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

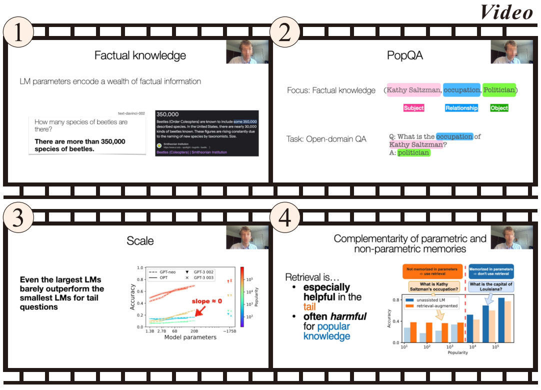
Dongqi Liu$^{\Omega*}$**, Chenxi Whitehouse**$^{\triangle}$**, Xi Yu**$^{\Omega}$**, Louis Mahon**$^{\ominus}$**, Rohit Saxena**$^{\ominus}$**,
Zheng Zhao**$^{\ominus}$**, Yifu Qiu**$^{\ominus}$**, Mirella Lapata**$^{\ominus}$**, Vera Demberg**$^{\Omega\Psi}$
$^{\Omega}$Saarland University, $^{\Psi}$Max Planck Institute for Informatics
$^{\triangle}$University of Cambridge, $^{\ominus}$University of Edinburgh
$^{\Omega}${dongqi,xiyu,vera}@lst.uni-saarland.de
$^{\triangle}$chenxi.whitehouse@cl.cam.ac.uk
$^{\ominus}${lmahon,rohit.saxena,zheng.zhao,yifu.qiu}@ed.ac.uk, mlap@inf.ed.ac.uk

Abstract

Transforming recorded videos into concise and accurate textual summaries is a growing challenge in multimodal learning. This paper introduces VISTA, a dataset specifically designed for video-to-text summarization in scientific domains. VISTA contains 18,599 recorded AI conference presentations paired with their corresponding paper abstracts. We benchmark the performance of state-of-the-art large models and apply a plan-based framework to better capture the structured nature of abstracts. Both human and automated evaluations confirm that explicit planning enhances summary quality and factual consistency. However, a considerable gap remains between models and human performance, highlighting the challenges of our dataset. This study aims to pave the way for future research on scientific video-to-text summarization. The project information is available at https://dongqi.me/projects/VISTA.

1 Introduction

Large multimodal models (LMMs), which integrate components from different modalities through cross-modal alignment training (Koh et al., 2023; Cheng et al., 2023; Li et al., 2024a; Ahn et al., 2024; Fu et al., 2025; Wu et al., 2025), have achieved considerable progress in video-to-text summarization tasks for general-purpose content such as YouTube, movies, and news videos (Li et al., 2020; Lin et al., 2023; Krubiński and Pecina, 2023; Hua et al., 2024; Chen et al., 2024a; Zhang et al., 2024a; Qiu et al., 2024; Patil et al., 2024; Mahon and Lapata, 2024a,b). However, many recent studies have highlighted that these LMMs exhibit reduced performance in scientific contexts, particularly when processing technical terminology and scientific visual elements like figures and tables (Li et al., 2024b; Lu et al., 2024; Yue et al., 2024; Bai et al., 2024; Liang et al., 2024; Patil et al., 2024; Huang et al., 2024). This performance gap might be largely attributed to the absence of specialized datasets for multimodal scientific content (Chen et al., 2024c; Hu et al., 2024; Pramanick et al., 2024; Zhang et al., 2024b).


Video
1
Factual knowledge
LM parameters encode a wealth of factual information
text-davinci-002
How many species of beetles are there?
There are more than 350,000 species of beetles.
350,000
2
PopQA
Focus: Factual knowledge
(Kathy Saltzman, occupation, Politician)
Subject
Relationship
Object
Task: Open-domain QA
Q: What is the occupation of Kathy Saltzman?
A: politician
3
Scale
Even the largest LMs barely outperform the smallest LMs for tail questions
GPT-neo
OPT
GPT-3 002
GPT-3 003
slope ≈ 0
Accuracy
Model parameters
Popularity
4
Complementarity of parametric and non-parametric memories
Retrieval is…
especially helpful in the tail
often harmful for popular knowledge
Not memorized in parameters ⇨ use retrieval
Memorized in parameters ⇨ don't use retrieval
What is Kathy Saltzman's occupation?
What is the capital of Louisiana?
unassisted LM
retrieval-augmented
Accuracy
Popularity

Summary

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 retrieval-augmentation[...] memories when necessary.

Figure 1: An example from VISTA: a conference presentation video (top) paired with the abstract of the corresponding paper (bottom). This data sample (Mallen et al., 2023) was presented at ACL 2023 and received the Best Video Recordings award.

Thus, we introduce **VISTA** (**Vi**deo to **S**cien**t**ific **A**bstract), an English dataset for video-to-text summarization in scientific domains. VISTA consists of 18,599 aligned pairs of conference presentation recordings and their corresponding paper abstracts, collected from leading conferences in computational linguistics (ACL Anthology including ACL, EMNLP, NAACL, EACL, Findings of *ACL) and machine learning (ICML and NeurIPS). Figure 1

[*] This work was partially conducted during the research visit at the University of Edinburgh.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6187–6210
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

illustrates an example selected from VISTA.

We use the abstract of the paper as a proxy for the summary of the video and benchmark VISTA using several state-of-the-art (SOTA) large models, including closed-source LMMs (`Claude 3.5 Sonnet`, `Gemini 2.0`, `GPT-o1`), as well as video-specific open-source LMMs (`Video-LLaMA`, `Video-ChatGPT`, `mPLUG-Owl3`, etc.; Zhang et al., 2023; Maaz et al., 2024; Lin et al., 2024a; Ye et al., 2025; Li et al., 2025, 2024c). For comparison, we also include strong baselines: text-to-text model `LLaMA-3.1` (Touvron et al., 2023) and audio-to-text model `Qwen2-Audio` (Chu et al., 2024). Experiments across zero-shot, QLoRA, and full fine-tuning settings reveal that in-domain fine-tuning improves summarization performance across different large models, and video-based models generally outperform text- and audio-based models on our dataset. However, end-to-end approaches may often struggle to capture the underlying structure of scientific abstracts (Liu et al., 2025).

