Early Exploration into AI-Assisted Visual Analytics for Dynamic Videos

Qi Guo* — Xi'an Jiaotong-Liverpool University
Junyi Li† — University of Liverpool
Jiayi Hong‡ — Laval University
Lijie Yao§ — Xi'an Jiaotong-Liverpool University

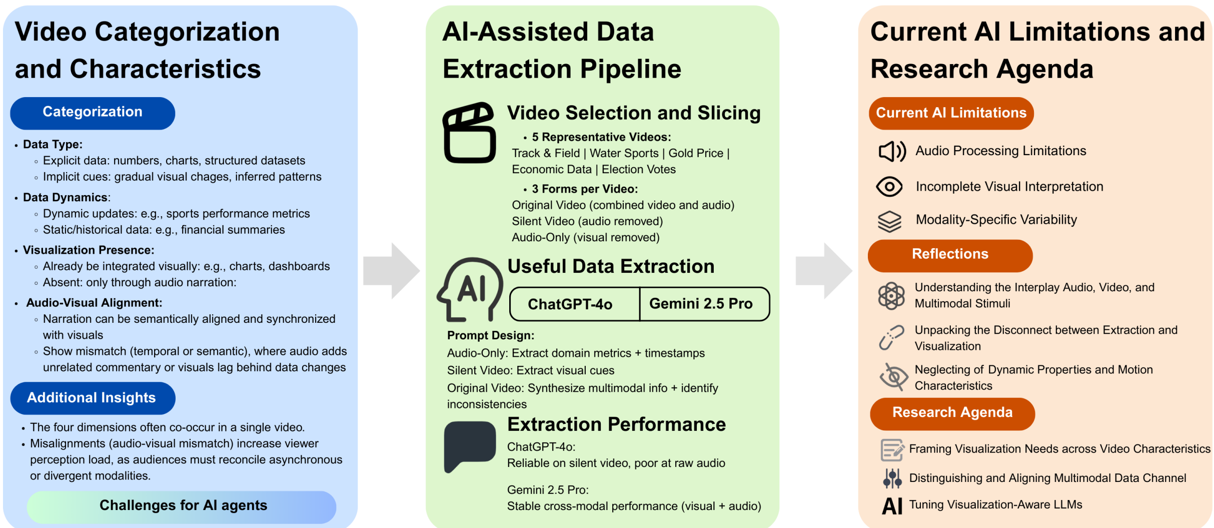

Fig. 1: The working pipeline of our early exploration.

Abstract

We present a preliminary investigation into the capabilities of current large language models (LLMs), i. e., ChatGPT and Gemini, in supporting visual analytics tasks for videos containing dynamically changing information. Videos are inherently multimodal, combining visual frames, audio narration, and sometimes text—often with inconsistencies or redundancies across channels—which poses challenges for reliable data extraction. While recent advances in video understanding have improved general-purpose AI performance, relatively little work has explored how generative AI can extract, prepare, and visualize data from videos through prompts, particularly where multimodal conflicts, dynamic updates, and moving entities are involved. To explore this space, we first categorize information-bearing videos along four dimensions: data type, data dynamics, visualization presence, and audio-visual alignment. We then apply LLMs to extract and structure information from representative video samples to support downstream visualization. We conclude with reflections and outline a research agenda for AI-assisted video-based visual analytics. Our OSF repository is at osf.io/ygn4c/.

Index Terms: Visualization, video analysis, large language model.

1 Introduction

Large language models (LLMs) such as GPT and Gemini have rapidly advanced, demonstrating strong general-purpose capabilities across tasks ranging from programming assistance to natural language conversation. However, their potential to support domain-specific tasks, especially those requiring contextual reasoning and nuanced understanding, remains an open area of investigation.

One such underexplored domain is video-based visual analytics. Videos are inherently multimodal, combining visual frames, audio narration, embedded graphics, text overlays, and dynamic entities. These channels offer rich contextual cues but also introduce challenges, such as misalignment between modalities, redundancy, and temporal complexity. In videos with changing entities, such as news segments, sports broadcasts, or explainer content, these challenges are amplified by rapid information updates and moving elements. While recent advances in video understanding have improved AI performance in captioning [11], summarization [8], and retrieval [12], relatively little work has examined how current models support analytical tasks involving structured information extraction and multimodal reasoning. Unlike generic video understanding, domain-oriented video analytics requires interpreting evolving data states, resolving multimodal inconsistencies, and preparing information in forms suitable for downstream analysis or visualization.

To begin addressing this gap, we conduct a preliminary investigation into how general-purpose LLMs perform in this context. We introduce a categorization framework for information-rich videos along four dimensions: *data type*, *data dynamics*, *visualization presence*, and *audio-visual alignment*, which enables us to characterize the complexity of video content and assess its implications for LLM-supported analysis. Using representative examples, we evaluate the capacity of current LLMs to extract, organize, and structure information for visual analytics tasks. Our findings highlight both the current capabilities and limitations of these models when applied to temporally dynamic, multimodal video content. In addition, we outline reflections and a research agenda in directions of LLM-assisted data extraction, preparation, and visualization for video content. We summarize an overview of our work in Fig. 1.

*This author contributes equally. e-mail: qiguo4123@gmail.com
†This author contributes equally. e-mail: junyi7906@gmail.com
‡e-mail: jiayi.hong@ift.ulaval.ca
§Corrresponding author, e-mail: yaolijie0219@gmail.com

2 RELATED WORK

We introduce the related work from two aspects: AI-assisted visualization tasks and video information analysis.

2.1 LLMs in General Visualization Tasks

LLMs have been explored by visualization researchers in supporting general visualization tasks from multiple perspectives, including design and analytics. For example, Shin et al. [10] proposed Visualizationary to demonstrate how LLMs can assist visualization designers by automating critique generation, while Wang et al. [13] developed DracoGPT to extract the visualization design preferences. Regarding visual analytics tasks, Zhao et al. [16] introduced LightVA, an LLM-agent-based framework for lightweight visual analytics, where large language models autonomously plan and execute tasks such as data wrangling, chart generation, and interpretation.

