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

Dongqi Liu[Ω*], **Chenxi Whitehouse**[△], **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

[△]University of Cambridge, [Θ]University of Edinburgh

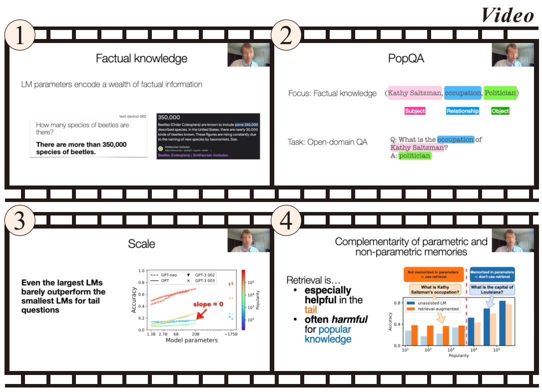
[Ω]{dongqi,xiyu,vera}@lst.uni-saarland.de

[△]chenxi.whitehouse@cl.cam.ac.uk

[Θ]{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.



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.

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).

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

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^3AV (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

[1]We discuss copyright in Appendix A.

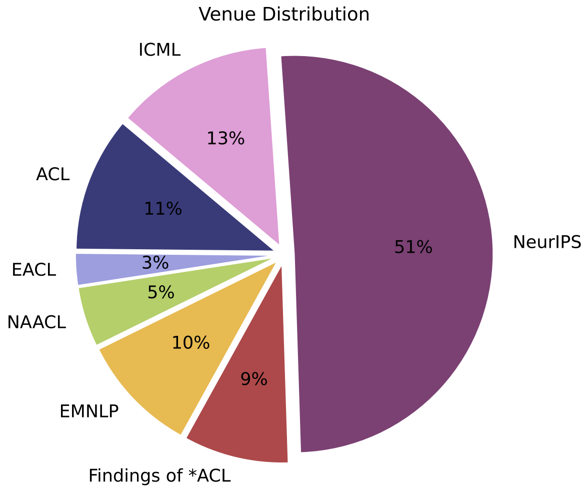


Figure 2: Venue distribution of the VISTA dataset.

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).

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

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.

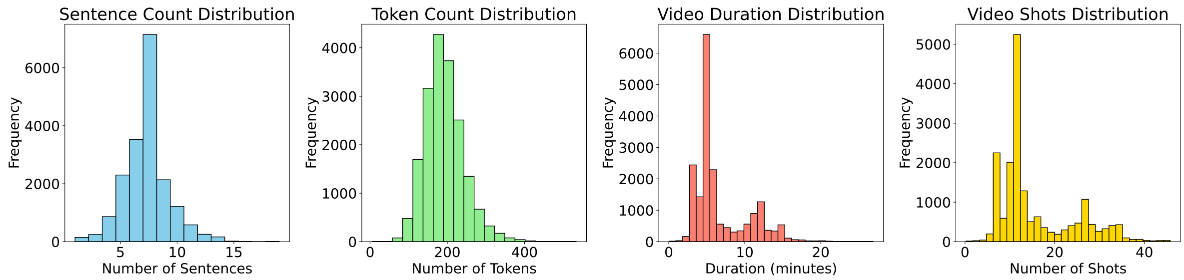


Figure 3: Distribution of summary sentences, summary tokens, video durations, and video shots in VISTA.

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

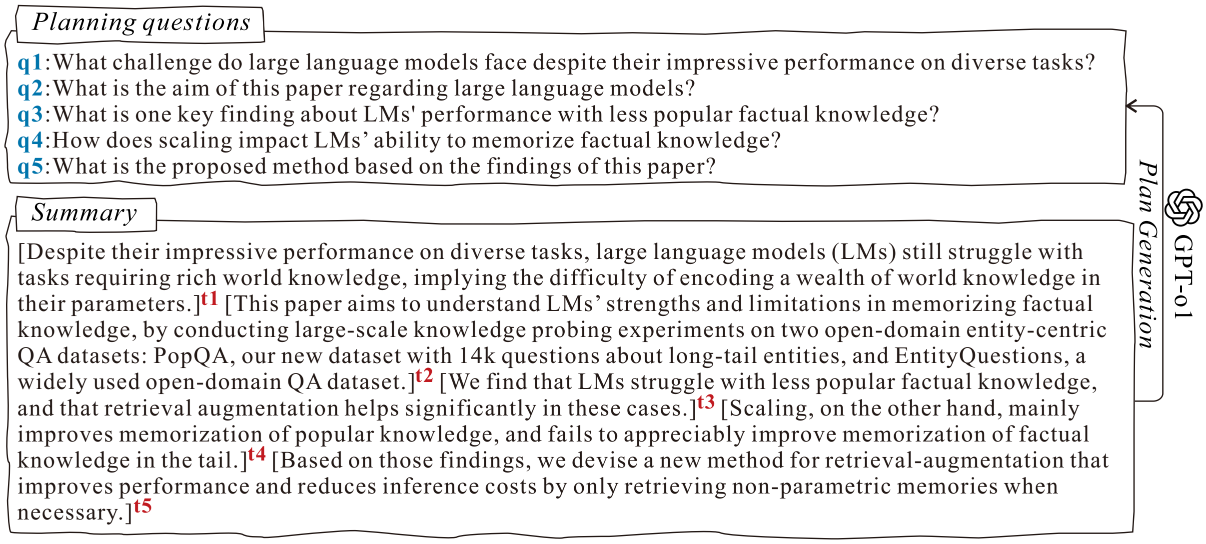


Figure 4: `GPT-o1` generates plans based on reference summaries. Each question q_i corresponds to a summary sentence t_i, which we assume constitutes its answer. Index i ranges from 1 to the number of summary sentences.

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).