To address this, we explore a plan-based approach, which has been shown to improve coherence and factual grounding through a predefined planning component (Narayan et al., 2021, 2023; Liu et al., 2025). Unlike direct end-to-end generation, plan-based methods can leverage the fact that scientific abstracts often follow a well-defined format (Takeshita et al., 2024). By explicitly modeling the latent structure of the summary through a sequence of intermediate plans, the summary generation process can be better guided. Empirical results confirm that the plan-based method outperforms existing SOTA models in terms of summary quality and factual accuracy. This work also lays the groundwork for future investigations into the multimodal summarization of scientific videos.

In summary, our contributions are as follows:

- We present VISTA, a novel large-scale multimodal dataset with 18,599 video-summary pairs, tailored for summarizing scientific presentations from video recordings.
- We establish benchmark performance on VISTA through a comprehensive evaluation of leading large (language/audio/multimodal) models.
- We leverage a plan-based approach that consistently improves summary quality and factual accuracy over SOTA models.
- We conduct error analysis, case studies, and human evaluations to identify the pivotal issues in the model-generated summaries.

2 Related Work

Video-to-Text Summarization generates coherent summaries by integrating multimodal information (Hua et al., 2024), supported by datasets like MSS (Li et al., 2017), VideoXum (Lin et al., 2024b), MMSum (Qiu et al., 2024), Hierarchical3D (Papalampidi and Lapata, 2023), and LfVS-T (Argaw et al., 2024), spanning tasks from instructional videos to general web content (Li et al., 2017; Zhou et al., 2018; Li et al., 2019, 2020; Liu and Wan, 2021; Fu et al., 2021; Liu et al., 2022; Krubiński and Pecina, 2023; Han et al., 2025; He et al., 2023; Hua et al., 2024; Islam et al., 2024; Qiu et al., 2024). Technical advancements include hierarchical attention models (Sanabria et al., 2018), extractive methods using multimodal features (Cho et al., 2021; Krubiński and Pecina, 2023), and hybrid extractive-abstractive frameworks (Ramakrishnan and Ngan, 2022; Papalampidi and Lapata, 2023). Transformer-based systems have further improved performance (Krubiński and Pecina, 2023; Li et al., 2020; Shang et al., 2021; Mahon and Lapata, 2024a). However, challenges in summarizing academic videos remain under-explored.

Scientific Text Summarization condenses complex scholarly content into concise formats (Cachola et al., 2020; Ju et al., 2021; Liu et al., 2023b; Liu and Demberg, 2023), supported by datasets like TalkSumm (Lev et al., 2019) for academic video transcripts, SumSurvey (Liu et al., 2024b) for survey papers, ACLSum (Takeshita et al., 2024) for ACL discourse, and SciNews (Liu et al., 2024a) for simplifying research for broader audiences. M^{3}AV (Chen et al., 2024c) supports tasks like ASR, TTS, and slide-script generation. Methods like RST-LoRA (Liu and Demberg, 2024) and RSTformer (Liu et al., 2023b) improve discourse and structural summarization, while CiteSum (Mao et al., 2022) and SSR (Fatima and Strube, 2023) focus on scalability and audience-specific customization. Despite these efforts, scientific summarization remains a challenging domain due to the inherent complexity and diversity of scholarly texts.

Plan-based Summarization employs structured representations to improve summary quality and reduce hallucinations (Narayan et al., 2021; Amplayo et al., 2021; Wang et al., 2022; Narayan et al., 2023; Liu et al., 2025). Research focuses on text-based planning with elements like entities (Narayan et al., 2021; Liu and Chen, 2021; Huot et al., 2024), keyword prompts (Creo et al., 2023), and question-answer pairs (Narayan et al., 2023). Examples include PlanVerb (Canal et al., 2022), which converts task plans into natural language via semantic tagging, and domain-specific approaches that align with knowledge structures for improved quality (Srivastava et al., 2024). Blueprint-based frameworks utilize intermediate plans to create coherent narratives for visual storytelling (Liu et al., 2023a). However, plan-based strategies for multimodal tasks, particularly video-to-text summarization, have received limited attention.

3 The VISTA Dataset

Data Acquisition and Cleaning VISTA is derived from computational linguistics and machine learning conferences, including ACL Anthology (ACL, EMNLP, NAACL, EACL, Findings of *ACL), ICML, and NeurIPS, covering content from 2020 to 2024. All materials (paper abstracts and video recordings) are contributed by the respective paper authors, ensuring narrative consistency. Since these metadata are stored in XML/JSON files on their respective websites, no further data preprocessing (e.g., extracting abstracts from PDFs) is required. We collect paper titles, author lists, paper abstracts, links to papers, and presentation videos, in accordance with platform terms for academic research purposes (or obtain written confirmation).[1] To maintain one-to-one video-to-text alignments, we exclude samples that may cover multiple papers (e.g., tutorials, invited talks) and videos shorter than one minute or longer than 30 minutes.

Quality Control We verify the data quality through both manual and automated checks. We discuss quality control guidelines and the results in Appendix Figure 10 and Appendix B, respectively.

Data Splits After quality control, our dataset comprises 18,599 samples, with venue distributions shown in Figure 2. To ensure balanced domain coverage in each subset, we proportionally sample to split the dataset into training (80%), validation (10%), and test (10%) sets. All subsequent experiments are conducted using these splits.

Dataset Comparison and Statistics Table 1 compares VISTA with several existing video-to-text summarization datasets. While many focus on open-domain (e.g., MMSum, Instruct-V2Xum) or areas like news (MLASK, MM-AVS) and activities (VideoXum), VISTA is tailored for summarizing scientific presentations. On average, it features longer inputs (6.8 minutes) than VideoXum (2.1 minutes) and MSS (3.4 minutes), as well as longer summaries (192.6 tokens), compared to YouCook2 (67.8 tokens) and VideoXum (49.9 tokens).