More directly related to our work, Hong et al.'s work [2] reveals the challenges in LLM's interpretation of data visualizations: when evaluating ChatGPT-4 and Gemini's ability to interpret modified visualizations (e. g., line charts with randomized monthly oil prices, scatterplots of males' height-weight data), they find LLMs focus on static visual features (e. g., line color, scatter size) rather than dynamic data changes (e. g., oil price trends over months, deviations in height-weight distribution).

2.2 LLM-supported Applications in Video-Related Tasks

Despite the challenges, researchers have begun exploring LLMs in video-related tasks. For instance, Zhou et al. [18] survey how generative AI, particularly LLMs, advances video generation and understanding. Video-3D LLM [17] adds 3D positional encoding to highlight spatial relationships, while VideoChat [6] links video models with LLMs via a neural interface trained on spatiotemporal cues. For information extraction, TempCompass [7] benchmarks temporal perception using conflicting video pairs, while CVRR-ES [3] proposes Dual-Step Contextual Prompting for complex reasoning. TimeChat [9] and MA-LMM [1] tackle long-term video understanding via time-sensitive modeling and memory augmentation. Gemini [4] can enhance the perception of data point trajectories through bundling and vector fields, but currently does not support strategies related to trajectories.

More directly relevant to visualization, Lee et al.'s Sportify [5] shows how large language models (LLMs) can generate personified narratives that are integrated with embedded visualizations to support question answering about basketball tactics in sports videos. Yao et al.'s work on SwimFlow [15] used standard CV algorithms to combine multiple videos into one with specific perspective requirements for embedding visualizations. Chen et al. proposed VisCommentator [19] to facilitate the creation of table tennis's augmented videos with data insights and visualizations recommendations.

Overall, although LLMs capable of processing both speech and vision have made significant progress, most of the existing multimodal systems still neglect the data extraction and alignment from both continuous visual frames and speech. In contrast, our work incorporates audio as a key modality, examining its informational content extraction, and supplementary, repetition, and alignment with visual frames.

3 VIDEO CATEGORIZATION & CHARACTERISTICS

To assess how well current LLMs can extract information from video content, we first conducted a broad search to collect a diverse set of videos and examined their characteristics. This section outlines our video collection process and categorization strategy.

Video Search. We gathered video samples through iterative snowball sampling, beginning with sports videos due to their prevalence in visualization research and their inherently dynamic, data-rich nature. As our search expanded, we observed that sports content was frequently embedded within broader news reporting. This led us to explore news coverage more systematically, where we identified additional domains featuring prominent data visualizations, including politics (e. g., election results), finance (e. g., stock prices, gold prices), infrastructure development (e. g., high-speed rail, highway projects), and weather reporting (e. g., temperature trends).

Apart from videos with explicit data encoding, we encountered content featuring implicit or gradual information changes, including disease progression, environmental change, animal behavior, and human psychology. Unlike data-integrated videos, such examples were often slower-paced, lacked formal data encodings, and required inference to identify evolving data states.

Video Categorization and Characteristics. To better understand how LLMs may support information extraction from video content, we categorized videos based on their underlying data characteristics. Our goal was to distinguish the types of information conveyed, the dynamics of data change, and the degree to which information is already visualized or aligned across modalities. For example, sports videos often present continuous, real-time data: an athlete's location is visible within the video frames, while embedded dashboards may show live performance metrics such as speed [15]. Audio narrations often reinforce or complement these visuals with spoken data, such as scores or rankings [19]. In contrast, finance videos (e. g., those discussing gold prices) tend to focus on historical data. Visualizations are often static and already embedded in the video, and narration typically provides explanatory context rather than reporting live data updates.

Based on observations from a range of such examples, we propose four dimensions to categorize information-bearing videos:

Data Type: whether the video presents explicit data (e. g., numbers, charts, structured datasets) or relies on implicit cues (e. g., gradual visual changes or inferred trends). We did not break down the data type further with continuous/numerical and categorial/qualitative data at the current stage. But we agree it is an interesting dimension to explore LLM's understanding of fine data in videos.

Data Dynamics: whether the data updates in real-time or remains static throughout the video.

Visualization Presence: whether visual representations of the data are already integrated into the video content (in visual form or only in auditory form).

Audio-Visual Alignment: whether the audio narration is semantically aligned and temporally synchronized with the visual content, or whether it introduces unrelated or asynchronous commentary.

Our four dimensions are not mutually exclusive and often co-occur within the same video. For instance, videos that present explicit data—such as financial summaries or sports coverage—frequently include embedded visualizations. However, these visual elements are not always temporally aligned with the content. In sports videos, for example, new data points emerge rapidly, yet the overlaid visualizations often remain static, failing to reflect real-time updates or key highlights. Similarly, in financial videos, narration may elaborate on factors unrelated to the on-screen visualizations, introducing semantic divergence.

Such misalignments may increase perception load for viewers, who need to reconcile asynchronous or unrelated modalities. We refer to this phenomenon as *audio-visual mismatch*, where narration and visuals diverge either temporally or semantically. This mismatch presents a challenge for AI agents attempting to extract coherent, context-aware information from video content—especially when tasked with structuring it for downstream analysis or visualization.

4 A PRELIMINARY EXPLORATION

Followed by the Hong et al.'s work [2], who evaluated the capabilities of ChatGPT and Gemini on the interpretation of data visualization, we conducted a preliminary exploration with these two LLMs (i. e., ChatGPT-4o and Gemini 2.5 Pro) to assess their capabilities in supporting visual analytics tasks for dynamic video content. Adopting a two-stream strategy, one author focused on visual features while the other focused on audio features, enabling separate evaluation of modality-specific characteristics prior to integration. Here we report our exploration from a high level; a detailed version is provided in Appendix A and B.

Step 1: Video Selection and Slicing. To investigate AI-assisted data extraction from videos, we curated a small, diverse corpus of five video clips across sports, finance, government reporting, science documentaries, and technology news. These domains capture varied information delivery styles including real-time metrics like sports scores, structured narration such as financial summaries, and embedded visual explanations.

From the visual perspective, we examined how motion, overlay graphics, and dashboards conveyed evolving content. Sports and election videos featured frequent updates across multiple data streams (e. g., athlete positions, vote tallies), while finance clips often focused on singular metrics evolved over time (e. g., gold price trends). In contrast, domains such as medicine or environmental science tended to rely on static or pre-rendered visualizations (e. g., animations or simulations) that lacked extractable raw data. We excluded these clips from further analysis due to their limited suitability for AI-based information parsing.

