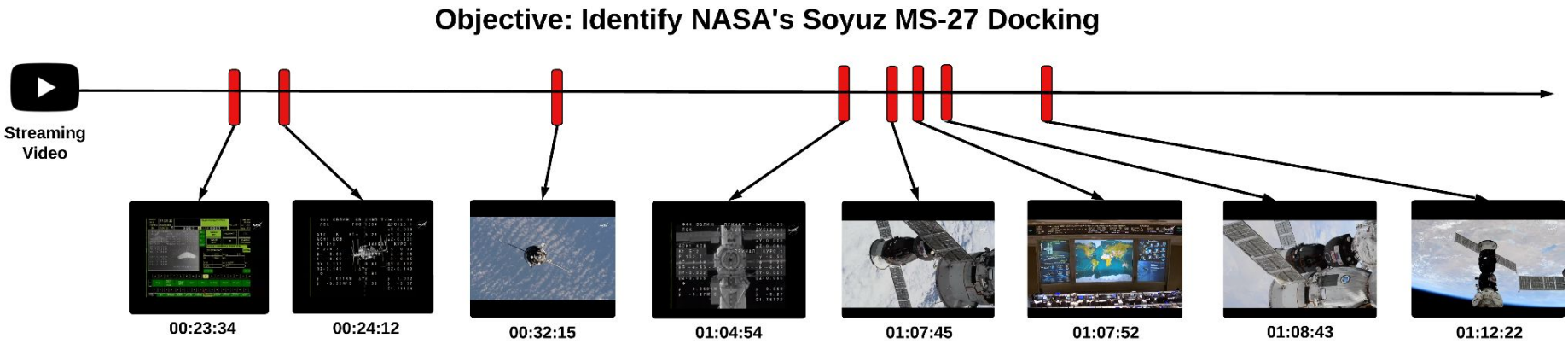
AHA - Predicting What Matters Next: Online Highlight Detection Without Looking Ahead

Aiden Chang[1], Celso De Melo[2], Stephanie Lukin[2]

KEY TAKEAWAY



RESEARCH QUESTION

How do we teach machines what's important in videos, like a human would know intuitively? Or, how do we perform highlight detection?