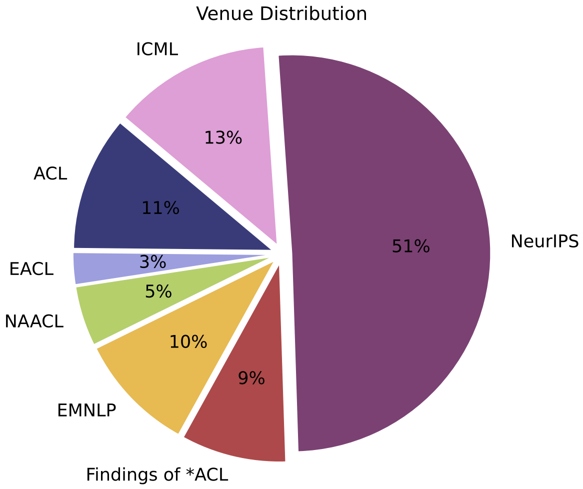


Figure 2: Venue distribution of the VISTA dataset.

Table 2 summarizes the VISTA dataset statistics: Videos average 6.76 minutes and 16.36 shots (we use PySceneDetect with `ContentDetector` to calculate video shots), while summaries contain 192.62 tokens on average across 7.19 sentences. The average dependency tree depth (Avg. Depth of Dep Tree) is 6.02, indicating the syntactic complexity of the summaries. Meanwhile, the Type-Token Ratio (TTR) is 0.62, reflecting lexical diversity. Both metrics are calculated using spaCy. Diversity metrics (Li et al., 2016), which measure the variety of unique n-grams, yield Distinct-1, Distinct-2, and Distinct-3 scores of 0.62, 0.93, and 0.97, respectively. Figure 3 visualizes key attributes: Most summaries remain under 250 tokens and 10 sentences, and most videos last fewer than 10 minutes with under 30 shots. In Appendix C, we present a random sample from the VISTA dataset.

4 Benchmarking VISTA

Task Overview We formalize the task of summarizing recorded scientific videos as follows: Let v and s denote a video (or its transcript/audio) and its paired summary from dataset $D = \{(v_1, s_1), (v_2, s_2), \ldots, (v_n, s_n)\}$, where n signifies the number of video-summary pairs. The objective is to train a (multimodal) model $\mathcal{M}$ to learn the conditional probability distribution

[1] We discuss copyright in Appendix A.

Dataset	Language	Domain	#Videos	VideoLen	SumLen
MSS (Li et al., 2017)	English, Chinese	News	50	3.4	—
YouCook2 (Zhou et al., 2018)	English	Cooking	2.0K	5.3	67.8
VideoStorytelling (Li et al., 2019)	English	Open	105	12.6	162.6
VMSMO (Li et al., 2020)	Chinese	Social Media	184.9K	1.0	11.2
MM-AVS (Fu et al., 2021)	English	News	2.2K	1.8	56.8
MLASK (Krubiński and Pecina, 2023)	Czech	News	41.2K	1.4	33.4
VideoXum (Lin et al., 2023)	English	Activities	14.0K	2.1	49.9
Shot2Story20K (Han et al., 2025)	English	Open	20.0K	0.3	201.8
BLiSS (He et al., 2023)	English	Livestream	13.3K	5.0	49.0
SummScreen^{3D} (Papalampidi and Lapata, 2023)	English	Open	4.5K	40.0	290.0
Ego4D-HCap (Islam et al., 2024)	English	Open	8.3K	28.5	25.6
Instruct-V2Xum (Hua et al., 2024)	English	Open	30.0K	3.1	239.0
MMSum (Qiu et al., 2024)	English	Open	5.1K	14.5	21.7
LfVS-T (Argaw et al., 2024)	English	YouTube	1.2K	12.2	—
VISTA (ours)	English	Academic	18.6K	6.8	192.6

Table 1: Comprison of video-to-text summarization datasets. #Videos = the number of videos, whereas VideoLen and SumLen refer to the average of video duration (in minutes) and the average number of summary tokens.

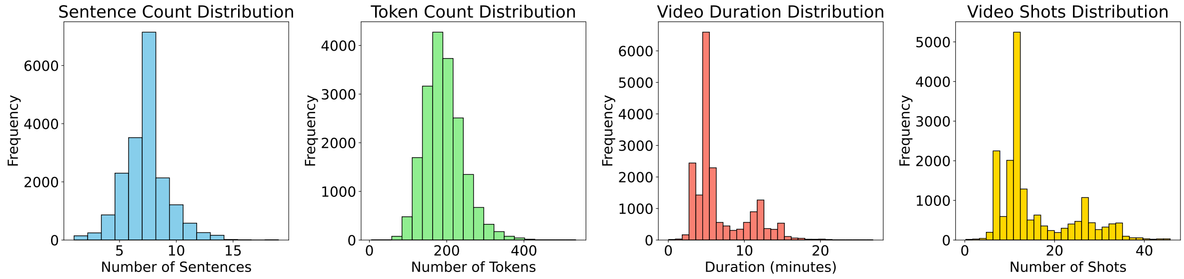


Figure 3: Distribution of summary sentences, summary tokens, video durations, and video shots in VISTA.

