EchoingPixels: Aliasing-Resistant Joint Token Reduction for Audio-Visual LLMs

Chao Gong[1,2] Depeng Wang[2] Zhipeng Wei[3]
Ya Guo[2] Huijia Zhu[2] Jingjing Chen[1]

[1]Fudan University [2]Ant Group [3]UC Berkeley

ICML 2026

Code & Models: `github.com/CharlesGong12/EchoingPixels`

Roadmap

- **Background** —token compression in (audio-visual) MLLMs
- **Two Challenges** —cross-modal saliency & positional aliasing
- **Method** —*EchoingPixels*: CS2 + Sync-RoPE
- **Experiments** —quality, efficiency, ablations, visualization

Part 1

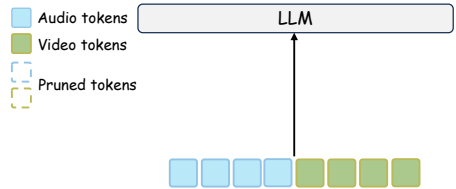
Background

Why token compression matters for AV-LLMs

Audio-Visual LLMs: Tokens Explode

- AV-LLMs (Qwen2.5-Omni, GPT-4o, Gemini) concatenate **video + audio + text** tokens.
- Qwen2.5-Omni at 1 FPS, 28×28 patch (**LLM input**): up to **144 visual** + **25 audio** tokens / sec
- A **30 s clip** ⇒ **>5,000 tokens** attention is $O(L^2)$ ⇒ prohibitive cost.

Bottleneck: **huge token sequences**, mostly redundant.



Vanilla AV-LLM: dense audio/visual tokens.

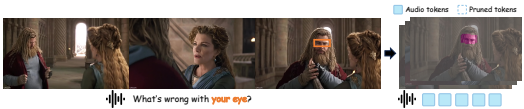
Part 2

Two Challenges

Cross-Modal Saliency + Positional Aliasing

Challenge #1 —Cross-Modal Saliency

The *relevance of an event in one modality* is often **defined by** the other.



Audio-guided saliency: spoken "*your eye*" ⇒ visual budget on the eye.

Heterogeneous saliency: ambient silence ⇒ shift budget to the visual stream.

Per-modality compression is **blind** to these synergies.

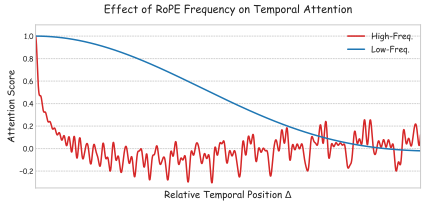
Challenge #2 —Positional Aliasing

Setup. Modern AV-LLMs use TMRoPE; *high-frequency* channels encode time.

- Adjacent retained tokens have effective stride T_s.
- Phase shift per channel: $\phi = T_s\theta$.
- **Nyquist:** need $T_s\theta \leq \pi$.

Under sparsification, T_s blows up $\Rightarrow$ high-freq channels **violate Nyquist** $\Rightarrow$ phase wraps mod 2π $\Rightarrow$ distant tokens look adjacent.

Result: temporal monotonicity is **silently corrupted**. AV sync and event ordering break.



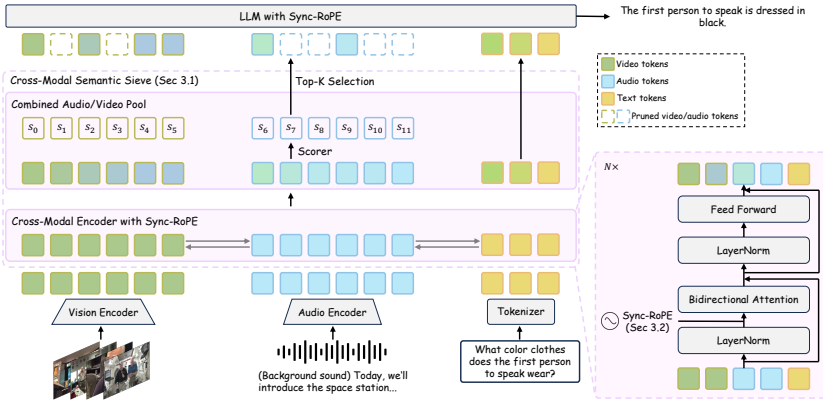
Red: vanilla high-freq RoPE oscillates over distance. Blue (ours): smooth and monotonic.