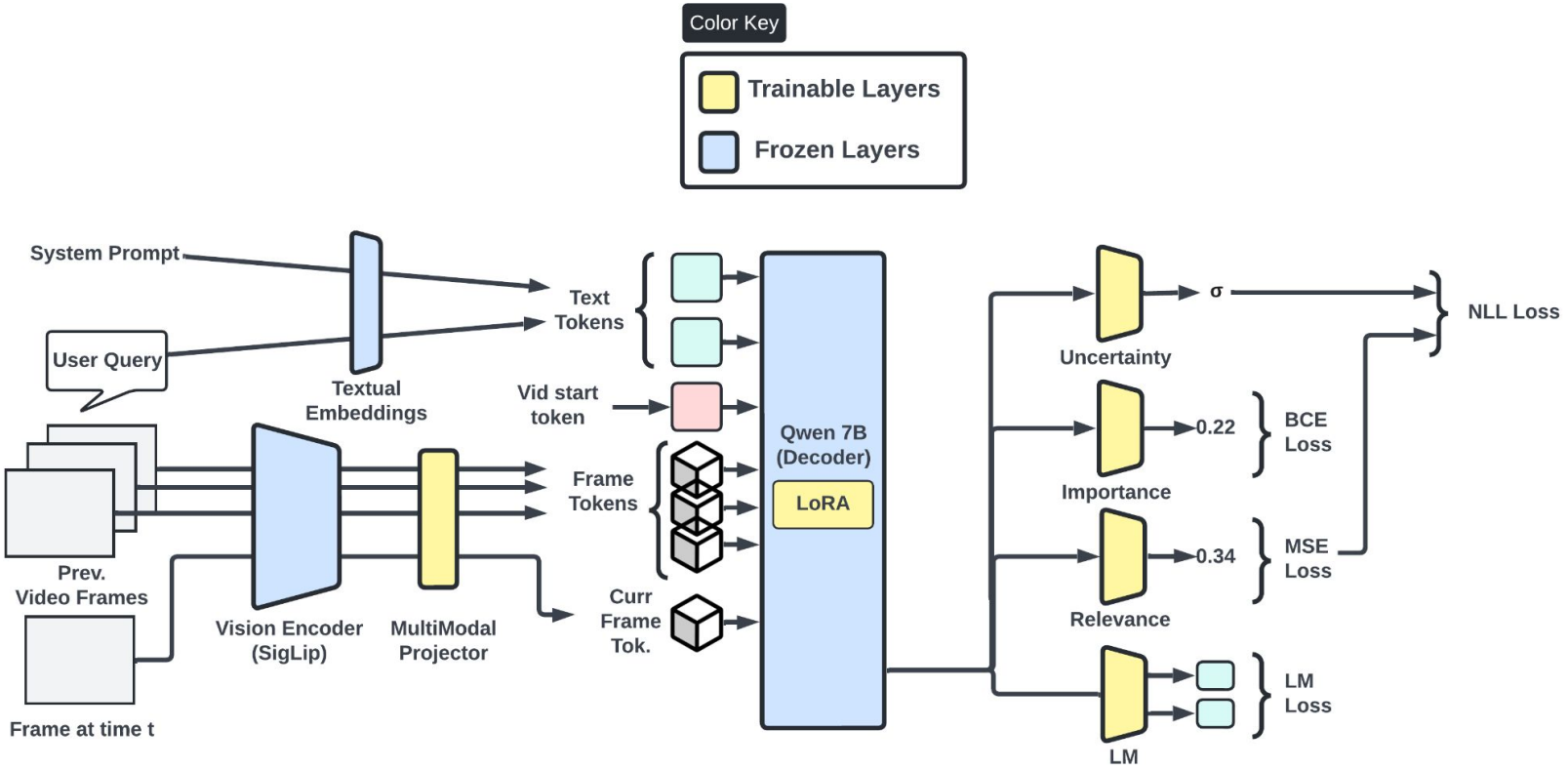
- What does it mean for something to be important?
- How do we define and measure this?

HIGHLIGHT DETECTION!

Why Current Highlight Detection Fails in Streaming Video

- Offline models cheat by looking into the future
- Video-LLMs use smoothing that violates online constraints
- KV cache grows without bound → OOM
- No dataset captures scalable human intuition

MODEL ARCHITECTURE



LOSS FUNCTIONS

We have:

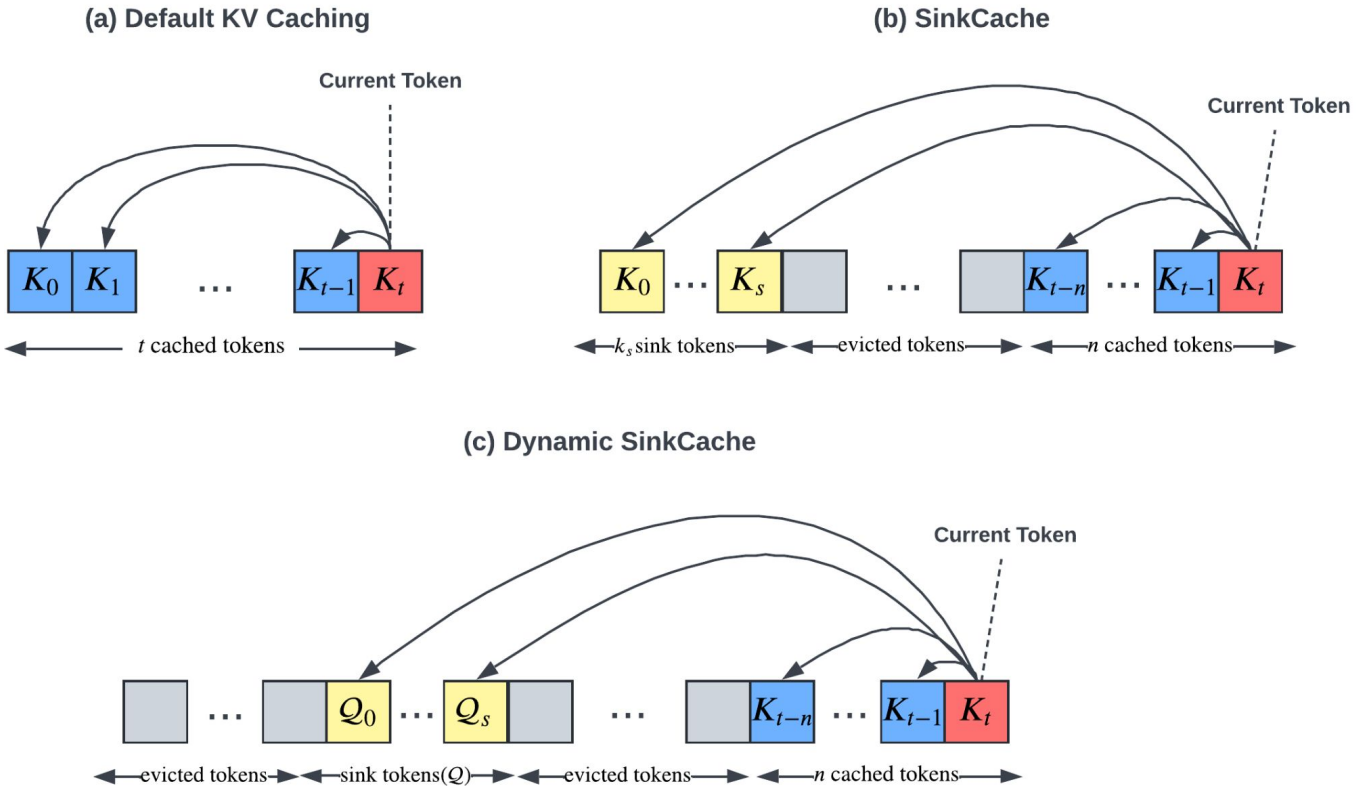
- TV loss(total variation loss)
- Info (information) loss
- Rel (relevance) loss
- LM loss
- Uncertainty Loss

$$L_{rel} \leftarrow L_{rel} + \lambda_{TV} L_{TV}$$

$$L_{video} = \lambda_{info} L_{info} + \lambda_{rel} L_{rel} + \lambda_{uncertainty} L_{uncertainty}$$

$$L_{total} = \lambda_{LM} L_{LM} + \lambda_{video} L_{video}$$

DYNAMIC SINK CACHE



BENCHMARK RESULTS

Table 1: TVSum Performance. We report top-5 mAP, τ, and ρ. 'Tuned?' indicates if fine-tuned on TVSum (Y) or not (N). Modalities: V (visual), T (text), A (audio). **Bold** is SOTA. (Per-category details: Appendix B.2).

Model	Tuned?	Modality	mAP	Kendall τ	Spearman ρ
Human [20]	N	V	–	0.177	0.204
PGL-SUM [18]	N	V	57.1	0.206	0.157
LLMVS [10]	N	V+T	–	0.211	0.275
UniVTG [17]	N	V	84.6	–	–
QD-DETR [37]	Y	V+A	86.6	–	–
TR-DETR [19]	Y	V+A	87.1	–	–
AHA (Zero-Shot)	N	V+T	91.6	**0.304**	**0.433**
AHA (Domain-Adapted)	N	V+T	**93.0**	0.285	0.406

Table 2: Overall HiSum performance on the full test set. **Bold** highlights our SOTA results.

Metric	SL-module [38]	iPTNet [39]	DSNet [40]	PGL-SUM [18]	**AHA (Ours)**
mAP@50	55.31	50.53	50.78	55.89	**64.19**
mAP@15	24.95	22.74	24.35	27.45	**32.66**

ARL SCOUT ANALYSIS