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

ACL 2025 VIENNA

Introduction

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

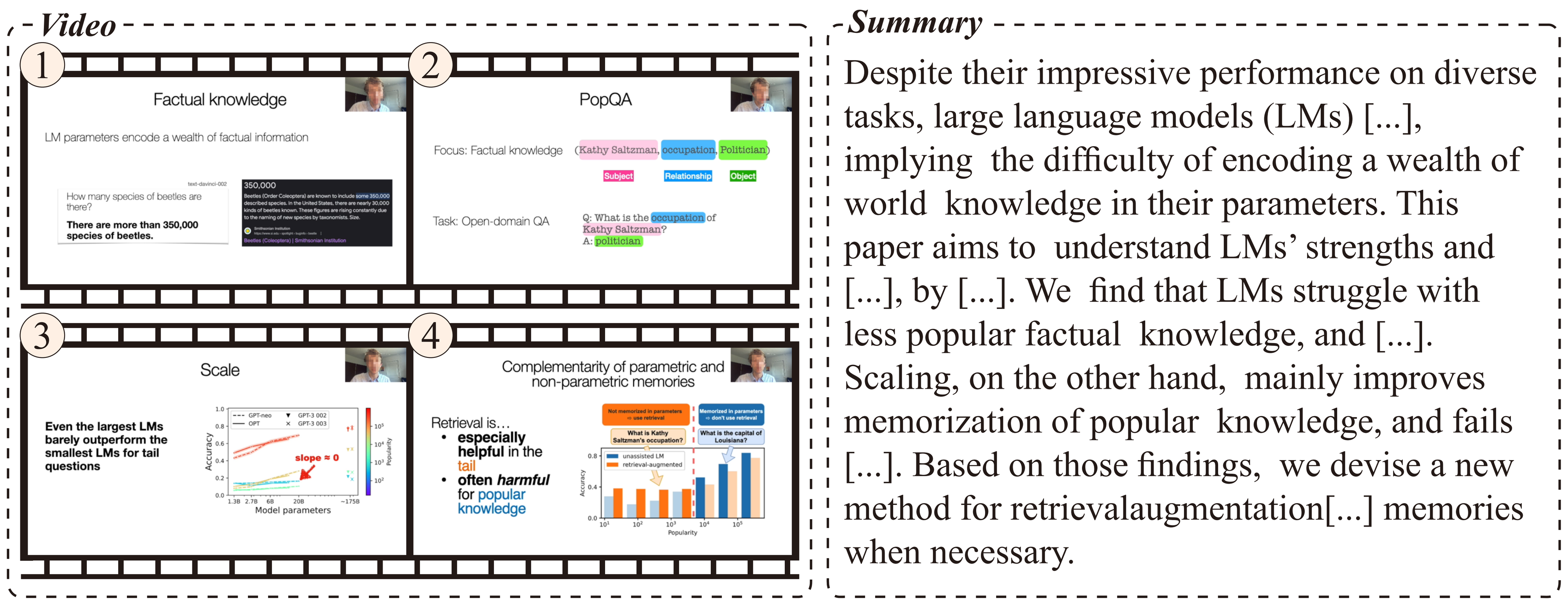

Figure 1. VISTA pairs presentation videos with paper abstracts

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:** Learn $P(s|v,p)$ (video v, summary s) instead of $P(s|v)$ mapping