From the audio perspective, we prioritized clips with explicit, structured narration that communicated data-rich content. In sports videos, commentary provided key details such as dive codes, rankings, or serve speeds—data not always available visually [19]. Financial and technical clips frequently included economic figures and specifications in voiceover, while scientific documentaries used narration to convey abstract relationships or statistical insights not directly shown on screen.

Based on this dual-modality assessment, we selected five exemplar topics: track and field, water sports, gold price reports, macroeconomic updates, and election results. Each clip was processed into three modalities using LLMs: the original version (audio+video), an audio-only version, and a silent video version, resulting in a dataset of 15 files. While both ChatGPT-4o and Gemini 2.5 Pro handled audio-video separation effectively, they were unable to segment video content by time intervals. We therefore manually identified and extracted high-information scenes for downstream analysis.

Step 2: Useful Data Extraction. We used the three forms (audio-only, silent video, and original video) of each representative video to examine the potential of generative AI in extracting structured information from video content. We separate the modalities of a video to evaluate how different information channels contribute to extractable insights.

For each form, we issued prompts tailored to the video's domain—targeting relevant metrics such as numerical indicators in finance, performance features in sports, or evolving trends in news coverage. Prompts were iteratively refined when needed to recover missed or ambiguous details. We documented all outputs, including the agents' chain-of-thought reasoning and structured tabular results. All stimuli and results are provided on OSF repository (see Appendix). To avoid cross-contamination, a strict memory-reset protocol was followed between stimulus conditions.

Audio-based prompts focused on spoken data—such as quantified metrics ("$34B revenue") or descriptive commentary ("entry angle"). Silent video prompts targeted visual cues, including posture, movement quality, splash size, and embedded text or graphical overlays. Prompts for the original video condition emphasized audio-visual alignment, asking agents to synthesize multimodal content and identify inconsistencies. An example prompt input, chain-of-thought, and output given by LLM can be seen in Fig. 2.

In our exploration, Gemini 2.5 Pro demonstrated consistent performance across video/audio modalities, effectively extracting and organizing information from both visual and auditory streams. In

VC_A_Diving_SilentVideo

Q: (1st round) Analyze this diving silent video, identify and record: the body posture of the athlete at the moment of diving (such as 'straight body' or 'bent body'); the type and height of the diving board / platform; the size and effect of the water splash (described as 'small', 'medium', or 'large'); the score board numbers shown in the picture, and other related information. Then summarize and present these data in a concise and orderly manner, output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references, and do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in english.
(2nd round) Was there no data point between 00:10 and 00:28? If there was one, please add it.
(3rd round) Please review this video again carefully to ensure that no data points are missed. Then output the data table and your chain of thought.

A: (3rd round) Systematic Review: I conducted a multi-pass review of the video. The first pass focused on the overall sequence of events. Subsequent passes focused on specific details: the environment, the athlete's equipment, the technical execution of the dive (broken down into take-off, flight, and entry), and the final results and reactions.
Initial Context (Coaches' Reaction): I began at 00:00. The video opens with the German coaches' reaction. Their body language and hand gestures, mimicking an entry, are not celebratory but analytical, immediately suggesting a flaw in the dive's finish. This is an important contextual data point.
Dive Execution & Equipment: I analyzed the multiple replay angles to extract key technical details.
Platform: The replay at 00:04 clearly shows a high, fixed 10-meter platform.
Equipment: The close-up slow-motion at 00:13 reveals the athlete is wearing black wrist guards. This is a significant detail, often used for support after injury or to protect against the high impact of platform diving.
Body Posture: The replays confirm the athlete maintains a compact "tuck" position throughout the dive's rotations (00:07, 00:17).
Performance Analysis (Splash & Error): The primary performance indicators were the splash and the technical reason for it.
Splash: The entry at 00:10 creates a very large splash, which is a clear visual indicator of a sub-optimal dive.
Technical Flaw: The crucial slow-motion replay at 00:22 provides the definitive reason for the large splash. It clearly shows the athlete's body has rotated *past the vertical axis* upon water entry. This is a classic "over-rotated" entry, a significant deduction for judges.

VC_A_Diving_SilentVideo

A: Here is the data table:

TimeStamp	Data Points	Data References
00:00	**Coaches' Reaction:** Analytical / Disappointed	The German coaches are shown with analytical expressions, using hand gestures that mimic the dive's entry, indicating a focus on a flaw.
00:04	**Diving Platform:** 10-Meter Platform	The athlete is shown diving from a high, fixed, concrete platform.
00:07	**Body Posture:** Tuck Position	The athlete performs the dive with her body bent at the waist and knees, holding her shins in a compact tuck.
00:10	**Water Splash:** Large	A large and turbulent splash is created upon entry into the water, indicating a non-vertical entry.
00:13	**Athlete Equipment:** Wrist Guards	The slow-motion replay shows the athlete is wearing black guards on both wrists for support.
00:22	**Key Point / Deduction Reason:** Over-rotated Entry	The slow-motion replay shows the athlete's body entering the water angled forward, past the vertical line, confirming an over-rotation.

Fig. 2: The chain-of-thought and the data table given by Gemini for a silent diving video. It includes three rounds of prompt words(Q), the third round of thoughts generated by the generative AI(A), and the final result(Data Table). Blue highlighting shows reasoning processes.

contrast, ChatGPT-4o produced more variable results: it performed reliably on silent videos with prominent visual features but struggled to parse and interpret audio content without external transcription tools. Across video/audio forms, we observed extraction failures in both models, such as incomplete outputs or garbled characters, particularly in audio-dominant clips.

5 Reflections and Research Agenda

Here we present the observation from our preliminary exploration, discuss our reflections, and outline a research agenda.

5.1 Observations

Our exploration is preliminary but still shows several limitations in how current LLMs process real-world video with data-related content across modalities:

Audio Processing Limitations: Despite expectations, ChatGPT-4o was unable to extract structured audio content such as timestamps, numerical values, or domain-specific cues, instead returning generic metadata (e. g., sampling rate) or vague sound descriptions. These outputs were unusable for visualization tasks. This limitation is partly due to the absence of Whisper-based transcription in the default ChatGPT-4o interface. While While accessible via API, we deliberately tested the agent in its standard setup to mirror common visualization workflows, where external infrastructure is rarely deployed.

