What Is That Talk About?
A Video-to-Text Summarization Dataset for Scientific Presentations

Dongqi Liu[Ω], Chenxi Whitehouse[Δ], Xi Yu[Ω], Louis Mahon[Θ], Rohit Saxena[Θ], Zheng Zhao[Θ], Yifu Qiu[Θ], Mirella Lapata[Θ], Vera Demberg[ΩΨ]

[Ω]Saarland University, [Ψ]Max Planck Institute for Informatics
[Δ]University of Cambridge, [Θ]University of Edinburgh

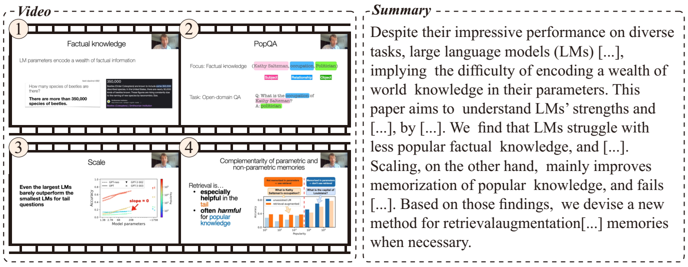
MAX PLANCK INSTITUTE FOR INFORMATICS

European Research Council
Established by the European Commission

ACL 2025 VIENNA

Introduction

We propose **VISTA**, the first multimodal summarization dataset consisting of scientific presentation videos paired with paper abstracts.

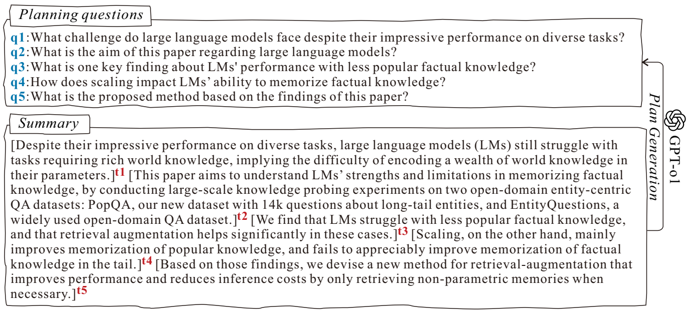


Despite their impressive performance on diverse tasks, large language models (LMs) [...], implying the difficulty of encoding a wealth of world knowledge in their parameters. This paper aims to understand LMs' strengths and [...], by [...]. We find that LMs struggle with less popular factual knowledge, and [...]. Scaling, on the other hand, mainly improves memorization of popular knowledge, and fails [...]. Based on those findings, we devise a new method for retrievalaugmentation[...] memories when necessary.

Figure 1. VISTA pairs presentation videos with paper abstracts

Plan-based Framework

- **Problem:** SOTA LMMs show problems with structural grounding -> incoherence, hallucination
- **Solution:** Introduce intermediate plan p as question sequence $\{q_1, q_2, \ldots, q_m\}$
- **Training & Inference:** Plan Generation (PG) and Summary Generation (SG) are trained separately on (v, p) and $([v; p], s)$, where v is the input and s is the summary. At inference, PG predicts $\hat{p}$ for v, and SG generates the summary from $[v; \hat{p}]$



q1:What challenge do large language models face despite their impressive performance on diverse tasks?
q2:What is the aim of this paper regarding large language models?
q3:What is one key finding about LMs' performance with less popular factual knowledge?
q4:How does scaling impact LMs' ability to memorize factual knowledge?
q5:What is the proposed method based on the findings of this paper?

[Despite their impressive performance on diverse tasks, large language models (LMs) still struggle with tasks requiring rich world knowledge, implying the difficulty of encoding a wealth of world knowledge in their parameters.]t1 [This paper aims to understand LMs' strengths and limitations in memorizing factual knowledge, by conducting large-scale knowledge probing experiments on two open-domain entity-centric QA datasets: PopQA, our new dataset with 14k questions about long-tail entities, and EntityQuestions, a widely used open-domain QA dataset.]t2 [We find that LMs struggle with less popular factual knowledge, and that retrieval augmentation helps significantly in these cases.]t3 [Scaling, on the other hand, mainly improves memorization of popular knowledge, and fails to appreciably improve memorization of factual knowledge in the tail.]t4 [Based on those findings, we devise a new method for retrieval-augmentation that improves performance and reduces inference costs by only retrieving non-parametric memories when necessary.]t5