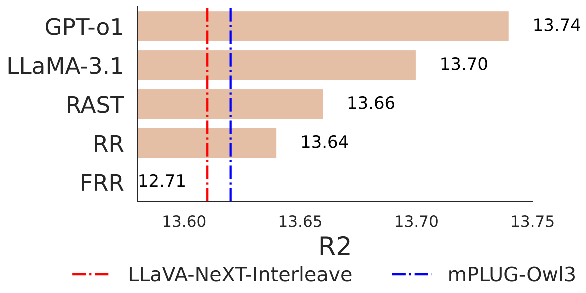


Figure 5: Noise in plan generation impacts summarization performance. FRR is a shorthand for Full Random Replacement, and RR for Random Replacement. RAST is a SOTA question generation method.

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.

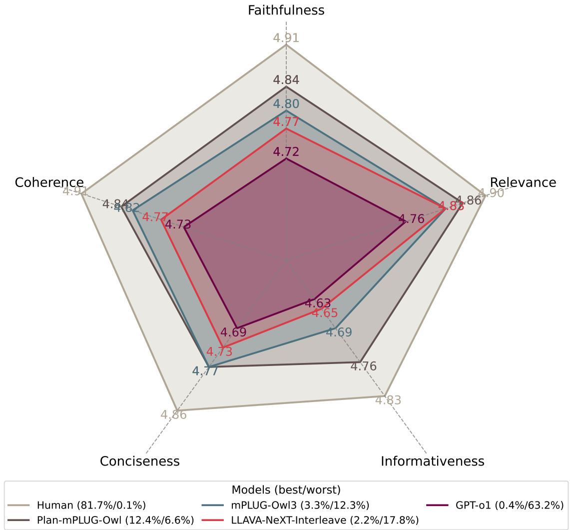


Figure 6: Human evaluation results. Human-written summaries consistently outperform all neural models.

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).

European Research Council
Established by the European Commission

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. 2024. Tuning large multimodal models for videos using reinforcement learning from AI feedback. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–940, Bangkok, Thailand. Association for Computational Linguistics.

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12489–12497.

Anthropic. 2024. Claude 3.5 - sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2024-12-06.

Dawit Mureja Argaw, Seunghyun Yoon, Fabian Caba Heilbron, Hanieh Deilamsalehy, Trung Bui, Zhaowen Wang, Franck Dernoncourt, and Joon Son Chung. 2024. Scaling up video summarization pretraining with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8332–8341.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Gerard Canal, Senka Krivić, Paul Luff, and Andrew Coles. 2022. Planverb: Domain-independent verbalization and summary of task plans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9698–9706.

Brian Chen, Xiangyuan Zhao, and Yingnan Zhu. 2024a. Personalized video summarization by multimodal video understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4382–4389.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024b. Mllm-as-a-judge: assessing multimodal llm-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024c. M^{3}AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Bangkok, Thailand. Association for Computational Linguistics.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. 2023. Can we edit multimodal large language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13877–13888, Singapore. Association for Computational Linguistics.

Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. 2021. StreamHover: Livestream transcript summarization and annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Aldan Creo, Manuel Lama, and Juan C Vidal. 2023. Prompting llms with content plans to enhance the summarization of scientific articles. *arXiv preprint arXiv:2312.08282*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Mehwish Fatima and Michael Strube. 2023. Cross-lingual science journalism: Select, simplify and rewrite summaries for non-expert readers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1843–1861, Toronto, Canada. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2025. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. MM-AVS: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926, Online. Association for Computational Linguistics.

Qi Gou, Zehua Xia, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li, and Nguyen Cam-Tu. 2023. Diversify question generation with retrieval-augmented style transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1677–1690, Singapore. Association for Computational Linguistics.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. 2025. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *The Thirteenth International Conference on Learning Representations*.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. 2024. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, Miami, Florida, USA. Association for Computational Linguistics.

Anwen Hu, Yaya Shi, Haiyang Xu, Jiabo Ye, Qinghao Ye, Ming Yan, Chenliang Li, Qi Qian, Ji Zhang, and Fei Huang. 2024. mPLUG-paperowl: Scientific diagram analysis with the multimodal large language model. In *ACM Multimedia 2024*.

Hang Hua, Yunlong Tang, Chenliang Xu, and Jiebo Luo. 2024. V2xum-llm: Cross-modal video summarization with temporal prompt instruction tuning. *arXiv preprint arXiv:2404.12353*.

Kung-Hsiang Huang, Hou Pong Chan, Yi R Fung, Haoyi Qiu, Mingyang Zhou, Shafiq Joty, Shih-Fu Chang, and Heng Ji. 2024. From pixels to insights: A survey on automatic chart understanding in the era of large foundation models. *IEEE Transactions on Knowledge and Data Engineering*.

Fantine Huot, Joshua Maynez, Chris Alberti, Reinald Kim Amplayo, Priyanka Agrawal, Constanza Fierro, Shashi Narayan, and Mirella Lapata. 2024. μPLAN: Summarizing using a content plan as cross-lingual bridge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2146–2163, St. Julian's, Malta. Association for Computational Linguistics.

Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. 2024. Video recap: Recursive captioning of hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18198–18208.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4091–4098, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs. In *International Conference on Machine Learning*, pages 17283–17300. PMLR.

Mateusz Krubiński and Pavel Pecina. 2023. MLASK: Multimodal summarization of video-based news articles. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 910–924, Dubrovnik, Croatia. Association for Computational Linguistics.

Mateusz Krubiński and Pavel Pecina. 2024. Towards unified uni- and multi-modal news headline generation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 437–450, St. Julian's, Malta. Association for Computational Linguistics.

Jonas Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. 2020. Learning kernel tests without data splitting. *Advances in Neural Information Processing Systems*, 33:6245–6255.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. Talk-Summ: A dataset and scalable annotation method for scientific paper summarization based on conference talks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2125–2131, Florence, Italy. Association for Computational Linguistics.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. 2025. LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand. Association for Computational Linguistics.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. VMSMO: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024c. Llama-vid: An image is worth 2 tokens in large language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVI*, page 323–340, Berlin, Heidelberg. Springer-Verlag.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2024b. Videoxum: Cross-modal visual and textural summarization of videos. *IEEE Transactions on Multimedia*, 26:5548–5560.

Danyang Liu, Mirella Lapata, and Frank Keller. 2023a. Visual storytelling with question-answer plans. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5800–5813, Singapore. Association for Computational Linguistics.