Incomplete Visual Interpretation: For both original and silent video forms, AI responses typically omitted salient visual elements (for example, a moving entity) or produced incorrect numerical values from the frames, such as malformed percentages or unreadable characters. Also, when asked the AI to track an athlete moving on a track, the tracking data given was incorrect, according to our replotting result.

Modality-Specific Variability: The AI-analysis performance varied depending on video characteristics—static scenes with structured visuals (e.g., swimming posture, vote count displays) yielded higher extraction accuracy, whereas continuous motion or unstructured audio (e.g., ambient news clips) posed greater challenges and led to limited to no useful data extraction.

Our observations reveal the need for visualization-oriented multimodal models, especially for multimodal corpora like videos. To be able to perform context-aware reasoning and reliable parsing across asynchronous or multiple input channels.

5.2 Reflections

From the outset, our ultimate goal is to explore a seamless pipeline: enabling LLMs to extract structured data from videos and also automatically translate that data into meaningful visualizations. However, our practical exploration showed that current mainstream AI agents could not reliably reach the data table stage, and the subsequent visualization step remains unclear. In this section, we discuss our reflections.

Understanding the Interplay of Audio, Video, and Multimodal Stimuli: From our exploration, we found that each modality—audio, visual, and combined video—plays a distinct role in data extraction, especially for some domain-specific videos, where a single modality would be the main source of data. For example, in finance clips, audio emerged as the primary source of structured information (e. g., "$34 billion revenue", "2% tax increase"), while visuals served as background context. And so it is in some visually rich content as well, like diving, commentary added interpretive depth (e. g., explaining "over-rotation" deductions) that visual content from frames could not provide. In contrast, silent video stimuli revealed limited extractable data, which may be caused by the underutilization of the AI capabilities. Although humans can simply "read" the implicit data from the silent video (e. g., an athlete moves quickly/slowly), it seems that AI agents are not able to identify such information through our designed prompts. For combined videos, audio can provide additional information, for example, in news videos, the audio says "150 mph top speed" compared to the vague impressions of "fast-moving trains" derived from visuals.

Unpacking the Disconnect between Extraction and Visualization: A core challenge emerged in our exploration: despite relying on two widely used, mainstream LLMs, neither was able to meet the baseline requirements for domain-specific tasks such as visualization. Even when we decomposed the task into smaller components and issued precise, carefully designed prompts, the outputs—particularly at the data extraction stage—remained insufficient. While one explanation may lie in current computational limitations, a more likely cause is the agents' limited understanding of visualization tasks and their underlying workflows. This gap highlights research opportunities: as visualization researchers, we are uniquely positioned to define, structure, and communicate the domain-specific reasoning required to support such tasks effectively. In particular, we can generate representative datasets and develop interaction strategies that align model capabilities for visualization tasks.

Neglecting of Dynamic Properties and Motion Characteristics: Unlike static images, video content unfolds over time, introducing dynamic properties between frames. Entities within the frame may move, shift, or change state, resulting in observable displacement and other temporal patterns. While object detection and tracking algorithms can reliably identify and follow moving entities, current AI agents struggle to extract structured motion data, such as trajectories, from videos containing multiple dynamic targets. One possible reason lies in the framing of tasks in mainstream AI research, which has traditionally emphasized answering "what" questions: identifying what an object is and what it is doing. This focus often overlooks the underlying data generated by motion itself—such as location sequences, movement paths, speed, and acceleration [14]. Capturing and reasoning about these motion characteristics remains a gap in current generative models and suggests a direction for more temporally aware AI capabilities in support of visual analytics.

5.3 A Research Agenda

Building on our observations and reflections, we outline a research agenda aimed at advancing AI-assisted visual analytics for dynamic, multimodal video content.

Framing Visualization Needs across Video Characteristics: As discussed in Sec. 3, video modality, structure, and temporal dynamics shape what information is available and how it can be visualized. Different domains introduce distinct analytical goals, and even within a domain, user needs vary depending on expertise or intent. While audience-specific factors lie beyond our current scope, they remain an important future direction. To better align AI assistance with practical analysis needs, we propose grounding visualization support in the structural properties of video content. A promising step is to develop a task taxonomy that links content characteristics to visualization demands—clarifying what data to extract, how to structure it, and how to represent it effectively.

Distinguishing and Aligning Multimodal Data Channels: Videos combine visual and auditory streams that often present asynchronous, overlapping, or conflicting information. As noted in Sec. 3, such mismatches, referred to as audio-visual mismatches, pose significant challenges for information extraction. Practically, data from these channels may be complementary, redundant, irrelevant, or temporally misaligned. A critical research step is to develop methods that can distinguish informative from non-informative content in each modality, extract non-redundant data, and synchronize the results across modalities to build a coherent, temporally aligned dataset for downstream visualization.

Tuning Visualization-Aware LLMs: Our exploration suggested that current LLMs lack a conceptual model of how visualization tasks unfold, struggling even when provided with precise, small-scale prompts for tasks like data extraction. This indicates a deeper disconnect between LLMs' general-purpose capabilities and the specialized reasoning required in visualization workflows. While retraining large models may be out of reach for most visualization researchers, we argue that the visualization community has a critical role to play in tuning future LLMs.

Rather than expecting LLMs to automatically generalize to visualization contexts, we propose a research opportunity centered on visualization-aware language models. This could include lightweight modular training, such as prompt-based adaptation, that exposes agents to core visualization principles, structured examples of visualization tasks, and domain-specific annotation schemas. Alternatively, hybrid systems could be developed, pairing existing large models with visualization-specific planning modules or toolkits that scaffold agent behavior across pipeline steps (e. g., data extraction, transformation, encoding).

While we do not claim deep expertise in LLM architecture, we see an opportunity, and a need, for interdisciplinary collaboration. We call on experts in AI, NLP, and multimodal reasoning to partner with visualization researchers in shaping agents that are not only capable of understanding charts but also the reasoning processes that underlie visualization design and analysis.