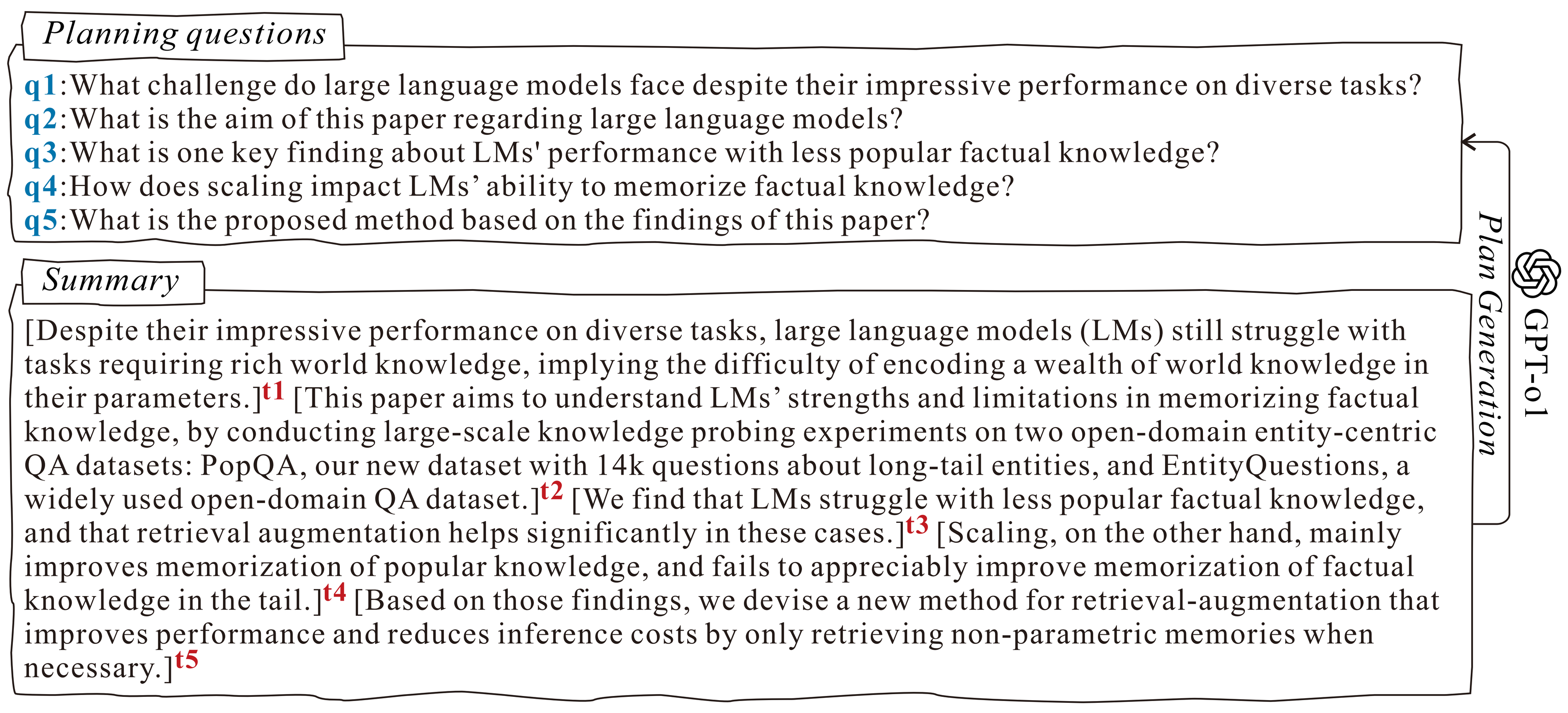

Figure 2. Plan extraction

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)
- **Data Splits:** Train (80%), Validation (10%), Test (10%)