Dongqi Liu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Dongqi Liu and Vera Demberg. 2024. RST-LoRA: A discourse-aware low-rank adaptation for long document abstractive summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2200–2220, Mexico City, Mexico. Association for Computational Linguistics.

Dongqi Liu, Xudong Hong, Pin-Jie Lin, Ernie Chang, and Vera Demberg. 2022. Two-stage movie script summarization: An efficient method for low-resource long document summarization. In *Proceedings of the Workshop on Automatic Summarization for Creative Writing*, pages 57–66, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Dongqi Liu, Yifan Wang, and Vera Demberg. 2023b. Incorporating distributions of discourse structure for long document abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5574–5590, Toronto, Canada. Association for Computational Linguistics.

Dongqi Liu, Yifan Wang, Jia Loy, and Vera Demberg. 2024a. SciNews: From scholarly complexities to public narratives – a dataset for scientific news report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14429–14444, Torino, Italia. ELRA and ICCL.

Dongqi Liu, Xi Yu, Vera Demberg, and Mirella Lapata. 2025. Explanatory summarization with discourse-driven planning. *Preprint*, arXiv:2504.19339.

Hui Liu and Xiaojun Wan. 2021. Video paragraph captioning as a text summarization task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 55–60, Online. Association for Computational Linguistics.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11807–11823, Singapore. Association for Computational Linguistics.

Ran Liu, Ming Liu, Min Yu, He Zhang, Jianguo Jiang, Gang Li, and Weiqing Huang. 2024b. SumSurvey: An abstractive dataset of scientific survey papers for long document summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9632–9651, Bangkok, Thailand. Association for Computational Linguistics.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024c. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

Louis Mahon and Mirella Lapata. 2024a. A modular approach for multimodal summarization of TV shows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8272–8291, Bangkok, Thailand. Association for Computational Linguistics.

Louis Mahon and Mirella Lapata. 2024b. Screenwriter: Automatic screenplay generation and movie summarisation. *arXiv preprint arXiv:2410.19809*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Yuning Mao, Ming Zhong, and Jiawei Han. 2022. CiteSum: Citation text-guided scientific extreme summarization and domain adaptation with limited supervision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10922–10935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Pinelopi Papalampidi and Mirella Lapata. 2023. Hierarchical3D adapters for long video-to-text summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1297–1320, Dubrovnik, Croatia. Association for Computational Linguistics.

Vaidehi Patil, Leonardo Ribeiro, Mengwen Liu, Mohit Bansal, and Markus Dreyer. 2024. REFINESUMM: Self-refining MLLM for generating a multimodal summarization dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13773–13786, Bangkok, Thailand. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. SPIQA: A dataset for multimodal question answering on scientific papers. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, et al. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921.

Aishwarya Ramakrishnan and Chun-Kit Ngan. 2022. A hybrid video-to-text summarization framework and algorithm on cascading advanced extractive- and abstractive-based approaches for supporting viewers' video navigation and understanding. In *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, pages 36–39.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *NeurIPS*.

Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal video summarization via time-aware transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1756–1765, New York, NY, USA. Association for Computing Machinery.

Aseem Srivastava, Smriti Joshi, Tanmoy Chakraborty, and Md Shad Akhtar. 2024. Knowledge planning in large language models for domain-aligned counseling summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17775–17789, Miami, Florida, USA. Association for Computational Linguistics.

Ashima Suvarna, Xiao Liu, Tanmay Parekh, Kai-Wei Chang, and Nanyun Peng. 2024. QUDSELECT: Selective decoding for questions under discussion parsing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1288–1299, Miami, Florida, USA. Association for Computational Linguistics.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ye Wang, Xiaojun Wan, and Zhiping Cai. 2022. Guiding abstractive dialogue summarization with content planning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3408–3413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. 2025. A comprehensive study of multimodal large language models for image quality assessment. In *European Conference on Computer Vision*, pages 143–160. Springer.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023a. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023b. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537, Singapore. Association for Computational Linguistics.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2025. mPLUG-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Litian Zhang, Xiaoming Zhang, Linfeng Han, Zelong Yu, Yun Liu, and Zhoujun Li. 2024a. Multi-task hierarchical heterogeneous fusion framework for multimodal summarization. *Information Processing & Management*, 61(4):103693.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Copyright

According to the statement displayed on the ACL Anthology website, *"Permission is granted to make copies for the purposes of teaching and research"*, allowing us to use the corresponding data. For ICML and NeurIPS, we (the authors) have obtained written confirmation granting permission to use the paper titles, author lists, paper abstracts, full papers, and presentation videos available on their websites for research purposes.

B Quality Control

Manual Control We randomly select 500 video-summary pairs to assess whether the summaries provide accurate descriptions of the videos. Two Ph.D. candidates in Computer Science or Computational Linguistics perform binary judgments on these pairs. Across all 500 samples, neither evaluator rejected any sample.

Automated Control To go beyond the limited scope of manual checks, we employ `GPT-o1` for automated assessment using the same binary criteria across all data samples. The model initially flagged 39 pairs as potentially invalid. These flags were likely caused by difficulties in interpreting domain-specific terms or rare expressions and sensitivity to variations in summary length. After further manual review, all 39 samples were confirmed as valid and retained in the dataset.

C Data Sample

The VISTA dataset contains carefully curated video-text pairs, predominantly sourced from published papers, aiming to ensure a high standard of quality and relevance. The accompanying texts are designed to function as summaries of their respective videos, offering a concise representation of their content (see Figure 7). Additionally, our dataset focuses on topics within the field of artificial intelligence, making it a good resource for research in AI-related video-to-text summarization and comprehension.

D Model Version Details

Table 7 provides the detailed version identifiers for the models evaluated in our study, showing both model names as referenced in the main text and the specific versions used in our experiments.