6 Conclusion

We conduct a preliminary exploration to assess how LLMs can help with data identification, extraction, preparation, and visualization within videos. Our results show that although the two prevalent LLMs can do some tasks, the given output is still far from expectations. Our ultimate goal is to advance automated techniques for data extraction, preparation, and visualization in videos with rich temporal and perceptual dynamics. We discuss and reflect on our results and outline a research agenda for future research.

Supplemental Materials

We provide the following supplemental materials to support our work: (a) An appendix of detailed descriptions of our exploration process for both video and audio streams. (b) A set of decks that illustrate our full prompts and the whole chain-of-thought and LLM outputs that we obtained. (c) An OSF repository where we uploaded all the videos assessed in our early exploration, the two materials listed above, and an author version of this paper. Our OSF repository is at osf.io/ygn4c/.

Acknowledgments

Lijie Yao is partially funded by the XJTLU RDF, grant № RDF-24-01-062, and XJTLU TDF, grant № TDF2425-R30-283.

References

[1] B. He, H. Li, Y. K. Jang, M. Jia, X. Cao, A. Shah, A. Shrivastava, and S.-N. Lim. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13504–13514, 2024. doi: 10.1109/CVPR52733.2024.01282 2

[2] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do LLMs Have Visualization Literacy? An Evaluation on Modified Visualizations to Test Generalization in Data Interpretation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–13, 2025. doi: 10.1109/TVCG.2025.3536358 2

[3] M. U. Khattak, M. F. Naeem, J. Hassan, M. Naseer, F. Tombari, F. S. Khan, and S. Khan. How Good is my Video-LMM? Complex Video Reasoning and Robustness Evaluation Suite for Video-LMMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 3651–3660, June 2025. 2

[4] Y. Kim and J. Heer. Gemini: A Grammar and Recommender System for Animated Transitions in Statistical Graphics. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):485–494, 2021. doi: 10.1109/TVCG.2020.3030360 2

[5] C. Lee, T. Lin, H. Pfister, and C. Zhu-Tian. Sportify: Question Answering with Embedded Visualizations and Personified Narratives for Sports Video. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):12–22, 2025. doi: 10.1109/TVCG.2024.3456332 2

[6] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao. VideoChat: Chat-Centric Video Understanding. *SCIENCE CHINA Information Sciences*, 2025. doi: 10.1007/s11432-024-4321-9 2

[7] Y. Liu, S. Li, Y. Liu, Y. Wang, S. Ren, L. Li, S. Chen, X. Sun, and L. Hou. TempCompass: Do Video LLMs Really Understand Videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772. ACL, Bangkok, Thailand, Aug. 2024. doi: 10.18653/v1/2024.findings-acl.517 2

[8] N. M. Ranjan, G. Mate, D. H. Patil, A. Jadhav, S. Adhav, and R. Umbare. Inflated 3D Video Summarization: A Comprehensive Review. In *International Conference on Innovative Mechanisms for Industry Applications*, pp. 170–176, 2023. doi: 10.1109/ICIMIA60377.2023.10426498 1

[9] S. Ren, L. Yao, S. Li, X. Sun, and L. Hou. TimeChat: A Time-sensitive Multimodal Large Language Model for Long Video Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024. doi: 10.1109/CVPR52733.2024.01357 2

[10] S. Shin, S. Hong, and N. Elmqvist. Visualizationary: Automating design feedback for visualization designers using llms. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–17, 2025. doi: 10.1109/TVCG.2025.3579700 2

[11] P. Song, D. Guo, J. Cheng, and M. Wang. Contextual Attention Network for Emotional Video Captioning. *IEEE Transactions on Multimedia*, 25:1858–1867, 2023. doi: 10.1109/TMM.2022.3183402 1

[12] M. Sun. An Intelligent Retrieval Method for Audio and Video Content: Deep Learning Technology Based on Artificial Intelligence. *IEEE Access*, 12:123430–123446, 2024. doi: 10.1109/ACCESS.2024.3450920 1

[13] H. W. Wang, M. Gordon, L. Battle, and J. Heer. Dracogpt: Extracting visualization design preferences from large language models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):710–720, 2025. doi: 10.1109/TVCG.2024.3456350 2

[14] L. Yao, A. Bezerianos, R. Vuillemot, and P. Isenberg. Visualization in Motion: A Research Agenda and Two Evaluations. *IEEE Transactions on Visualization and Computer Graphics*, 28(10):3546–3562, Oct. 2022. doi: 10.1109/TVCG.2022.3184993 4

[15] L. Yao, R. Vuillemot, A. Bezerianos, and P. Isenberg. Designing for Visualization in Motion: Embedding Visualizations in Swimming Videos. *IEEE Transactions on Visualization and Computer Graphics*, 30(3):1821–1836, Mar. 2024. doi: 10.1109/tvcg.2023.3341990 2

[16] Y. Zhao, J. Wang, L. Xiang, X. Zhang, Z. Guo, C. Turkay, Y. Zhang, and S. Chen. Lightva: Lightweight visual analytics with llm agent-based task planning and execution. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6162–6177, 2025. doi: 10.1109/TVCG.2024.3496112 2

[17] D. Zheng, S. Huang, and L. Wang. Video-3D LLM: Learning Position-Aware Video Representation for 3D Scene Understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, June 2025. 2

[18] P. Zhou, L. Wang, Z. Liu, Y. Hao, P. Hui, S. Tarkoma, and J. Kangasharju. A Survey on Generative AI and LLM for Video Generation, Understanding, and Streaming. *arXiv preprint arXiv:2404.16038*, 2024. Preprint. 2

[19] C. Zhu-Tian, S. Ye, X. Chu, H. Xia, H. Zhang, H. Qu, and Y. Wu. Augmenting Sports Videos with VisCommentator. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):824–834, 2022. doi: 10.1109/TVCG.2021.3114806 2, 3

Early Exploration into AI-Assisted Visual Analytics for Dynamic Videos

Appendix

In this appendix, we offer additional material, including a detailed description of our preliminary exploration for both video and audio streams, as well as a full version of the prompt, chain-of-thought, and results that we obtained. The OSF repository is at osf.io/ygn4c/.