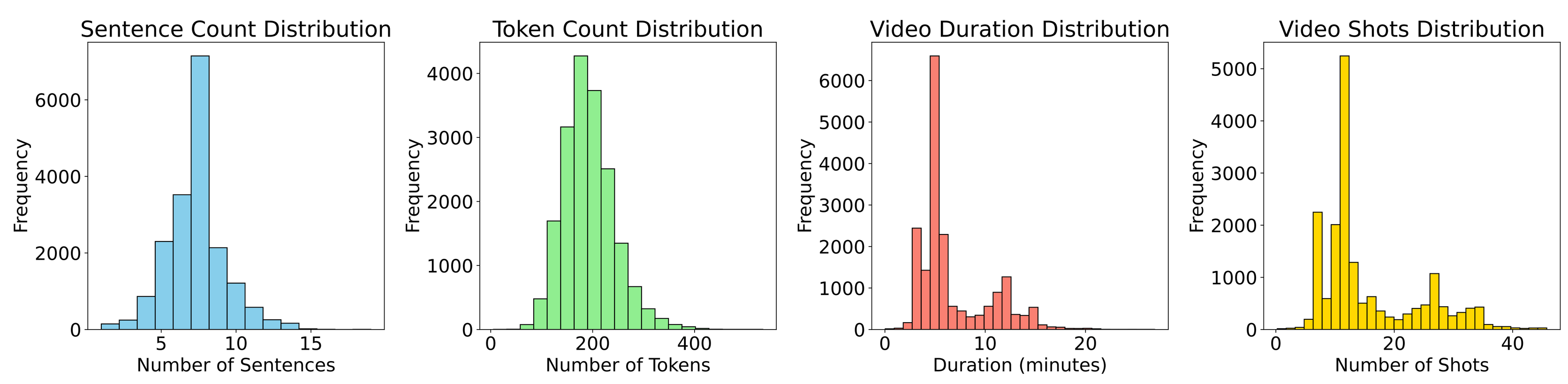

Figure 3. Dataset attribute distributions

- **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

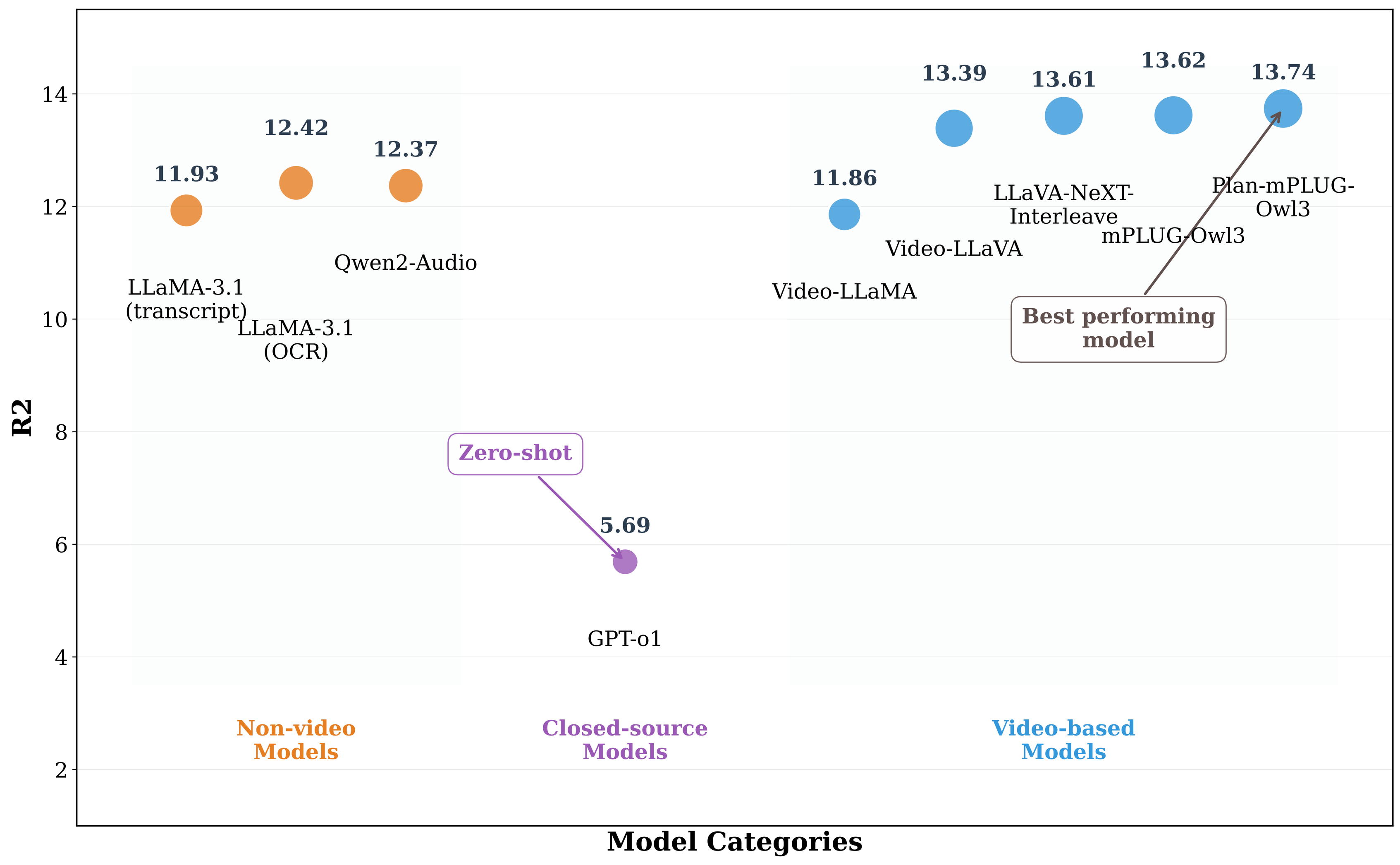

Figure 4. Model performance comparison

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:** Plan-mPLUG-Owl3 achieves best performance among other models in both evaluations