E Hyper-parameters Settings

For all fine-tuning experiments, we utilize the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$, and a weight decay of 0.1, combined with a warm-up ratio of 0.15. The initial learning rate is set to 5e-5, with cosine learning rate scheduling. DeepSpeed is configured with ZeRO-3 Offload. We set the random seed to 2025 and apply a dropout rate of 0.1. In the QLoRA setting, the rank r is set to 32, the scaling factor α is set to 64, and the dropout rate for the low-rank matrices is 0.1. All other parameters follow the default settings of the `Transformers` library.

During training, we save the checkpoint with the highest Rouge-2 F1 score on the validation set as the final model. All experiments are conducted over 16 epochs with a batch size of 16 and early stopping (all models converged before 16 epochs). For model inference (including zero-shot learning), we employ a beam search with a beam of size 4, a

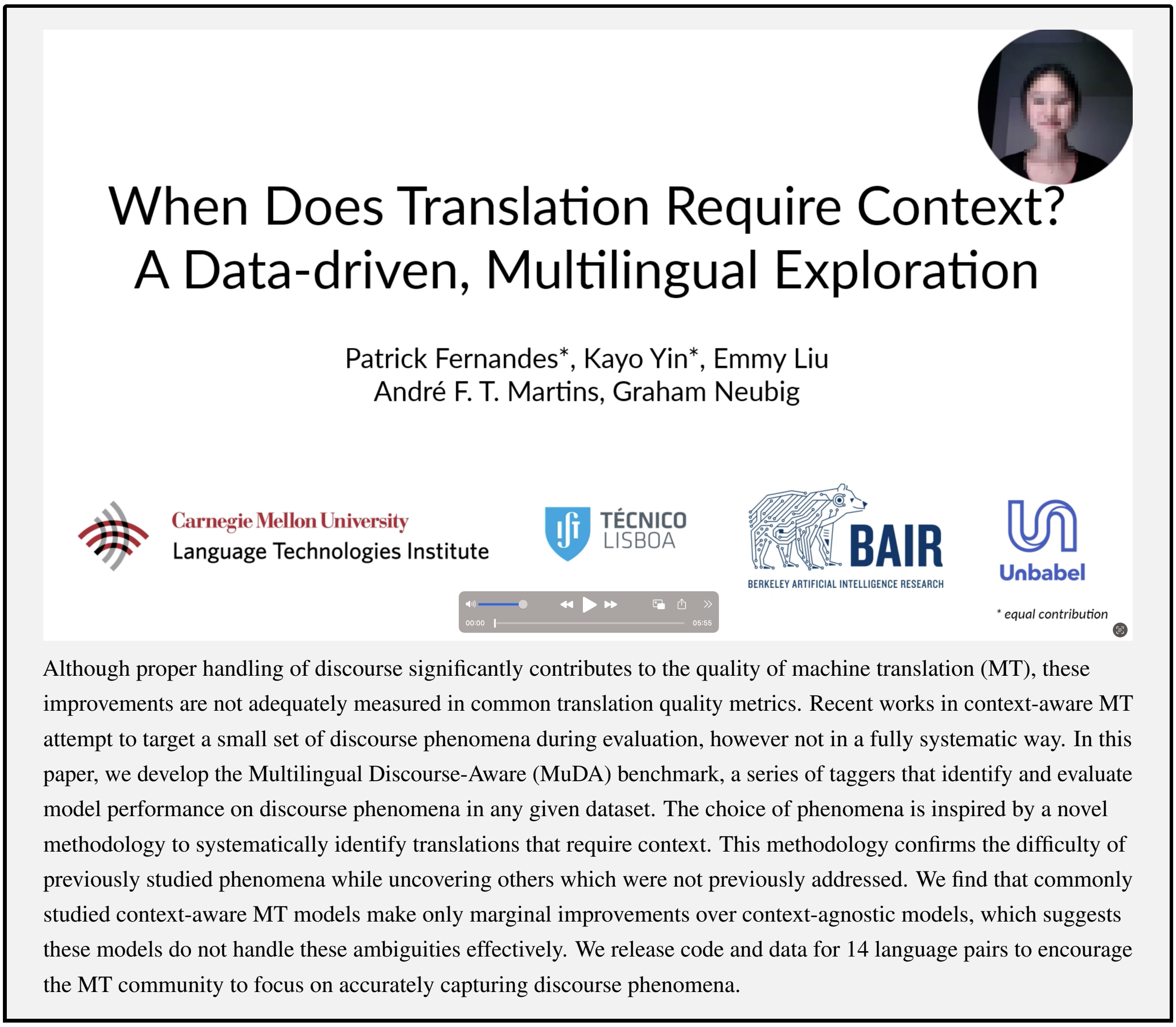


Figure 7: A random sample from the VISTA dataset, originating from Fernandes et al. (2023).

length penalty of 3.0, a no-repeat n-gram size of 3, and the maximum number of new tokens generated is limited to 256. For video-based LMMs, the sampling rate is set to 0.1 fps, and the number of extracted frames is set to 32.

For closed-source models, results are obtained via API requests during the experimental period from 01/09/2024 to 10/02/2025. The hyperparameter settings for these API requests include a temperature of 1, top_p of 1, a frequency penalty of 0.2, and a presence penalty of 0.2. All other parameters adhere to the default settings specified by their respective platforms.

F Automatic Evaluation Metrics

In line with common practice in video-to-text summarization research, we evaluate the model-generated summaries using the following metrics:

- ROUGE (Lin, 2004): measures n-gram overlap between machine-generated and human reference texts. We report F1 scores for Rouge-1 (R1), Rouge-2 (R2), and Rouge-Lsum (RLSUM).
- SacreBLEU (Post, 2018): assesses linguistic consistency and fluency between generated and reference texts.
- METEOR (Banerjee and Lavie, 2005): calculates the harmonic mean of unigram precision and recall, placing greater emphasis on recall for a balanced evaluation.
- BERTScore (Zhang et al., 2020): uses contextual embeddings from BERT to evaluate semantic similarity between texts.
- CIDEr-D (Vedantam et al., 2015): evaluates the consensus between generated summaries and references by using TF-IDF weighting combined with a decay factor to reduce the impact of repeated terms.