Part 3

EchoingPixels

Two co-designed modules: CS2 + Sync-RoPE

Framework Overview



- **CS2** (Cross-Modal Semantic Sieve) —extractive selection on the joint A/V/T stream.
- **Sync-RoPE** —spectral low-pass filter, applied in **both** CS2 encoder and main LLM.

CS2 —Design Rationale

Three requirements for "selecting the right tokens":

1. **Cross-modal synergy.**
   Sieve must operate on the *joint* A/V stream, not per-modality.
2. **Instruction pre-fusion.**
   Text tokens should *absorb* multimodal info *before* pruning, or it's lost.
3. **Global temporal context (bidirectional).**
   Importance can be retrospective; standard causal attention can't see the future.

⇒ A **trainable copy of the first *N* LLM layers**, with attention switched to **bidirectional**.

CS2 —Selection & Training

Contextualize. A bidirectional encoder $\mathcal{E}$ over the joint stream:

$$\hat{T} = \mathcal{E}\big(\mathrm{Concat}(T_v, T_a, T_t)\big)$$

Score. A 2-layer MLP scorer (text always kept):

$$s_i = \mathrm{MLP}(\hat{T}_i),\ \ i \in \{1, \ldots, L_v{+}L_a\}$$

Select. Global Top-K on the **unified A/V pool**:

$$I_{\mathrm{sel}} = \mathrm{TopK}\big(\{s_i\},\ \ k{=}r(L_v{+}L_a)\big)$$

- **Dynamic budget** —no fixed per-modality ratio.
- **Positions preserved** —required by Sync-RoPE.
- Top-K is non-differentiable ⇒ **Straight-Through Estimator (STE)**.

STE in one line

Forward: hard 0/1 mask y_i from Top-K.

Backward: set $\frac{\partial y_i}{\partial s_i} := 1$;

chain rule then gives

$$\frac{\partial y_i}{\partial \mathbf{w}_{\mathrm{MLP}}} = \frac{\partial s_i}{\partial \mathbf{w}_{\mathrm{MLP}}}.$$

Sync-RoPE —A Spectral Low-Pass Filter

The problem (recap). On the sparse grid:

$$\Delta_1\theta \equiv \Delta_2\theta \pmod{2\pi} \;\Rightarrow\; R(\Delta_1) \approx R(\Delta_2)$$

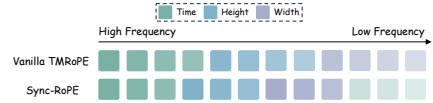
Nyquist-safe condition:

$$T_s\theta \le \pi \iff \theta \le \pi / T_s$$

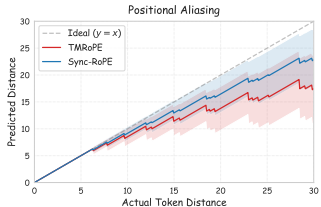
Fix. Repartition the d channels as

$$[\, t_{\text{high}},\ h,\ w,\ t_{\text{low}} \,]$$

- Keep t_{high} for compatibility & dense local resolution.
- Reassign temporal modeling to an **ultra-low band** Θ_{low} with $\max(\Theta_{\text{low}}) \ll \pi/\mathbb{E}[T_s]$.
- Guarantees a **bijective, monotonic** phase mapping globally.



TMRoPE: only high-freq for time.
Sync-RoPE: add ultra-low-freq channels to time.



Linear recovery ⇒ no aliasing.

Part 4

Experiments

Quality | Efficiency | Ablations | Visualization

Setup

- **Backbones:** Qwen2.5-Omni-3B / -7B (strong open-source AV-LLMs).
- **Token budgets:** 5%, 10%, 20% (3B); 10% (7B).
- **Benchmarks:**
  - AV: **WorldSense, Daily-Omni, Video-MME w/ audio**
  - Video-only: Video-MME, MLVU
- **Baselines:** Full Model, IntraModal (our combined unimodal baseline), FastV, PyramidDrop, OmniZip.
- **Training data:** only **390k samples** —initialization from pretrained LLM layers is data-efficient.