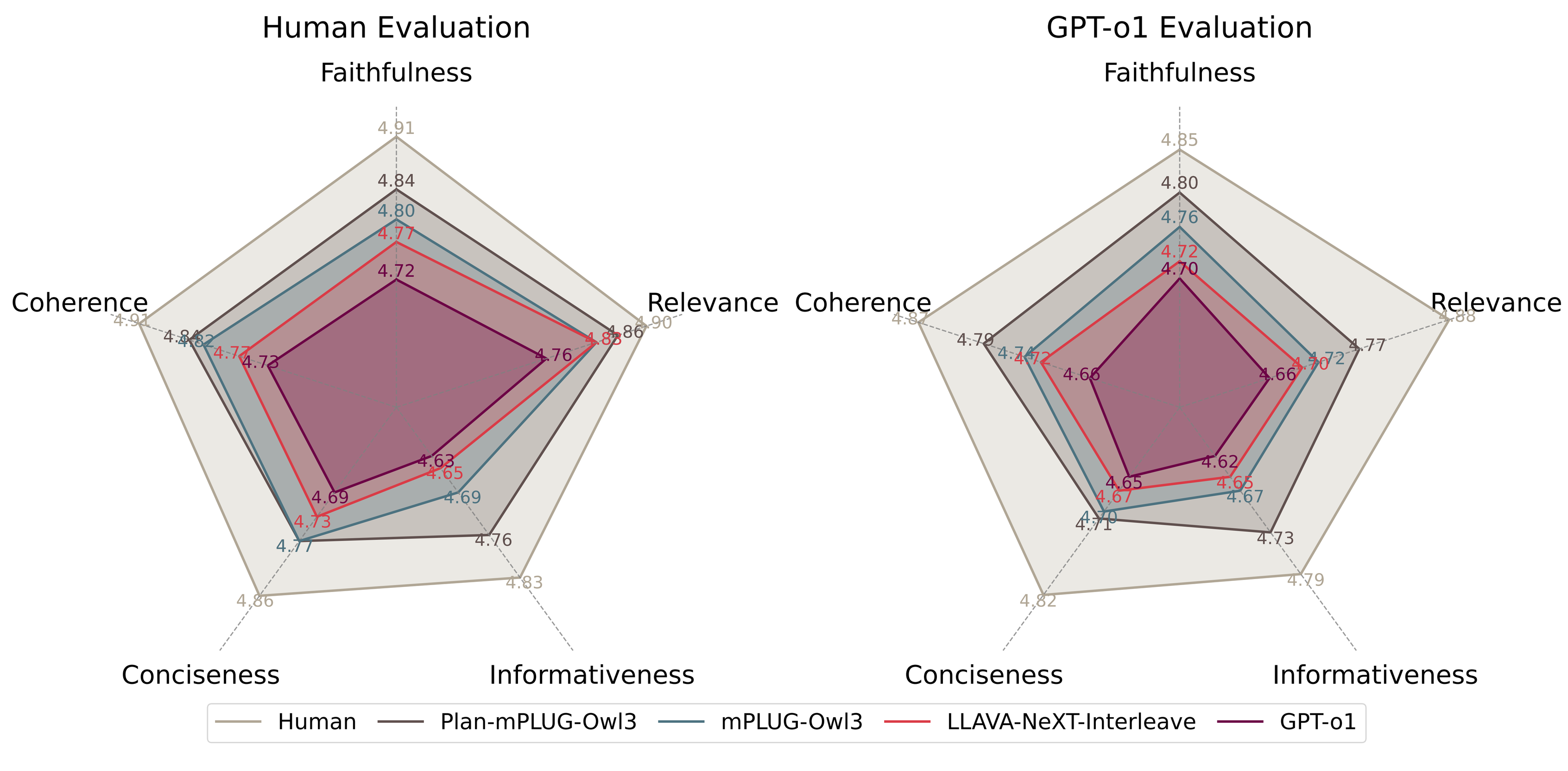

Figure 5. Human and GPT-o1 evaluation results

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