Model	Version	Model Size
GPT-o1 (Achiam et al., 2023)	o1-2024-12-17	Unknown
Gemini 2.0 (Team et al., 2023)	Gemini 2.0 Flash	Unknown
Claude 3.5 Sonnet (Anthropic, 2024)	claude-3-5-sonnet-20241022	Unknown
LLaMA-3.1 (Touvron et al., 2023)	LLaMA-3.1-8B-Instruct	8B
Qwen2-Audio (Chu et al., 2024)	Qwen2-Audio-7B-Instruct	7B
Video-LLaMA (Zhang et al., 2023)	VideoLLaMA2-7B-16F	7B
Video-ChatGPT (Maaz et al., 2024)	Video-ChatGPT-7B	7B
Video-LLaVA (Lin et al., 2024a)	Video-LLaVA-7B-hf	7B
LLaMA-VID (Li et al., 2024c)	LLaMA-VID-7B-Full-224-Long-Video	7B
LLaVA-NeXT-Interleave (Li et al., 2025)	LLaVA-NeXT-Interleave-Qwen-7B	7B
mPLUG-Owl3 (Ye et al., 2025)	mPLUG-Owl3-7B-241101	7B

Table 7: Model version details.

- VideoScore (He et al., 2024): focuses on text-to-video alignment, evaluating how accurately video content matches the given text prompts using fine-grained multi-aspect scoring.
- FactVC (Liu and Wan, 2023): calculates the factual consistency of text with video content by aligning coarse-grained video-text similarity and precision-based fine-grained matching. The original values of FactVC range from 0 to 1, and in our experiments, we scale them by 100 to convert them into percentages.

G Plans Quality Validation

To validate the quality of the silver-standard plans generated by GPT-o1, we conduct a manual evaluation on 100 randomly selected samples. The evaluation is carried out by the same annotators involved in our human evaluation setup. Each annotator is asked to make a binary judgment on whether the generated plan question satisfied two validity criteria: (1) Local Coherence: The question is well-formed and semantically related to the summary; and (2) QUD-Alignment: Each sentence in the summary could plausibly serve as an answer to the question, consistent with the QUD framework.

We observe strong inter-annotator agreement (Fleiss' $\kappa = 0.853$), indicating a high degree of consistency in decisions. In addition to this, we perform a manual error analysis to screen for systematic biases or recurrent flaws, such as overly generic phrasing, hallucinated entities, or structural redundancy. No such patterns are observed.

H Impact of Video Context on Summary Generation

We examine the impact of different video context configurations on summary generation, comparing mPLUG-Owl3 with Plan-mPlug-Owl3. Unlike earlier experiments that use the full video as input, here only the first or last 10% or 30% of the video is provided as input. We report results in the full fine-tuning setting.

Context	Model	R2	RLsum	VideoScore	FactVC
All	mPLUG-Owl3	13.62	32.91	3.28	71.94
	Plan-mPlug-Owl3	13.74	33.25	3.33	75.41
First 10%	mPLUG-Owl3	6.31	25.44	2.37	51.02
	Plan-mPlug-Owl3	7.37	27.38	2.52	52.39
First 30%	mPLUG-Owl3	9.42	28.88	2.78	54.10
	Plan-mPlug-Owl3	10.59	30.13	2.78	55.37
Last 10%	mPLUG-Owl3	6.53	27.34	2.51	53.64
	Plan-mPlug-Owl3	7.62	29.73	2.77	55.93
Last 30%	mPLUG-Owl3	7.32	29.17	2.82	57.36
	Plan-mPlug-Owl3	10.72	31.29	2.98	62.05

Table 8: Model performance under different video context configurations (full fine-tuning). The video content at the end is more helpful for summary generation.

The results in Table 8 indicate that partial video context consistently underperforms compared to using the full video. Using the last part of the video generally produces better results than using the first part, as concluding sections often summarize key findings while opening sections primarily introduce background information. Additionally, utilizing 30% of the video outperforms using only 10%, highlighting that more content generally yields better outputs. Across all configurations, the Plan-mPlug-Owl3 model consistently outperforms mPLUG-Owl3.

I Impact of Text Context on Plan Generation

The generation of plan questions in our experiments is influenced by the target sentence and its context. In our main experiments, plan questions are generated based on the target sentence and its preceding summary text (`Previous-Context`), in line with the original Question Under Discussion (QUD) requirements (Wu et al., 2023a,b; Liu et al., 2025). We now assess configurations that generate questions only based on the target sentence (`No-Context`) or the entire summary (`All-Context`).

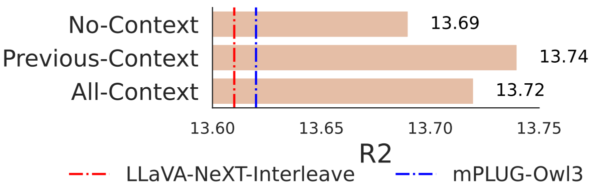


Figure 8: Impact of text context for plan generation.

As shown in Figure 8, performance differences between different context configurations are relatively small (yet superior to models without planning components shown as red and blue dashed lines). `No-Context` shows the lowest performance but is the most cost-effective, as it requires the shortest input length for `GPT-o1` during question generation. `All-Context` achieves slightly better results but at the highest computational cost due to the long input length. `Previous-Context` is aligned with QUD and strikes a good balance, achieving the best performance for a moderate cost.

J Controllable Generation

An advantage of plan-based models is their ability to control the output summaries by modifying the plans used for generation. We investigate how modifying the structure and composition of these plans impacts the generated summaries, specifically comparing their performance against direct summary generation control through instructions. To this end, we design two controlled experiments:

- *Summary Readability*: How question complexity affects readability, tailored for lay readers or expert readers.
- *Summary Length*: How the number of questions influences summary length, by removing 10%, 30%, and 60% of questions.

Condition	`Plan-mPlug-Owl3`		`GPT-o1`	
	R2	FRE	R2	FRE
No change	13.74	30.62	5.69	26.37
Lay questions	13.38	35.17	4.26	28.94
Expert questions	13.24	23.54	4.13	24.33

Table 9: Control experiment for summary readability. FRE = Flesch Reading Ease.

