer accuracy and format correctness with equal weights. Video is sampled at 1 FPS with at most 32 frames. Audio output generation is disabled, and audio-in-video reasoning is enabled.

Figure 5 shows an upward trend in reward and a downward trend in the compression gap ratio, despite local fluctuations. These trends are consistent with improved task performance and a smaller reward difference between the compressed and full-token branches.

A.2 Annotator Instructions

Annotators were student research assistants from the authors' institution with experience in multimodal annotation. They were compensated according to local institutional standards and informed that the videos and annotations would be used for audiovisual reasoning research.

Annotators inspected both the visual stream and audio track. They wrote detailed multimodal captions covering visible entities, actions, scene transitions, speech, non-speech sounds, and temporal relations. They were instructed not to add private identity information beyond what was already visible or audible in the public video.

For each video, annotators wrote multiple-choice questions requiring both modalities. The four categories were defined as follows:

- **Audio-Visual Event Progression:** what acoustic event precedes or follows a visual event.
- **Audio-Visual Scene or Temporal Transition:** how audio changes across visual scene transitions.
- **Cross-Scene Audio-Visual Alignment:** which audio segment corresponds to a visual moment across scenes.
- **Fine-Grained Audio-Visual Contrast:** how subtle acoustic cues differ across visual states.

Each question required one correct option, plausible distractors, and supporting evidence from the caption. During quality control, reviewers removed questions with ambiguous answers, questions solvable from vision alone, malformed options, and answers unsupported by the video.

Benchmark	Representative categories	Category focus
UGC-AVQA	Event progression, scene or temporal transition, cross-scene alignment, fine-grained contrast	Comparing, aligning, or contrasting audio and visual evidence over time.
WorldSense	Object counting, spatial relations, attribute recognition, event sorting, temporal localization, audio recognition	Broad understanding across visual, acoustic, temporal, and semantic tasks.
OmniVideoBench	Counting, causal reasoning, temporal understanding, reference reasoning, fine-grained perception, sentiment analysis	General long-video reasoning, often using speech to locate or disambiguate visual evidence.
Daily-Omni	Event sequence, AV event alignment, context understanding, reasoning, inference, comparison	Everyday video understanding, including event order, scene context, and audiovisual matching.

Table 8: **Question taxonomies.** UGC-AVQA centers on audiovisual association; the other benchmarks cover a broader range of video understanding tasks.

A.3 Question Taxonomy Comparison

We compare UGC-AVQA's question categories with those in the annotation files of existing audiovisual benchmarks. Its four categories focus on relating acoustic and visual evidence across events and time, beyond locating a visual interval or recognizing an isolated sound.

As Table 8 shows, the existing benchmarks cover audiovisual tasks within broader taxonomies. WorldSense includes audio recognition and counting alongside visual attributes, spatial relations, and object counts. OmniVideoBench organizes questions around general abilities such as counting, causal reasoning, and temporal understanding. Daily-Omni includes audiovisual event alignment within a broader set of everyday video tasks. UGC-AVQA makes dependence on both modalities the organizing principle of its question categories.

The question types illustrate these differences. A WorldSense question may ask where an object is relative to a person, primarily testing spatial perception. An OmniVideoBench question may ask how many objects appear during a passage of narration, using audio to locate a moment and vision to answer. A Daily-Omni question may ask which sound is synchronized with a shot. Section 3.2 discusses four UGC-AVQA cases in detail, showing how the questions connect vocal reactions, spoken cues, visual changes, and event order. The audio removal check further selects questions that the filtering model cannot answer from muted video.