A VIDEO STREAM

A.1 Video Selection and slicing

We found that sports requires analyzing the positional changes of movement, finance needs to extract known data points based on audio and video images, and government needs to extract known data from video images. However, the sought-after medicine, experiment, natural movement, and behavioral analysis are in the form of animated videos, and the videos have already been visualized, so no other data need to be visualized. Therefore, we finally selected five videos, namely track and field competitions, swimming competitions, gold price trends, economic data trends, and election vote counting. Artificial intelligence (chatgpt4o) then completed audio-video separation to obtain original videos, pure audio, and silent videos, totaling 15 videos. In the videos we have selected, Sports and government both show continuous changes in multiple data points, while finance shows continuous changes in a single data point.

After the screening process, we watched the video again to find the segments where the data we needed appeared, and then we sliced the video. I tried using artificial intelligence (chatgpt4o) to perform the slicing. I sent the video that needed to be sliced to artificial intelligence (chatgpt4o) and gave it two instructions: (1) "Please provide me with video clips that contain data points for visualization purposes" (2) "I extract data points for visualization, Motion tracking. I would like it in the form of video clips." Artificial intelligence (chatgpt4o) can receive and read the video, but when executing the slicing instructions, it will display that artificial intelligence cannot handle long videos, even if it cannot slice the first two seconds.However, artificial intelligence (chatgpt4o) can handle audio-video separation very well. Just give it the instruction "Separate the audio and video of this video", and artificial intelligence (chatgpt4o) will provide you with a silent video in MP4 format and a pure audio in MP3 format, and they are completely correct.

We use artificial intelligence (chatgpt4o) for the extraction and analysis of video data. First, we edit the serial numbers of the screened videos, pure audio, and silent videos, and group them. After completing the preparatory steps, send them respectively to artificial intelligence (chatgpt4o), and issue prompts. Record the results given by artificial intelligence (chatgpt4o) as well as the thinking process.

A.2 Editorial Number

In order to conduct the analysis in an orderly manner, we need to analyze the following video editing sequence names (1) Video clips (marked as VC_x_ for example, VC_A_ = Gold Price, VC_B_ = US Election...): Original video clips you cut from a long video, along with the accompanying audio; (2) silent video (marked as VC_x_SilentVideo): The video without audio obtained from the video clips; (3) Audio (marked as VC_x_Audio): The sound extracted from the video clips.

A.3 Operation Overview

After the sequence names have been edited, we will consider the following sequence as a set of correct sequences for operation, namely [VC_A_Audio, VC_A_SilentVideo, VC_A], [VC_B_Audio, VC_B_SilentVideo, VC_B], [VC_C_Audio, VC_C_SilentVideo, VC_C]... We need to repeat the entire task cycle listed below for each element, namely VC_A_Audio, VC_A_SilentVideo, VC_A, etc.

We sent each element to Ghatgpt 4o and gave it prompts, such as "Please extract data from this video. Please explain your reasoning process step by step", "What useful data points can be obtained from the background audio of the video", "Extract data points from the content in the audio", etc. After receiving the prompts, Ghatgpt 4o would provide its thinking process and response. We tried and recorded different prompt phrases and repeatedly operated until we found the most effective prompt phrase and the corresponding thinking process.

A.4 Detailed records

During the operation, when analyzing the pure audio and original video of the track and field competition, we attempted several times to use Artificial intelligence (chatgpt 4o) to read the data in the audio content, but all attempts failed. Therefore, for other types, we only analyzed the original video and silence video, and did not consider the data in the audio content.

A.4.1 Analysis Record of Pure Audio

a. VC_A_Audio = Track of track and field competition. The audio is sent to artificial intelligence (ChatGPT 4o) and prompt is given (1) What is the main content of this audio? Artificial intelligence (ChatGPT 4o) provides the most basic information, such as sampling rate, channels, and sample width, and indicates that the Whisper transcription model is currently unavailable in this environment. Artificial intelligence (ChatGPT 4o) does not able to read pure audio.

A.4.2 Analysis Record of Original Video

a. VC_A = Track Running: In this video, eight athletes compete separately. The video is sent to AI (ChatGPT 4o) and instructions are given: (1) "Please extract the data from this video. Explain your reasoning step by step." Artificial intelligence (ChatGPT 4o) provided a relatively detailed chart of positional changes, but this is only data extracted from the video frames. (2) "What useful data points can be obtained from the background audio of the video?" Artificial intelligence(ChatGPT 4o) provided an analysis from four aspects: voice and dialogue, environmental sounds, time markers, and musical and emotional cues. However, it did not extract the data from the audio. (3) "Extract data points from the audio content." Artificial intelligence(ChatGPT 4o) provided the most basic features of the audio, such as volume changes, rhythm strength, etc., but could not extract the audio data and provided other approaches. (4) "Extract the audio part of the video and convert it into subtitles."and (5) "Convert the extracted audio content into text." The results given by artificial intelligence(ChatGPT 4o) all extracted the audio part from the video, but the conversion to the text part could not be completed.

b. VC_B = Price of gold.Send the video to artificial intelligence (ChatGPT 4o) and issued the prompts (1) "Please extract the data from this video. Explain your reasoning step by step." When no clear prompts are given, artificial intelligence(ChatGPT 4o) defaults to analyzing the movement positions of the characters. Issue the second quality (2) "Please extract the important data points of the audio in the video for me." Artificial intelligence(ChatGPT 4o) successfully extracted the audio and transcribed it in WAV format. It clearly states