Condition	`Plan-mPlug-Owl3`		`GPT-o1`	
	R2	Avg. #Tokens	R2	Avg. #Tokens
No deletion	13.74	202.39	5.69	267.32
Delete 10%	11.05	178.47	4.32	220.49
Delete 30%	10.41	137.72	3.17	192.42
Delete 60%	8.01	100.32	2.98	185.28

Table 10: Control experiment for summary length.

We note that the plan-based method employs an explicit planning component where each sentence is guided by a corresponding question that facilitates fine-grained control over the summary's style or content. Specifically, after `PG` produces the plan, we use `GPT-o1` to edit it and then feed the edited questions back to `SG` for the final output. For `GPT-o1`, which operates in a zero-shot manner, we prepend constraints directly in the prompt. Specifically, `GPT-o1` generates an initial summary in one pass and then applies additional prompt-based instructions during a secondary rewriting step to control the output. Both control experiments (Table 9 and Table 10) reveal similar trends: While performance declines for both models, the plan-based method is more robust and controllable.

In the readability control experiment (Table 9), both models show reductions in R2, but `Plan-mPlug-Owl3` declines less, averaging an R2 loss of 0.43 compared to 1.50 for `GPT-o1`. Furthermore, `Plan-mPlug-Owl3` controls readability more effectively, achieving a higher Flesch Reading Ease (FRE) score[3] of 35.17 for lay questions, compared to 28.94 for `GPT-o1`, and a lower FRE score of 23.54 for expert questions.

In the length control experiment (Table 10), R2 scores decline as content is removed, but the plan-based model aligns more closely with target compression ratios, producing summaries averaging 100.32 tokens under 60% deletion, while `GPT-o1` generates longer summaries (185.28 tokens).

[3]The FRE score, which ranges from 0 to 100, measures text readability, with higher scores indicating easier-to-read content, and lower scores reflecting greater complexity.

K Case Study and Error Analysis

For our case study, we randomly select a sample (Kübler et al., 2020) from the test split. The analysis in Table 11 reveals differences in summary quality across models and against the human-written text. Specifically, GPT-o1 often produces concise summaries but at the cost of precision. For example, it incorrectly claims that "data splitting helps control test thresholds," which is a hallucination — while data splitting ensures a tractable null distribution, it does not explicitly control test thresholds. Furthermore, its summaries frequently oversimplify complex concepts, reducing the depth of explanations and omitting crucial distinctions, such as the role of dependency calibration in the proposed method. Similarly, mPLUG-Owl3 introduces factual inaccuracies, such as stating that data splitting "ensures a reliable null distribution." This phrasing misleadingly implies that reliability is an inherent property of data splitting, whereas the correct point is that it makes the null distribution tractable rather than necessarily more reliable.

Plan-mPlug-Owl3 is more factually accurate than the other models. It correctly captures the main idea of full-sample hyperparameter learning and testing without data splitting. However, it still introduces subtle distortions, such as falsely suggesting a "trade-off" between test power and tractability, which misrepresents the actual relationship. These inaccuracies, while less severe than those in GPT-o1 and mPLUG-Owl3, highlight the model's tendency to infer unstated causal links, leading to potential misinterpretations. Despite the relative strengths of Plan-mPlug-Owl3, all generated summaries fall short of human-written text. The model-generated outputs consistently struggle with informativeness, coherence, and factual accuracy. These shortcomings underscore the ongoing challenge of improving automated summarization systems to better align with human standards in both accuracy and clarity.

Controlled generation experiments reveal that hallucination issues are further amplified when imposing constraints on readability and length. Under readability control (Table 12), GPT-o1 is more likely to introduce fabricated or misleading content when forced to generate more complex outputs. This occurs because it lacks an explicit mechanism to ensure factual consistency while adapting to varying readability demands. Rather than relying on implicit internal heuristics, Plan-mPlug-Owl3 has an explicit planning mechanism which makes it less likely to introduce unsupported claims. Planning provides an additional layer of control, helping the model maintain factual alignment even as readability demands change. A similar trend is observed in length control experiments (Table 13). As the compression ratio increases, GPT-o1 struggles to balance conciseness and informativeness, sometimes hallucinating missing details to compensate for omitted content. This suggests that purely instruction-based control (i.e., prompting the model to shorten outputs) does not effectively enforce content retention, leading to greater inconsistencies. In contrast, the plan allows Plan-mPlug-Owl3 to selectively retain essential elements, reducing the risk of generating misleading content; it can also avoid answering deleted questions, to a certain extent.

These findings reinforce the advantages of plan-based control over instruction-based prompting. While neither approach fully eliminates hallucinations, planning provides a structured mechanism to manage content selection, ensuring greater alignment with the input source compared to freeform generative adjustments.

L LMM-as-Judge Evaluation

To facilitate large-scale comparisons of model outputs, we adopt a method inspired by LLM-as-Judge (Liusie et al., 2024; Liu et al., 2024c; Zheng et al., 2024; Liu et al., 2025), extending it to use a large multimodal model (Chen et al., 2024b). The proposed LMM-based evaluator incorporates both textual and video modalities and assesses the same summary quality dimensions used in our human evaluation for trend-level comparisons. Specifically, we use GPT-o1 as the evaluator, following the hyperparameter settings in Appendix E. To minimize potential bias from prior queries, the conversation history is reset before each evaluation. The instructions for the LMM-as-Judge evaluation are provided in Appendix O.