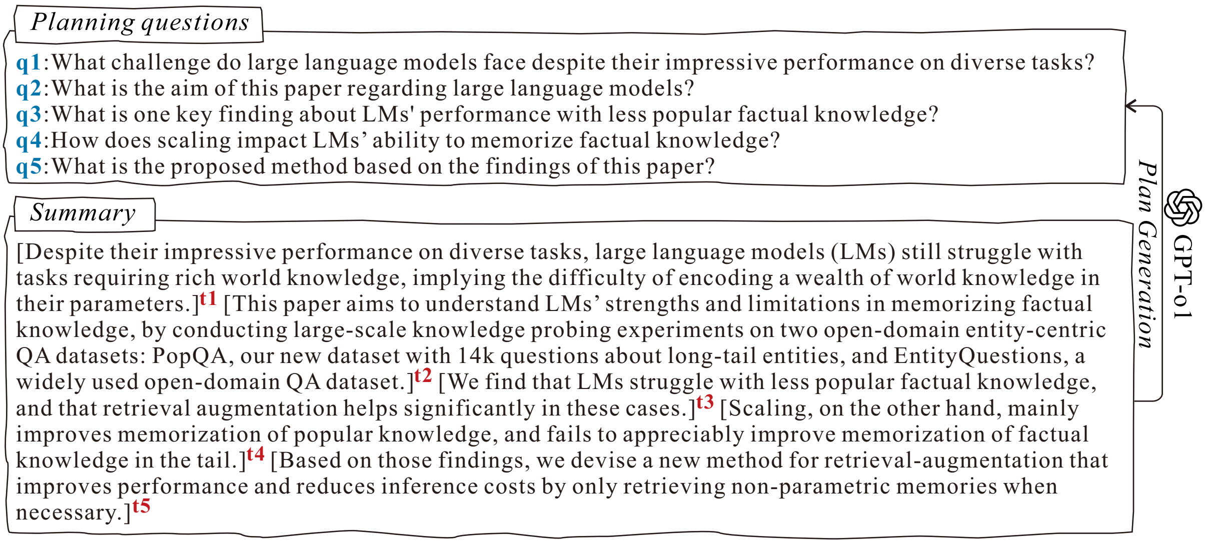
Training / Validation / Test Set	14,881 / 1,859 / 1,859
Avg. Video Length (mins) / Shots	6.76 / 16.36
Avg. #Summary Sent / Tokens	7.19 / 192.62
Avg. Depth of Dep Tree	6.02
Type-Token Ratio	0.62
Distinct-1 / -2 / -3	0.62 / 0.93 / 0.97

Table 2: Key statistics of the VISTA dataset, showcasing the average video length and shot count, summary characteristics (sentence and token counts), syntactic complexity (dependency tree depth), and lexical diversity (Type-Token Ratio and Distinct n-gram scores).

$P(s \mid v)$. Given a new video, the trained model $\mathcal{M}$ is expected to generate an appropriate summary.

A challenge in video-to-text summarization is structuring the generated summaries in a coherent and faithful manner. Directly learning the mapping from v to s could lead to inadequate outputs, as the model lacks explicit guidance on how to organize and present the extracted information (Mahon and Lapata, 2024a). Scientific abstracts often follow a relatively well-defined structure, making them suitable for a more structured generation approach (Takeshita et al., 2024). We follow previous work (Narayan et al., 2021, 2023) in adopting a plan-based framework that introduces an intermediate representation to capture latent structure more effectively than simpler end-to-end approaches. Specifically, given input v, we first generate a plan p, which consists of a sequence of automatically generated questions $\{q_1, q_2, \ldots, q_m\}$, each corresponding to a sentence to be verbalized in the summary. The plan explicitly controls the structure of the summary as a whole and the content of each of its sentences (which are meant to answer the questions in the plan). The model is then trained to learn the extended conditional probability distribution $P(s \mid v, p)$, ensuring that the generated summaries follow the structure and flow of plan p.

Plan Generation We hypothesize that summary sentences can be viewed as responses to plan questions, where the plan consists of an ordered sequence of questions directly associated with the target content. This idea is inspired by the theory of Question Under Discussion (QUD; Roberts (2012); Wu et al. (2023b); Suvarna et al. (2024)), which



Planning questions

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?

Summary

[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 4: GPT-o1 generates plans based on reference summaries. Each question q_i corresponds to a summary sentence t_i, which we assume constitutes its answer. Index i ranges from 1 to the number of summary sentences.

posits that discourse often revolves around a set of questions that guide the structure and interpretation of the conversation.

We leverage GPT-o1 (Achiam et al., 2023) to generate silver-standard plans based on reference summary sentences and their preceding context. As shown in Figure 4, for example, question q_3 is generated based on target sentence t_3 and the summary sentences preceding it (i.e., t_1 and t_2), and so on. As a result, the question sequence preserves the order of sentences in the reference summaries, ensuring that the plan maintains a natural and coherent flow consistent with the structure of reference summaries. The prompt used to generate plan questions is provided in Appendix Figure 12. We discuss the quality of the silver-standard plans through manual investigation in Appendix G.

Summarization Model We train two independent modules corresponding to Plan Generation (PG) and Summary Generation (SG). The PG module is trained on pairs of (v, p) samples. The SG module is trained on tuples $([v; p], s)$, where $[v; p]$ is the concatenation of the input v and its plan p. During inference, the trained PG module predicts plan $\hat{p}$ for input v, and the tuple $[v; \hat{p}]$ is fed into the SG module to generate the final summary. Both modules have the same backbone but are trained independently.

5 Experiments

Baseline Models We benchmark our dataset using three learning settings: Zero-shot learning, QLoRA fine-tuning (Dettmers et al., 2024), and full-parameter fine-tuning. For zero-shot learning, we test closed-source multimodal models, including GPT-o1 (Achiam et al., 2023), Gemini 2.0 (Team et al., 2023), Claude 3.5 Sonnet (Anthropic, 2024), as well as open-source video LMMs such as Video-LLaMA (Zhang et al., 2023), Video-ChatGPT (Maaz et al., 2024), Video-LLaVA (Lin et al., 2024a), LLaMA-VID (Li et al., 2024c), LLaVA-NeXT-Interleave (Li et al., 2025), and mPLUG-Owl3 (Ye et al., 2025). These open-source video LMMs process videos by extracting multimodal features, such as visual and/or audio components, using cross-modal attention mechanisms to align and integrate information across modalities.