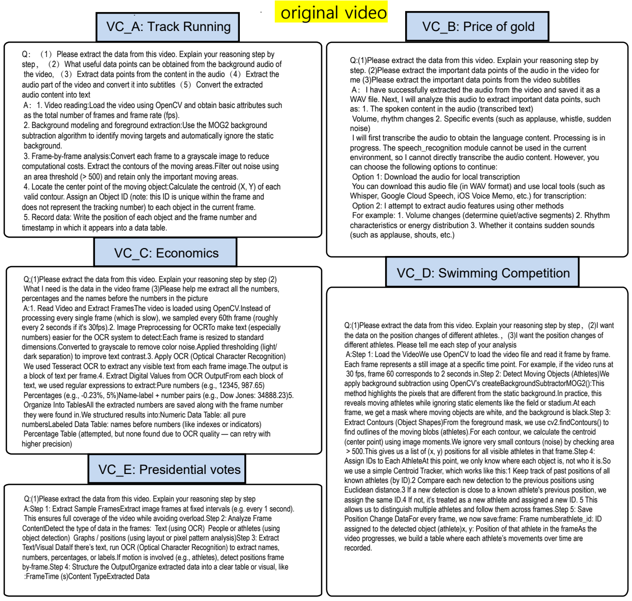

Fig. 3: chain-of-thought of the ChatGPT4o for original video

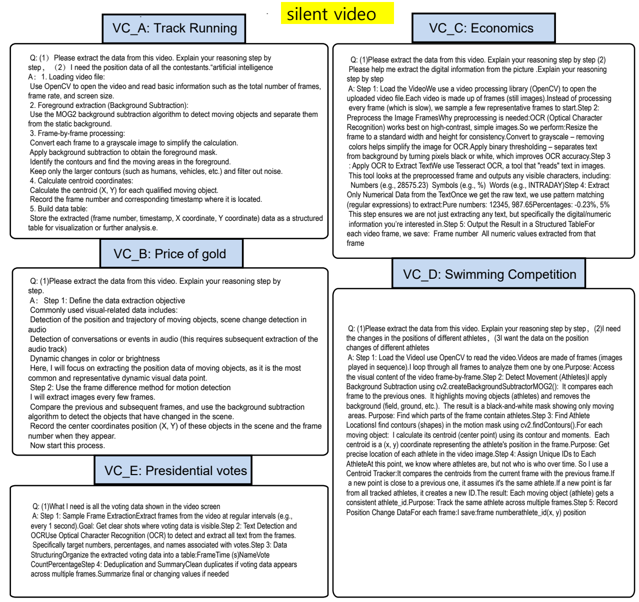

Fig. 4: chain-of-thought of the ChatGPT4o for original video

that the speech_recognition module cannot be used in the current environment, and also provides some alternative methods such as Whisper, Google Cloud Speech, and iOS Voice Memo, etc. Since this video contains subtitles, issue the third instruction (3) ”Please extract the important data points from the video subtitles.” The result directly shows that the operation cannot be completed.

c.VC_C = Economy: In this video, we need artificial intelligence(chatgpt 4o) to extract the data on the economic changes of different countries in the picture. Send the video to artificial intelligence (chatgpt 4o) and give the following prompts (1) ”Extract the data from this video. Explain your reasoning step by step”. By default,artificial intelligence(chatgpt 4o) analyzes the position changes of characters in videos. (2) ”What I need are the data in the video frame.” The table given by ai has obvious character and symbol errors, such as *28, 77.74a3,488.23; ”-.(3)”Please help me extract all the numbers, percentages and the names before the numbers in the picture.” The table provided by artificial intelligence (ChatGPT 4o) contains obvious extraction errors.

d.VC_D= Swimming Competition: This video features multiple athletes competing. I sent it to artificial intelligence (chatgpt 4o) and gave the following prompts: (1) ”Please extract the data from this video. Explain your reasoning step by step.” Artificial intelligence (chatgpt 4o) provided a diagram of the diving process. (2) ”I want the data on the position changes of different athletes.” Artificial intelligence (chatgpt 4o) only gave a few sets of data and an ellipsis. So I issued another prompt: (3) ”I want the position changes of different athletes. Please tell me each step of your analysis.” The data provided was still incomplete and only included five sets.

e.VC_E = Presidential votes: The data we need from this video is for artificial intelligence (ChatGPT 4o) to extract the real-time changing data of the two candidates from the screen. After sending the video, give the following instructions: (1) ”Please extract the data from this video. Explain your reasoning step by step.” The table provided by artificial intelligence has obvious errors and symbol issues, such as @AThe,.Guardian.

A.4.3 Analysis Record of silent Video

a. VC_A_silentvideo = Track Running. Send the silent video to artificial intelligence (ChatGPT 4o) and give the prompt (1): ”Please extract the data from this video. Explain your reasoning step by step.” Artificial intelligence (ChatGPT 4o) provided a table with the data but did not include the numbers for different athletes. So, the prompt (2) was given: ”I need the position data of all the contestants.” Artificial intelligence then provided a table with the data for each athlete.

b. VC_B_silentvideo = Price of gold. Send this silent video to artificial intelligence (ChatGPT 4O) and issued prompt (1): ”Please extract the data from this video. Explain your reasoning step by step.” Artificial intelligence operation failed and no result was given.

c. VC_C_silentvideo = Economy. This silent video was sent to artificial intelligence (ChatGPT 4o) and prompts were given (1) ”Please extract the data from this video. Explain your reasoning step by step.” and (2) ”Please help me extract the digital information from the picture. Explain your reasoning step by step.” Artificial intelligence indicated that there were obvious errors in the table, such as 3, 488.17 vi - 11,575.90 JooLavorsna amigos.

d.VC_D_silentvideo = Swimming competition. This silent video was sent to artificial intelligence (ChatGPT 4o) and prompts were given: (1) ”Please extract the data from this video. Explain your reasoning step by step.” Artificial intelligence (ChatGPT 4o) reported an error and was unable to run. (2) ”I need the changes in the positions of different athletes.” Artificial intelligence (ChatGPT 4o) reported an error and indicated ”Unable to recognize frame count.” The same prompt was given again: (3) ”I want the data on the position changes of different athletes.” This time, the artificial intelligence (ChatGPT 4o) provided a position table but only had 3 sets of data.

e.VC_E_silentvideo = Presidential votes. This silent video was sent to artificial intelligence (ChatGPT 4o) and prompts were given:(1) ”What I need is all the voting data shown on the video screen.” The table provided by the AI has obvious errors and symbol problems, such as The,., we PROJECTS A] suardianVEOX NE.

B AUDIO STREAM

We conducted the process of video filtering in two sequential steps to ensure the selection of ecologically valid and data-rich video clips for subsequent analysis. In this section, we detail the initial video corpus, the rationale behind their inclusion, and the method and criteria for the final selection.