We validate the agreement between GPT-o1 and human ratings by comparing its ratings with human evaluations on the same 50 samples from the VISTA test set. We calculate Fleiss' Kappa between GPT-o1 and mean human ratings across the dimensions of Faithfulness (κ=0.732), Relevance (κ=0.803), Informativeness (κ=0.730), Conciseness (κ=0.792) and Coherence (κ=0.721) at instance level. These results indicate that human

Model	Output
Reference	Modern large-scale kernel-based tests such as maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) optimize kernel hyperparameters on a held-out sample via data splitting to obtain the most powerful test statistics. While data splitting results in a tractable null distribution, it suffers from a reduction in test power due to a smaller test sample size. Inspired by the selective inference framework, we propose an approach that enables learning the hyperparameters and testing on the full sample without data splitting. Our approach can correctly calibrate the test in the presence of such dependency and yield a test threshold in closed form. At the same significance level, our approach's test power is empirically larger than that of the data-splitting approach, regardless of its split proportion.
GPT-o1	Modern kernel-based tests like MMD and KSD use data splitting to tune hyperparameters. This splitting helps control test thresholds but reduces power because it splits the data into smaller parts. We propose a new method to tune hyperparameters without splitting the data. Our method provides better test power than data splitting.
mPLUG-Owl3	Large-scale kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) usually split the data to tune hyperparameters. While this approach ensures a reliable null distribution, it reduces test power due to smaller sample sizes. We propose a method that avoids data splitting, enabling full-sample hyperparameter learning and testing. This method improves test power at the same significance level compared to data splitting, with a closed-form threshold for the test.
Plan Questions	q1: How do kernel-based tests like MMD and KSD typically optimize hyperparameters? q2: What is the key idea behind the proposed approach? q3: What are the advantages of the proposed method compared to data-splitting methods?
Plan-mPlug-Owl3	Modern kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) typically split the data to optimize hyperparameters, trading off test power for a tractable null distribution. Inspired by the selective inference framework, we propose a novel approach that eliminates the need for data splitting, allowing full-sample hyperparameter learning and testing. Our method addresses dependencies introduced by learning and provides a closed-form test threshold while achieving higher test power than data-splitting methods, regardless of split proportions.

Table 11: Comparative case study of model outputs. Hallucinations are highlighted in red color.

Model	Output
GPT-o1	Modern kernel-based tests, such as MMD and KSD, use data splitting to adjust hyperparameters. This data splitting ensures proper test thresholds but weakens test power because the data is divided into smaller portions. We propose a new method that avoids splitting the data and provides better test power.
Plan Questions	q1: What do MMD and KSD do with data splitting? q2: What is the main purpose of the proposed method? q3: Why is the proposed method better than data splitting?
Plan-mPlug-Owl3	Kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) typically divide the data into subsets to adjust hyperparameters. This approach ensures a manageable null distribution but reduces test power. Inspired by selective inference, we propose a new method that avoids splitting the data. This method uses the entire dataset for hyperparameter tuning and testing, textcolorredeliminating dependencies introduced by learning. It also provides a clear, closed-form test threshold and achieves higher test power compared to data-splitting methods, regardless of the split proportions.

Table 12: Decreasing reading difficulty: plan modifications vs instruction adjustments. Hallucinations are in red.

Model	Output
GPT-o1	Kernel-based statistical methodologies, exemplified by maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD), conventionally employ data partitioning mechanisms to facilitate hyperparameter optimization. While this stratagem ensures a well-calibrated null distribution, it inherently compromises the statistical power of the test due to data fragmentation. We advance a paradigm wherein hyperparameter selection is conducted holistically, circumventing data splitting and consequently preserving statistical power.
Plan Questions	q1: Through what operational principles do MMD and KSD navigate the interplay between hyperparameter tuning and null distribution calibration? q2: How does the selective inference framework inform the methodological innovations of the proposed approach? q3: What are the quantifiable and theoretical advantages of the proposed strategy in circumventing the limitations imposed by data-splitting conventions?
Plan-mPlug-Owl3	Modern kernel-based tests like maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD) typically split the data to optimize hyperparameters, trading off test power for a tractable null distribution. Inspired by the selective inference framework, we propose a novel approach that eliminates the need for data splitting, allowing full-sample hyperparameter learning and testing. Our method addresses dependencies introduced by learning and provides a closed-form test threshold while achieving higher test power than data-splitting methods, regardless of split proportions.

Table 13: Increasing reading difficulty: plan modifications vs. instruction adjustments. Hallucinations are in red.

Model	Output
GPT-o1	Kernel-based tests like MMD and KSD use data splitting for hyperparameter tuning, which controls thresholds but reduces power. We propose a method that avoids splitting, improving test power.
Plan Questions	q1: How do kernel-based tests like MMD and KSD typically optimize hyperparameters? q2: What are the advantages of the proposed method compared to data-splitting methods?
Plan-mPlug-Owl3	Kernel-based tests, such as maximum mean discrepancy (MMD) and kernelized Stein discrepancy (KSD), rely on data splitting for hyperparameter tuning. Our method delivers analytically derived test thresholds and consistently outperforms data-splitting techniques in test power, regardless of how the data is divided.

Table 14: Deleting 30% of plan questions vs. instruction adjustments. Hallucinations are highlighted in red.

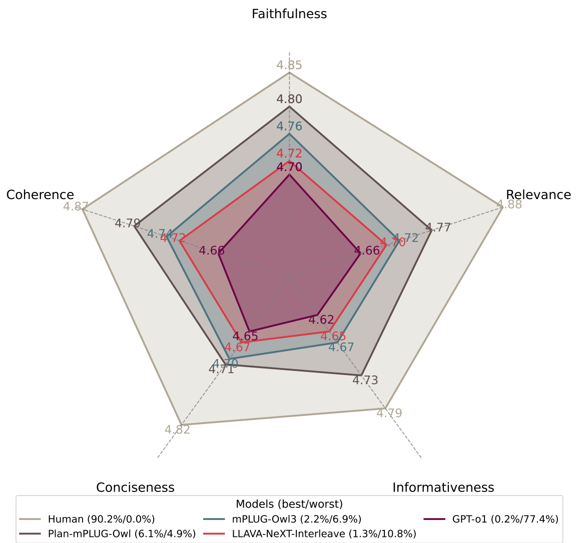


Figure 9: LMM-as-Judge evaluation results showing that human-written summaries consistently outperform neural models.

evaluators and `GPT-o1` achieve substantial levels of agreement across these dimensions. Following this, we expand the evaluation to include all samples in our test set.