We also assess LLaMA-3.1 (Touvron et al., 2023) and Qwen2-Audio (Chu et al., 2024) to examine if text- or audio-based models can accomplish the summarization task without taking video information into account. For LLaMA-3.1, we explore two variants: In LLaMA-3.1$_{transcript}$, we extract audio from video files using moviepy and transcribe it with OpenAI's Whisper-1 to generate text input for the model. In LLaMA-3.1$_{OCR}$, we apply EasyOCR to extract on-screen text from video frames and use the OCR-generated text as input for summarization. Similarly, for Qwen2-Audio, we use moviepy to convert video files into audio and treat the audio as input. Exact model versions are provided in Appendix D. Based on our benchmarking results, we select the best-performing model as the backbone for the plan-based strategy and evaluate its performance. Prompts for the above models

Method	Model	Open-source	R1	R2	RLsum	SacreBLEU	Meteor	BERTscore	CIDEr-D	VideoScore	FactVC
Zero-shot Learning	LLaMA-3.1$_{transcript}$	✓	23.68	4.22	21.39	2.70	14.62	80.93	1.17	1.53	34.32
	LLaMA-3.1$_{OCR}$	✓	24.02	4.37	21.42	2.63	14.59	80.33	1.19	1.50	34.06
	Qwen2-Audio	✓	23.52	4.29	21.53	2.49	14.77	80.62	1.15	1.59	34.31
	Claude 3.5 Sonnet	✗	27.71	5.59	24.14	3.14	17.53	82.57	1.32	1.91	50.11
	Gemini 2.0	✗	27.82	5.66	24.29	4.22	17.83	82.64	1.47	2.02	52.02
	GPT-o1	✗	27.90	5.69	24.37	4.38	17.90	82.63	1.61	2.17	51.36
	Video-LLaMA	✓	20.18	3.19	21.24	1.76	13.73	81.31	1.08	1.63	32.25
	Video-ChatGPT	✓	20.36	3.52	21.43	1.79	14.01	81.35	1.11	1.63	33.21
	Video-LLaVA	✓	25.29	4.50	22.52	2.82	15.13	81.39	1.17	1.65	36.45
	LLaMA-VID	✓	25.31	4.77	22.53	2.88	15.27	81.32	1.14	1.64	36.39
	LLaVA-NeXT-Interleave	✓	25.41	4.82	22.68	2.92	15.25	81.40	1.18	1.73	40.12
	mPLUG-Owl3	✓	25.57	4.82	22.84	2.99	15.33	81.39	1.21	1.77	42.07
	Plan-mPlug-Owl3*	✓	**25.62**†	**4.95**†‡	**22.97**†‡	**3.14**†‡	**15.39**†‡	**81.45**‡	**1.27**†‡	**1.86**†‡	**47.37**†‡
QLoRA Fine-tuning	LLaMA-3.1$_{transcript}$	✓	32.24	11.38	30.39	8.03	21.57	82.39	3.86	2.81	53.22
	LLaMA-3.1$_{OCR}$	✓	33.01	12.11	30.52	8.04	21.55	82.41	3.92	2.77	53.19
	Qwen2-Audio	✓	32.17	12.05	30.77	7.87	21.86	82.36	4.11	2.80	54.27
	Video-LLaMA	✓	30.74	9.44	28.33	6.45	22.49	82.61	3.99	2.77	52.05
	Video-ChatGPT	✓	31.68	10.50	30.40	7.63	23.67	82.62	4.02	2.78	55.02
	Video-LLaVA	✓	33.16	12.64	30.37	8.17	23.92	82.81	4.26	2.83	59.13
	LLaMA-VID	✓	33.31	12.73	30.49	8.22	23.90	83.01	4.31	2.88	62.20
	LLaVA-NeXT-Interleave	✓	33.37	12.77	30.56	8.30	23.95	83.47	4.47	2.93	66.14
	mPLUG-Owl3	✓	33.40	12.82	30.66	8.29	23.97	83.49	4.47	2.92	70.08
	Plan-mPlug-Owl3	✓	**33.52**†‡	**13.01**†‡	**31.10**†‡	**8.33**	**24.11**†‡	**83.53**†	**4.52**	**3.11**†‡	**73.11**†‡
Full Fine-tuning	LLaMA-3.1$_{transcript}$	✓	33.37	11.93	30.86	8.27	25.12	83.71	4.87	3.21	63.38
	LLaMA-3.1$_{OCR}$	✓	34.02	12.42	31.72	8.51	15.11	84.09	4.89	3.32	65.84
	Qwen2-Audio	✓	33.82	12.37	31.63	8.33	25.09	83.62	4.83	3.22	66.62
	Video-LLaMA	✓	32.19	11.86	31.68	8.41	24.99	83.83	4.77	3.04	64.21
	Video-ChatGPT	✓	32.47	12.11	32.21	8.72	25.09	83.91	4.82	3.11	66.09
	Video-LLaVA	✓	33.28	13.39	32.78	9.10	25.42	83.97	4.87	3.13	66.12
	LLaMA-VID	✓	33.47	13.53	32.80	9.21	25.41	84.03	4.91	3.17	68.30
	LLaVA-NeXT-Interleave	✓	33.75	13.61	32.88	9.26	25.63	84.11	5.01	3.23	73.42
	mPLUG-Owl3	✓	34.22	13.62	32.91	9.32	25.72	84.22	5.03	3.28	71.94
	Plan-mPlug-Owl3	✓	**34.53**†‡	**13.74**†‡	**33.25**†‡	**9.56**†‡	**25.88**†‡	**84.37**†‡	**5.15**†‡	**3.33**†‡	**75.41**†‡

Table 3: Model performance on VISTA dataset. In Plan-mPlug-Owl3*, only the PG module is trained. Plans generated by the PG on the test set serve as input to the SG module for zero-shot inference (no training is applied to the SG module). Symbols † and ‡ indicate that the performance of Plan-mPlug-Owl3 is significantly ($p < 0.05$) different from LLaVA-NeXT-Interleave (third best) and mPLUG-Owl3 (second best), when using the paired t-test.

are offered in Appendix M (Figures 11–14).

Experimental Setup To ensure a fair comparison, all models, including baselines, plan-based models, and ablation models, are evaluated under identical hyperparameter settings unless explicitly stated otherwise. All models are tested using identical prompt instructions. Detailed hyperparameter configurations are presented in Appendix E.