Figure 2. Plan extraction

The VISTA Dataset

- **Scale:** 18,599 video-abstract pairs from leading AI conferences
- **Sources:** ACL Anthology (ACL, EMNLP, NAACL, EACL), ICML, NeurIPS (2020-2024)
- **Quality Control:** Manual validation (500 samples) + automated assessment (GPT-o1, All samples)

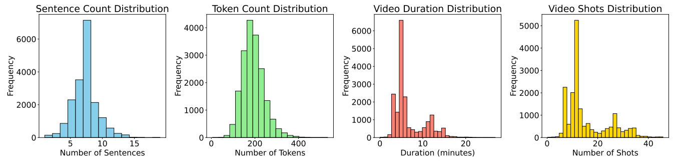


Figure 3. Dataset attribute distributions

- **Videos**: Avg. 6.76 minutes, 16.36 shots per video
- **Summaries**: Avg. 192.62 tokens, 7.19 sentences per summary
- **Complexity**: Dependency tree depth 6.02, TTR 0.62

Main Results

- **Plan-based superiority:** Planning model outperforms all baselines
- **Modality ranking:** Video + Audio > Video > Audio > Transcript
- **Modality interplay:** Video excels alone (rich cues), audio adds timing info, but transcripts are often noisy and hinder alignment
- **Planning benefit:** Planning also boosts summarization for text- and audio-only models

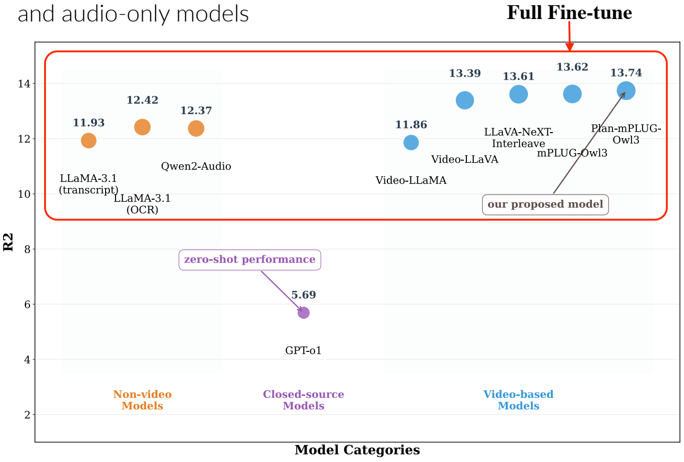


Figure 4. Model performance comparison

Human and GPT-o1 Evaluation

- **Multi-aspect assessment:** Faithfulness, Relevance, Informativeness, Conciseness, Coherence
- **Human superiority:** Humans consistently outperform all models across all evaluation criteria
- **Plan-based advantage:** Our proposed model achieves best performance among other models in both evaluations

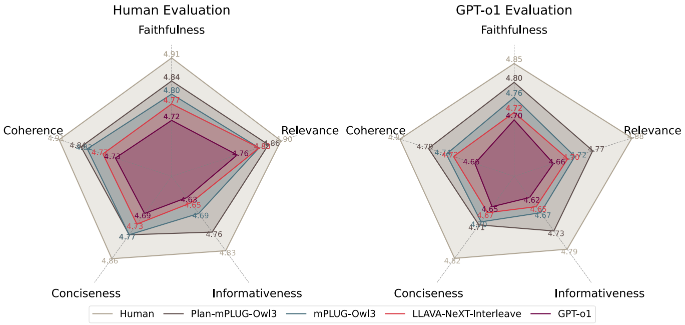


Figure 5. Human and GPT-o1 evaluation results

Conclusion

- **Dataset:** VISTA provides 18,599 video-summary pairs, a novel large-scale dataset for scientific video-to-text summarization
- **Benchmarking:** Comprehensive evaluation of 13+ SOTA LMMs across multiple settings (zero-shot, QLoRA, full fine-tuning)
- **Method:** Plan-based summarization improves quality and factual accuracy over strong multimodal baselines

Project Info

dongqi.me/
projects/VISTA

https://dongqi.me

Funded by the European Research Council (Grant No. 948878)

dongqi.me@gmail.com