Main Results

Model	Bud.	Audio-Visual			Avg	Video		Avg	Rel.
		WSen.	D-Om.	$VMME_a$		VMME	MLVU		Perf.
Full Model (3B)	100%	45.4	59.6	63.1	56.1	60.9	67.2	64.1	100%
IntraModal	25%	37.4	46.2	52.1	45.2	51.4	63.4	57.4	84.2%
FastV	20%	40.1	49.9	51.9	47.3	51.8	56.3	54.1	84.6%
PyramidDrop	20%	39.1	53.0	56.2	49.4	53.9	59.7	56.8	88.3%
OmniZip	20%	39.5	50.6	57.9	49.3	N/A	N/A	N/A	87.9%
EchoingPixels (3B)	20%	45.0	60.7	60.7	55.5	58.6	68.3	63.5	**99.0%**
EchoingPixels (3B)	10%	43.5	57.6	58.4	53.2	55.4	67.4	61.4	95.2%
EchoingPixels (3B)	5%	40.9	52.9	55.7	49.8	54.7	66.1	60.4	91.0%
Full Model (7B)	100%	46.1	60.5	66.3	57.6	63.6	68.4	66.0	100%
EchoingPixels (7B)	10%	47.4	57.0	64.1	56.2	53.8	62.9	58.4	**94.1%**

EchoingPixels at **20% budget retains 99.0%** of full-model quality; at **5%, still 91.0%**.

Efficiency Analysis (Qwen2.5-Omni-3B, Daily-Omni)

Bud.	Fwd (ms)	Spd.	Mem (GB)	Red.
100%	517.1	1.00×	32.0	1.00×
20%	231.7	2.23×	14.2	2.26×
10%	194.0	2.67×	13.1	2.45×
5%	174.8	**2.96**×	12.3	**2.61**×

Take-away:
Up to **2.96**× forward speedup;
Up to **2.61**× peak memory cut.

Numbers *include* the CS2 module overhead.

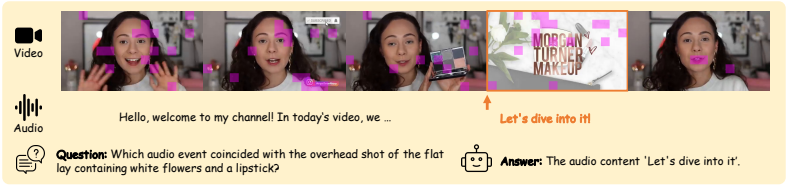
Key Ablations (Daily-Omni, 3B @ 20%)

Variant	Overall
Causal Attention	56.06
Intra-Modal Attention	57.73
w/o Pre-fusion	57.39
Per-modality Top-K	58.23
Similarity (heuristic)	41.85
Gumbel-Softmax	59.06
w/o Sync-RoPE	57.98
Full *EchoingPixels*	**60.65**

Every component matters:

- **Bidirectional** $>$ causal ($+4.6$).
- **Joint stream** $>$ intra-modal ($+2.9$).
- **Text pre-fusion** helps ($+3.3$).
- **Dynamic budget** $>$ per-modality ($+2.4$).
- **Learned** $\gg$ similarity heuristic ($+18.8$!).
- **STE** $>$ Gumbel-Softmax ($+1.6$).
- **Sync-RoPE** adds $+2.7$.

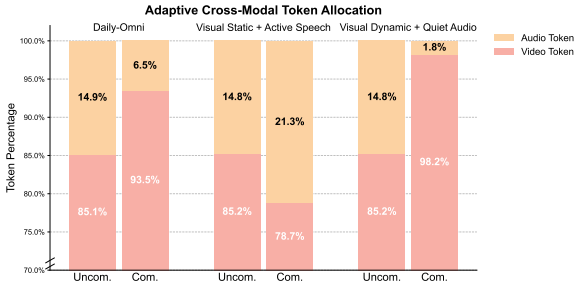
Qualitative —Fine-Grained Perception Survives Compression



Temporal alignment: audio "*Let's dive into it!*" linked to the matching shot.

OCR: sparse tokens preserved on tiny "*Cancer Research UK*" text.

Qualitative —Adaptive Budget Allocation



- Static visual scene ⇒ **budget shifts to audio**.
- Quiet audio ⇒ **budget shifts to video**.
- Driven by joint-modality saliency —not a fixed ratio.

This is what *joint Top-K* on the unified pool gives us.

Conclusion

- We identified **Positional Aliasing** —a previously overlooked failure mode of *any* sparse token reduction.
- *EchoingPixels* couples two co-designed modules:
  - **CS2** —joint-stream extractive selection with bidirectional context.
  - **Sync-RoPE** —Nyquist-aware spectral low-pass on the temporal channels.
- Results: **99% quality at 20% tokens**, **up to 2.96× speedup**, **2.61× memory cut**.

Code & Models: `github.com/CharlesGong12/EchoingPixels`

Thank you!

github.com/CharlesGong12/EchoingPixels