Evaluation Metrics We report a set of evaluation metrics to measure informativeness, alignment, and factual consistency in summaries. For informativeness, we utilize ROUGE (Lin, 2004), SacreBLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), and CIDEr-D (Vedantam et al., 2015). Specifically, we provide the F1 scores for Rouge-1 (R1), Rouge-2 (R2), and Rouge-LSum (RLSUM). Alignment to the input video is evaluated with VideoScore (He et al., 2024), and factual consistency with FactVC (Liu and Wan, 2023). Detailed descriptions of these metrics are given in Appendix F.

6 Results and Analysis

General Results Table 3 compares model performance across three learning settings: Zero-shot, QLoRA fine-tuning, and full-parameter fine-tuning. Overall, fine-tuning on in-domain data yields substantial performance gains across all evaluation metrics. Full fine-tuning consistently outperforms QLoRA. While closed-source models such as GPT-o1 and Gemini typically lead in zero-shot performance, open-source models like mPLUG-Owl3 and Plan-mPlug-Owl3 achieve competitive or even superior results when fine-tuned, especially in semantic alignment (BERTScore) and video-text consistency (VideoScore).

We observe that video-based LMMs consistently outperform text-based and audio-based models. While models such as LLaMA-3.1$_{transcript}$, LLaMA-3.1$_{OCR}$, and Qwen2-Audio yield comparable results, they lag behind video-grounded models in overall performance. In particular, mPLUG-Owl3 achieves SOTA results across most metrics, highlighting the crucial role of visual information in enhancing summarization quality.

Plan-mPlug-Owl3 is the plan-based approach built on mPLUG-Owl3, outperforming all open-source baselines in both zero-shot and fine-tuned settings. For zero-shot inference, the Plan-mPlug-Owl3* variant, which fine-tunes only the Plan Generation (PG) module, surpasses other models in summary quality, factual consistency, and semantic alignment. With full-parameter fine-tuning, Plan-mPlug-Owl3 achieves the highest overall scores across models, showing improvements in factual accuracy (+3.47 in FactVC) and quality (+0.34 in RLsum) compared to mPLUG-Owl3. However, all models (including the plan-based method) exhibit hallucinations (FactVC) and alignment (VideoScore) issues, and there are still significant differences (p-value of the paired t-test is less than 0.05) between the human performance in this task, with reference summaries scoring 88.54 on FactVC and 4.62 on VideoScore.

Impact of Modality Interplay To explore the impact of different modality combinations on our multimodal tasks, we conduct an experiment using Video-LLaMA (Zhang et al., 2023). Seven modality combinations are considered, including unimodal inputs (video, audio, transcript) and their pairwise or joint combinations. For each configuration, only the corresponding modality modules are updated while the remaining ones are kept frozen. The summarized results are shown in Table 4.

The results consistently show that video is the strongest standalone modality, likely due to its rich spatial-temporal information. Audio offers complementary prosodic and timing cues, but lacks semantic visual grounding. The transcript, while semantically rich, often introduces long, noisy, and unstructured inputs, particularly from ASR systems, that can overwhelm the model's attention and interfere with alignment. These findings suggest that current video-based LMMs face challenges in effectively aligning and fusing token-heavy, noisy textual inputs with corresponding visual or audio information.

Impact of Plan Generation Ablations We analyze the plan generation ablation by comparing it with simpler baselines: Lead-3_Q, Tail-3_Q, and Random-3_Q. In these ablation baselines, plans are generated by selecting the first three, last three, or three randomly chosen summary sentences, respectively. Each selected sentence serves as a target for generating a question, with its preceding sentences providing the context. For instance, in the Lead-3_Q setting, the first sentence is used as the target (without any preceding context), prompting the first question in the plan, while subsequent sentences incorporate earlier ones as context. Additionally, we compare the case where QUD is not considered. That is, we directly let GPT-o1 generate all plan questions at once based on the reference summary (NoQUD).

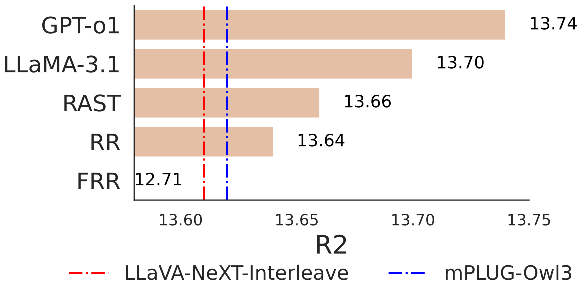


Figure 5: Noise in plan generation impacts summarization performance. FRR is a shorthand for Full Random Replacement, and RR for Random Replacement. RAST is a SOTA question generation method.

Table 5 underlines the performance differences across different plan generation ablations. For NoQUD, it underperforms compared to the QUD-based approach. The Lead-3_Q strategy performs better overall compared to Tail-3_Q and Random-3_Q, indicating that initial sentences offer stronger contextual continuity for generating plan questions.

Impact of Plan Quality We assess how the quality of the plan questions affects model performance. We apply GPT-o1 as a question generator in a zero-shot setting in our previous experiments. For comparative analysis, we additionally incorporate Llama-3.1 and a state-of-the-art question generation algorithm (RAST) from Gou et al. (2023) to generate the plan questions. In addition, we apply a Random Replacement (RR) method, where questions generated by GPT-o1 are randomly replaced with irrelevant ones. The number of replaced questions per summary ranges from one to the entire set. We also introduce full random replacement (FRR), where questions generated by GPT-o1 are all replaced with random irrelevant questions.[2]

Figure 5 reveals that the quality of plan questions does influence the summarization performance: Using GPT-o1 to generate questions outperforms

[2] The prompt for generating irrelevant questions is given in Appendix Figure 15.

