Attention-Guided Audio Compression For Multimodal LLMs

Prerana Rane[1], Amitesh Vatsa[2], Yash Pethe[3], Ogan Batu Aktolun[4], Kevin Li[3], Ishan Singh[3]

[1]IEEE Senior Member [2]Indian Institute of Technology (BHU) [3]Independent Researcher [4]University of Texas at Austin

Problem Definition

Challenge:

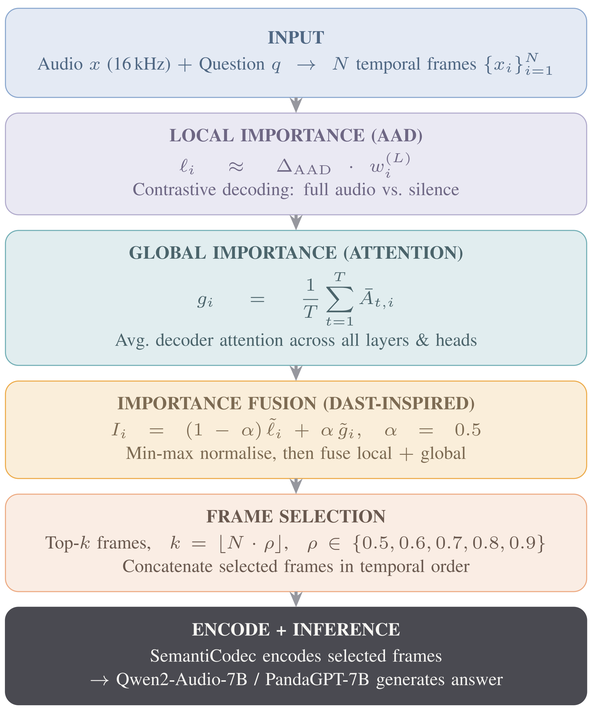
- Audio inputs create high token counts in multimodal LLMs — a 10 s clip requires **1,000 tokens** at 100 tok/s.
- Compression can cause **hallucinations** when semantic information is lost during tokenization.
- No metric exists to evaluate semantic preservation of audio in LLMs.

Our Approach: Attention-guided frame selection using fused local (AAD) and global (decoder attention) importance scores, evaluated with a new Answer Consistency metric.

Method: Attention-Guided Pipeline

We propose an attention-guided audio compression pipeline that selectively retains semantically important frames before encoding, reducing token load while preserving task-relevant content for multimodal LLMs.

- **Audio-Aware Decoding (AAD)** quantifies per-frame local semantic relevance to the input question via contrastive decoding (full audio vs. silence).
- **Global decoder attention** aggregates attention weights across all layers, heads, and generated tokens for a holistic importance view.
- **DAST-inspired fusion** linearly combines the signals (α=0.5) after min-max normalisation and selects top-k frames for SemantiCodec encoding.



Answer Consistency Metric

No existing metric evaluates semantic loss for audio in LLMs. We introduce **Answer Consistency (A.C.)**:

$$\text{A.C.} = \frac{1}{M}\sum_{s=1}^{M} \mathbf{1}\left[a^s_{\text{orig}} = a^s_{\text{comp}}\right]$$

- Binary labels (yes/no); ranges $0 \to 1$.
- Independent of correctness — measures proportion of samples where compressed audio produces the same answer as original.
- A.C.=1 means compression never changes the model's answer.

Results: ClothoAQA (Qwen2-Audio-7B)

$N = 2490$, mean $\pm$ std across 3 independent seeds. r = frame rate (fps), ρ = keep ratio, A.C. = Answer Consistency. All metrics in %.

Method	r	ρ	Acc	F1	A.C.
Baseline	—	1.0	74.1 $\pm$1.04	75.1 $\pm$0.83	100.0 $\pm$0.00
SemantiCodec	25	1.0	73.8 $\pm$0.96	74.7 $\pm$0.91	98.5 $\pm$0.10
Attention	**25**	**0.9**	**73.1** $\pm$1.30	73.6 $\pm$0.38	**98.8** $\pm$0.15
SemantiCodec	50	1.0	74.2 $\pm$0.95	74.6 $\pm$0.93	98.3 $\pm$0.30
Attention	50	0.9	73.0 $\pm$1.21	73.5 $\pm$1.01	98.6 $\pm$0.31
SemantiCodec	100	1.0	74.2 $\pm$0.95	74.6 $\pm$0.93	98.3 $\pm$0.46
Attention	100	0.9	72.8 $\pm$1.18	73.3 $\pm$1.04	98.4 $\pm$0.46

Key finding: At ρ=0.9, compute is reduced by 10% while maintaining **98.8%** answer consistency — semantic content is preserved.

Answer Consistency Paradox

Compression Threshold Ablation ($N = 1000$, mean $\pm$ std across 3 independent runs with random seeds)

Method	r	ρ	Acc	ΔAcc	A.C.
Baseline	—	1.0	76.9 $\pm$1.27	—	100.0 $\pm$0.00
Attention	25	0.7	74.1 $\pm$1.01	-2.8	98.5 $\pm$0.44
Attention	50	0.7	73.4 $\pm$0.90	-3.5	98.9 $\pm$0.40

Answer Consistency Paradox

At r=50, ρ=0.7: *lowest* accuracy (73.4%) yet *highest* A.C. (98.9%). Both original and compressed audio elicit the same **incorrect** answer — a systematic compression bias, not random error.

Results: AVHBench Audio→Video Hallucination

PandaGPT 7B, N=5816 samples, "Yes" is positive class. All metrics in %

Method	r	ρ	Acc	Rec	F1	ΔAcc
Baseline	—	—	49.91	92.43	64.85	—
SemantiCodec	100	1.0	50.88	91.20	64.99	+0.97
Attention	100	0.8	50.70	90.85	64.82	+0.79
Attention	**100**	**0.5**	**52.11**	88.91	65.00	**+2.20**
SemantiCodec	50	1.0	51.32	90.67	65.06	+1.41
Attention	50	0.8	50.44	90.14	64.52	+0.53
Attention	50	0.5	51.67	88.38	64.65	+1.76
SemantiCodec	25	1.0	51.58	91.37	65.37	+1.67
Attention	25	0.8	49.91	92.61	64.90	0.00
Attention	25	0.5	51.50	88.91	64.70	+1.59

Key finding: Aggressive compression (ρ=0.5) *helps* hallucination detection. Best: **+2.20%** at r=100, ρ=0.5.

Conclusion & Future Work

Findings:

- Attention-guided compression demonstrated improvements in semantic preservation across all configurations.
- ρ=0.9 maintains near-baseline accuracy with a **10% compute reduction**.
- Critical threshold: ρ<0.7 causes $\geq$2.8% accuracy degradation across all frame rates.
- The **Answer Consistency Paradox** reveals systematic compression-induced bias rather than random noise.
- Semantic codecs generalise across models (Qwen2-Audio-7B, PandaGPT-7B) and datasets (ClothoAQA, AVHBench).
- AVHBench's audio-driven video hallucination task using PandaGPT 7B embeddings yields +2.20% hallucination reduction.

Future Work:

- Extend evaluation to GPT-4V and Gemini.
- Investigate mechanisms underlying the consistency paradox.
- Explore adaptive ρ selection per task and audio type.
- Apply to additional AVHBench tasks beyond audio→video hallucination.