Compared to fine-tuned models, `GPT-o1` assigns the lowest scores to its own responses (see Figure 9). Human-written summaries consistently receive the highest scores and are generally regarded as the best. Aligning with our human evaluations, `GPT-o1` also recognizes that the plan-based model outperformed other models. We further conduct paired t-tests to find that human summaries outperform all neural models across all metrics with statistical significance ($p < 0.05$). Moreover, the plan-based model demonstrates significantly better performance ($p < 0.05$) than other neural models across all metrics except for conciseness. Our results also indicate that although the plan-based method can improve the performance of end-to-end models to some extent, there is still a gap between machine-generated and human summaries, which also reflects the challenging nature of our dataset.

M Prompts Used in Our Study

Quality Control Guidelines

Evaluate each video-text pair to determine whether the text provides a concise and accurate summary of the corresponding video.

- **Concise:** Ensure the text is brief, focused, and free of unnecessary details.
- **Accurate:** Verify that the text faithfully represents the video's content.

Make binary judgments (`Valid` or `Invalid`) for each pair. If flagged as `Invalid`, provide a brief justification.
Answer:
Judgment: (`Valid` or `Invalid`)
Justification: (`Justification if flagged as invalid`)

Figure 10: Quality control guidelines.

Summary Generation (without plan)

Generate a summary for the provided content.
Content: {`Video/Audio/Transcript/OCR`}
Summary:

Figure 11: Prompt to generate summaries without plans.

Question Generation

Generate a coherent and contextually relevant question based on the provided context and target sentence, ensuring that the target sentence can be treated as an answer to the generated question.
Context: {`Context Text`}
Target: {`Target Sentence`}
Question Sentence:

Figure 12: Prompt for question generation.

Prompt for PG model

Generate a list of questions for the provided {`Video/Audio/Transcript...`}.
Content: {`Video/Audio/Transcript...`}
Questions:

Figure 13: Prompt for PG model.

Prompt for SG model

Generate a summary for the following {`Video/Audio/Transcript...`} based on the plan questions.
Content: {`Video/Audio/Transcript...`}.
Plan Questions: {`Questions`}
Ensure that the generated summary sequentially answers the plan questions.
Summary:

Figure 14: Prompt for SG model.

Irrelevant Question Generation

Randomly generate a question with a question mark.
Question Sentence:

Figure 15: Prompt used by GPT-o1 to generate irrelevant questions.

Summary Readability Modification

Rewrite the following text to further adjust the style or detail.
Here is the text to be rewritten: {Text}
Refine the above text to be more {lay/expert} style.
Modified Text:

Figure 16: Summary readability modification.

Summary Length Modification

Rewrite the following text to further adjust the style or detail.
Here is the text to be rewritten: {Text}
Shorten the above text by about {10% / 30% / 60%}. Focus on the key points and remove less critical details.
Modified Text:

Figure 17: Summary length modification.

Plan Readability Modification

Rewrite the following questions to further adjust the style or detail.
Here are the questions to be rewritten:
1. {Q1}
2. {Q2}
...
Refine the above questions to be more {lay/expert} style.
Modified Questions:

Figure 18: Plan readability modification.

N Human Evaluation Guidelines

Prerequisites To participate in this evaluation, you must meet the following two criteria: (1) be a Master's or Ph.D. student in Computer Science or Computational Linguistics, and (2) demonstrate English proficiency at C2 level or higher.[a] If you do not meet both criteria, we kindly ask you to refrain from participating in this task. Eligible participants are encouraged to follow the instructions below carefully.

Instructions The following section provides detailed descriptions of the evaluation metrics and criteria used in this study. Please review the accompanying source video and the candidate summaries thoroughly. After evaluating each summary, assign scores based on the five criteria below, using a 1-to-5 Likert scale where higher scores indicate better quality:

- **Faithfulness:** Assess the accuracy of the summary in representing the content of the source video. A faithful summary should adhere closely to the source material, avoiding contradictions, misinterpretations, or unverified information.
- **Relevance:** Measure how well the summary includes the topics and themes central to the source video. A relevant summary should focus on the content that is most pertinent to the original video.
- **Informativeness:** Evaluate the extent to which the summary captures the main points and essential details of the source video. An informative summary should provide a clear and comprehensive understanding of the video's core ideas and findings.
- **Conciseness:** Determine the efficiency of the summary in conveying information. A concise summary should avoid redundancy and extraneous details while retaining all critical information from the source video.
- **Coherence:** Examine the logical flow and overall structure of the summary. A coherent summary should present information in an organized and easy-to-follow manner, ensuring that ideas connect naturally and transitions between points are smooth.

Rating System For each metric, use the following Likert scale:

- 1 (Worst): Does not meet the criteria at all.
- 2 (Poor): Meets the criteria minimally.
- 3 (Fair): Meets the criteria adequately.
- 4 (Good): Meets the criteria well.
- 5 (Best): Fully meets the criteria.

Overall Ranking After assigning scores to each summary for the individual criteria, rank all candidates from best to worst based on their overall quality. Consider the summaries' performance across all criteria when determining the final rankings.

[a] https://en.wikipedia.org/wiki/C2_Proficiency

Figure 19: A snapshot of the experimental instructions provided to human evaluators.

O Prompt for `GPT-o1` to Evaluate Summary Quality

Source Video: `{Source Video}`
Candidate Summary: `{Candidate Summary}`
You are tasked with evaluating the quality of the candidate summary based on the provided source video. Please adhere strictly to the following evaluation guidelines and scoring criteria to ensure a consistent and objective evaluation.
Evaluation Guidelines: `{Guidelines}`
Instructions for Output:

- Provide your evaluation using the following format, outputting scores only.
- Assign a score from 1 to 5 for each dimension, with 1 being the lowest and 5 being the highest.

Output Format:

- Faithfulness: [Score]
- Relevance: [Score]
- Informativeness: [Score]
- Conciseness: [Score]
- Coherence: [Score]

If you encounter ambiguity in evaluating any dimension, prioritize adherence to the evaluation guidelines and provide the most accurate score possible based on the provided information. Do not include any additional comments or justifications in your response.

Figure 20: Prompt for `GPT-o1` to evaluate summary quality.