Modality	Zero-shot Learning				QLoRA Fine-tuning				Full Fine-tuning			
	R2	RLsum	VideoScore	FactVC	R2	RLsum	VideoScore	FactVC	R2	RLsum	VideoScore	FactVC
Video only	2.68	20.34	1.55	28.93	8.83	27.51	2.65	50.66	10.78	30.02	2.91	60.87
Audio only	2.14	19.72	1.41	26.84	7.52	26.34	2.48	45.79	9.23	27.93	2.73	58.02
Transcript only	2.02	18.01	1.34	25.53	6.91	24.33	2.39	44.87	8.44	25.81	2.35	54.11
Video + Audio	**3.19**	**21.24**	**1.63**	**32.25**	**9.44**	**28.33**	**2.77**	**52.05**	**11.86**	**31.68**	**3.04**	**64.21**
Video + Transcript	1.87	18.94	1.39	27.76	7.35	24.82	2.51	48.63	9.01	27.19	2.65	58.91
Audio + Transcript	1.64	18.55	1.35	27.48	7.23	24.73	2.38	47.15	8.57	25.82	2.54	55.39
Video + Audio + Transcript	1.92	19.13	1.47	28.60	7.37	25.29	2.52	50.72	9.22	27.21	2.61	59.30

Table 4: Performance comparison of different modality combinations.

Model	R2	RLsum	VideoScore	FactVC
`Plan-mPlug-Owl3`	13.74	33.25	3.33	75.41
`NoQUD`	13.66	33.02	3.28	73.32
Lead-3$_Q$	12.87	30.64	2.95	71.26
Tail-3$_Q$	11.62	30.51	2.88	63.82
Random-3$_Q$	11.57	30.48	2.87	64.28

Table 5: Performance comparison of different plan generation ablations under full fine-tuning settings.

the rest. The FRR method performs the worst, as irrelevant questions disrupt the alignment between the plan and summary content. We also find that the plan-based method exhibits a certain degree of robustness, as it performs reasonably well even when the plans contain some degree of noise (RR vs. FRR). These findings emphasize the importance of question relevance and quality in structuring the output summaries.

Planning Beyond Vision While our primary objective is to evaluate the planning framework in the context of video-to-text summarization, it is valuable to assess its applicability to unimodal, non-visual models. To this end, we conduct supplementary experiments applying the planning method to three models that do not utilize video inputs: (1) `LLaMA-3.1`$_{transcript}$ (ASR-based textual input), (2) `LLaMA-3.1`$_{OCR}$ (OCR-based textual input), and (3) `Qwen2-Audio` (audio-based input). For each model, we compare baseline performance (i.e., without planning) against the planning counterpart. As summarized in Table 6, planning consistently improves performance across all settings and evaluation metrics. A paired t-test confirms that these improvements are statistically significant ($p < 0.05$).

These findings demonstrate that the planning method does not function solely as a domain-specific enhancement but rather as a generalizable scaffold that supports better discourse structure, even in the absence of visual input. We hypothesize that, for text- and audio-based models, planning mitigates the lack of spatial-temporal signals by providing discourse-level anchors, such as intent-driven prompts (e.g., “What problem is being addressed?”), that guide the model’s summarization trajectory.

Notably, despite these gains, video-based planning models such as `Plan-mPLUG-Owl3` still outperform their non-visual counterparts by a notable margin. Nonetheless, our findings reinforce the idea that structured planning improves summarization quality beyond the video domain. In Appendix H, we further explore the effect of video content on our summarization task, varying the length of the video given as input to the model. We also perform experiments with different textual contexts for generating plan questions in Appendix I, and with controlled generation in Appendix J. Additionally, we present an error analysis of model output in Appendix K.

7 Human Evaluation

We conduct a human evaluation on 50 randomly selected instances from the VISTA test set. Annotators include master’s and doctoral students in computer science or computational linguistics with advanced English proficiency. They receive compensation per our university’s standard rate and are blind to the source of each summary to ensure impartial assessment. We compare `Plan-mPlug-Owl3`, `mPLUG-Owl3`, `LLAVA-NeXT-Interleave`, and `GPT-o1` against human reference summaries. Three independent annotators are asked to review the source video and evaluate corresponding model outputs (and the human upper bound) on a 1–5 Likert scale for Faithfulness, Relevance, Informativeness, Conciseness, and Coherence (higher scores indicate better quality). They are also asked to provide an overall ranking. In total, participants rated 750 samples ($50 \times 5 \times 3$). Appendix N contains the full evaluation instructions.

Figure 6 presents the performance of each

Model	Setting	R2	RLsum	VideoScore	FactVC
LLaMA-3.1$_{transcript}$	Zero-shot Learning	4.22 → **4.56**	21.39 → **22.01**	1.53 → **1.75**	34.32 → **40.78**
	QLoRA Fine-tuning	11.38 → **11.62**	30.39 → **30.55**	2.81 → **3.02**	53.22 → **60.47**
	Full Fine-tuning	11.93 → **12.24**	30.86 → **31.38**	3.21 → **3.25**	63.38 → **65.21**
LLaMA-3.1$_{OCR}$	Zero-shot Learning	4.37 → **4.59**	21.42 → **21.89**	1.50 → **1.72**	34.06 → **40.24**
	QLoRA Fine-tuning	12.11 → **12.33**	30.52 → **30.78**	2.77 → **2.98**	53.19 → **60.38**
	Full Fine-tuning	12.42 → **12.75**	31.72 → **32.19**	3.32 → **3.38**	65.84 → **67.53**
Qwen2-Audio	Zero-shot Learning	4.29 → **4.51**	21.53 → **22.18**	1.59 → **1.77**	34.31 → **40.52**
	QLoRA Fine-tuning	12.05 → **12.19**	30.77 → **31.04**	2.80 → **3.01**	54.27 → **61.44**
	Full Fine-tuning	12.37 → **12.68**	31.63 → **32.12**	3.22 → **3.25**	66.62 → **68.25**

Table 6: Performance of baseline vs. planning models in non-video settings across different learning regimes. Each cell shows the result *before* → *after* applying the planning method.

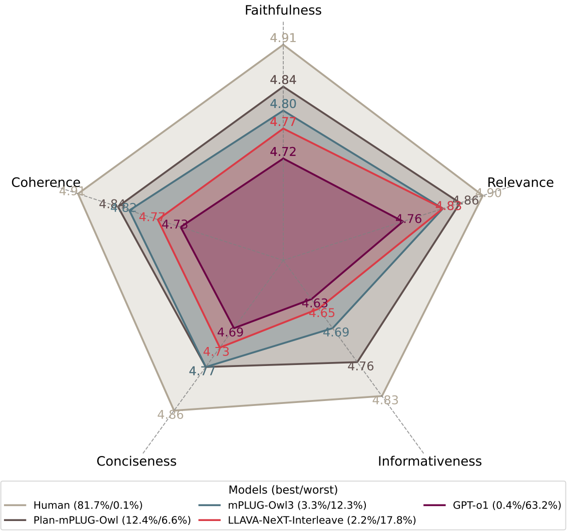


Figure 6: Human evaluation results. Human-written summaries consistently outperform all neural models.

model, along with the proportion of instances where models are rated best or worst. Fleiss' Kappa scores for Faithfulness ($\kappa = 0.767$), Relevance ($\kappa = 0.842$), Informativeness ($\kappa = 0.721$), Conciseness ($\kappa = 0.792$), and Coherence ($\kappa = 0.813$) indicate a substantial level of agreement, with an average agreement score of $\kappa = 0.787$. Overall, human-written summaries outperform all neural summarization models in quality, as they are perceived as substantially more faithful, coherent, concise, and informative. Human-written summaries are 81.7% more likely to be rated as best compared to model-generated summaries.

Among the four neural models, `GPT-o1` performs worst, being rated as worst 63.2% of the time. `LLAVA-NeXT-Interleave` follows suit, with a 17.8% chance of receiving the worst ranking. The plan-based model, `Plan-mPLUG-Owl3`, outperforms `mPLUG-Owl3` and demonstrates superior performance across all metrics. Additionally, it stands out among neural summarization systems for its higher likelihood of generating high-quality summaries. Paired t-tests show that human answers are considered significantly better than all neural models in all metrics ($p < 0.05$), revealing a clear gap between automatic systems and human performance on the VISTA dataset. The plan-based method is significantly better ($p < 0.05$) than other neural models in faithfulness, coherence, and informativeness, although it falls short of human performance. We also evaluate all samples of the test set with an LMM-as-Judge and obtain results that are broadly consistent with human evaluation. We describe the details of this study in Appendix L.

8 Conclusion

This paper introduces VISTA, a novel dataset specifically curated for the task of summarizing scientific video presentations into concise and coherent textual summaries. Comprehensive evaluations across multiple large (language/audio/multimodal) models demonstrate that this task poses significant challenges due to the complexity and multimodal nature of scientific presentations. To address these challenges, we operate a plan-based summarization approach that incorporates discourse-aware planning prior to summary generation. This method consistently improves summary quality, factual coverage, and coherence across multiple settings. In addition to presenting the dataset, our study reveals that even the strongest current models still fall short of matching human performance by a noticeable margin. We believe that VISTA could provide a robust and extensible foundation for future research on video-to-text summarization.

Ethical Considerations

All data in our dataset are sourced from publicly accessible resources, strictly adhering to relevant copyright regulations. Each data sample explicitly includes the corresponding source URL and author attribution. Throughout the processes of data processing, experimental analysis, model training, and evaluation, no instances of privacy infringement were identified. In human evaluations, all participants volunteered willingly and were fairly compensated. We provided a safe and comfortable environment for our participants and complied with ACL's Policy on Publication Ethics throughout our studies.

Limitations

Data All the summary and video data used in this study are open source. While our sources are generally of high quality and exhibit a broad range of diversity, we have not investigated inherent biases in the data. Moreover, as these data represent only a small fraction of real-world data, our findings may not extend to all video-to-text summarization scenarios.

Task In our task, we consider the paper abstract as a proxy for the summary of the corresponding video. This hypothesis has been supported by our two-stage quality control process, which ensures a strong alignment. However, we acknowledge that there may be nuanced differences between the abstract and a textual summary derived solely from the video. That said, authors often present the abstract as a summary of the video, as it conveys the key contributions, objectives, and findings of the research, which are typically central to the content discussed.

Model We have tested the plan-based approach on the video-based, audio-based, and text-based large models in our experiments. Our work does not aim to prove that the plan-based method is effective in all models of different modalities. Moreover, plan-based methods can take many different forms, and our work does not aim to identify the optimal planning approach for our dataset.

Scope Our study focuses on video-to-text summarization within scientific domains. We have not investigated applying the plan-based method to other natural language processing (NLP) tasks, such as multimodal machine translation, multimodal question answering, or multimodal reasoning. Although the plan-based approach could likely be adapted to these tasks with minimal effort, such possibilities remain unexplored and warrant future investigation.

Automated Evaluation While we employ a suite of automated metrics and hallucination detection methods to assess model performance on the test set, these metrics have inherent limitations and may fail to capture all aspects of model quality.

Human Evaluation Similar to many earlier studies (Papalampidi and Lapata, 2023; Krubiński and Pecina, 2023, 2024; Patil et al., 2024), we only evaluate 50 video-summary pairs, a subset that may not represent the entire dataset. Additionally, while all evaluators are graduate students, they are not necessarily experts in video-to-text summarization and possess varying levels of reading and assessment skills. Consequently, although their evaluations are valuable, they should not be treated as the only indicator of performance.

LMM-as-Judge Although the LMM-based judge paradigm enables large-scale and relatively consistent evaluations, it may inherit biases from its pretraining data, and its black-box nature makes the rating process difficult to interpret. Data contamination also remains a concern if `GPT-o1` is trained on overlapping data. We validate `GPT-o1`'s ratings with human evaluations on a small subset of samples, but this may not fully capture the model's reliability across diverse topics, domains, or summary styles. Therefore, results should be interpreted with caution and supplemented by human judgment where possible.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). Lapata acknowledges the support of the UK Engineering and Physical Sciences Research Council (Grant EP/W002876/1).