B.1 Video Selection and slicing

The initial video corpus was curated to encompass diverse domains with distinct data characteristics. We prioritized videos characterized by explicit and specific data points in their audio tracks, with a particular focus on two scenarios: (1) cases where the visual content lacks data, and the audio provides or supplements such data; and (2) instances where the data in the visual content is temporally unsynchronized with that in the audio. We selected these categories of scenarios based on their validity and structured approach to data presentation: sports, finance and technology sectors, documentaries.

In the domain of sports, we selected clips from events such as diving, tennis, and swimming. These videos are distinguished by real-time expert commentary that provides a layer of quantitative and qualitative analysis over the visual action. For instance, diving commentary frequently includes specific dive codes (e.g., 107B), degree of difficulty, and a breakdown of judges' scoring, offering a rationale for the scores awarded that is absent from the visuals alone. Similarly, commentary in tennis and swimming often highlights key performance indicators such as serve speed, stroke rates, and split times, providing a continuous stream of data that contextualizes athletic performance. This audio-visual pairing allows for a detailed examination of performance metrics that are either not visually apparent or are given meaning only through narration.

Videos from the finance and technology sectors, such as market analyses and product launch events by Apple or Huawei, were also included. These types of videos are characterized by scripted, data-dense presentations, where the audio track is the primary vehicle for conveying complex information. For example, financial reports feature expert discussions on macroeconomic indicators and market trends, while technology launches detail specific performance and technical updates of their products. In these instances, the visual content often serves as a backdrop to an audio narrative filled with quantifiable claims and expert analysis.

Finally, we chose scientific and nature documentaries. These films combine visuals with fact - based narration, exemplify the use of didactic narration to translate complex scientific concepts and empirical data into an accessible format. The voice-over provides a structured data layer, explaining and quantifying the phenomena displayed, and often presents abstract data, such as ecological statistics, that cannot be fully conveyed by visuals alone.

B.2 Refinement of Video Clips

The initial stage of this research involved the methodical preparation of the video data. The final selection included two categories of senarios: diving and finance. We selected a total of four video clips, one of which is categorized as diving, and the other three as finance. We named these four videos VC_A_Diving, VC_B_Finance, VC_C_Finance, and VC_D_Finance, and the durations of the videos are 35 seconds, 26 seconds, 22 seconds, and 37 seconds respectively. These videos were then subjected to a precise segmentation process manually utilizing the video editing software, CapCut. Initially, we extracted the audio stream from the video content using the "Audio Separation" function within CapCut. Subsequently, to ensure the primacy of the spoken content, we performed voice separation on the extracted sounds. This process effectively filtered out ambient background noise and other non-vocal sounds, yielding a purified human voice track. Upon the completion of these procedures, three distinct categories of files were systematically generated and archived for each of the original clips: the original, unaltered video (e.g., noted as "VC_A_Diving"), a silent version of the video from which the audio track had been removed (e.g., noted as "VC_A_Diving_SilentVideo"), and the isolated human speech, which was exported and saved in the MP3 format (e.g., noted as "VC_A_Diving_Audio").We did the data preparation to ensure the quality and suitability of the dataset for subsequent analysis.

From the initial corpus, we refined four clips to prioritize controllable multimodal data characteristics critical for video analysis and LLM-driven tasks. Our selection criteria included: (1) varied patterns of audio-visual synchronization, including matches (where video frames align with audio content) and mismatches (where such alignment is absent); (2) the presence of explicit data points (quantitative or descriptive) in audio, video, or both that haven't been visualized yet. The specific reasons why we ultimately chose these videos are as follows:

VC_A_Diving: The video captures a female athlete from performing a dive at the Tokyo 2020 Olympics. This clip was finally selected because it epitomizes a scenario where audio commentary provides crucial explanatory data that is not visually apparent. The visual component of the clip primarily displays the diver's physical action—the takeoff, the rotations in the air, and the entry into the water—followed by her interaction with her coach and the display of the final score. While the visuals show the dive itself, they do not inherently explain the technical merits or flaws of the performance. The audio track, however, offers a rich layer of expert analysis that is essential for a comprehensive understanding. The commentator provides several key data points: (1) Qualitative Assessment: The commentator immediately identifies a technical fault, stating that it was "another one of those over-rotations." This judgment is reinforced later with the comment, "that wasn't one of her best, unfortunately, just losing it on the entry." (2) Biomechanical Context: The commentary provides information about the physical challenges, noting the diver is "using lots of support on those wrists" due to past injuries and that divers hit the water at "about 30-35 miles an hour." This information is not visually deducible. (3) Performance Analysis: The audio breaks down the dive's execution, observing that she was "nice and tight on the tuck" but "just ran out of air really before hitting the water." (4) Score Justification: The commentary directly links the observed flaws to the final score, stating she was "marked down for that," and points to the "40.00 score" as the outcome of the imperfect execution. In summary, this clip presents a clear instance where the audio and visual streams are mostly synchronized during the action but also function in a complementary manner. The audio does not merely describe the visual; it interprets and quantifies it, offering a layer of data-rich narrative that explains the "why" behind the visual "what". This makes it an ideal case for forming a complete understanding of an event.

VC_B_Finance: The clip, sourced from a CNBC news report, discusses the financial performance of the Walt Disney Company's theme park division. It was selected because it presents a clear case where the audio track is the exclusive carrier of quantitative data, while the visual stream is purely illustrative. The visual content consists entirely of general background footage, showing guests enjoying various attractions at Disney theme parks. These scenes serve to establish a thematic context—the subject is Disney parks—but they are devoid of any inherent or explicit data points. In stark contrast, the audio narration is dense with specific, quantifiable financial metrics. It is through the audio alone that the following key data are communicated: revenue data, growth metric, profitability analysis and investment figures. The primary reason for selecting this video is the distinct informational asymmetry between its audio and visual channels. It exemplifies a scenario where the visual information is ambient and contextual, while the audio provides a structured, data-rich narrative.

VC_C_Finance: This video clip also comes from a CNBC news report, which compares the performance of the U.S. Acela train line with high-speed rail in China.The visual part of the video clip includes multiple backup shots of trains in motion, such as interior views of the trains, and exterior shots of the Acela train running on the tracks and crossing bridges, etc. Similar to VC_B_Finance, the reason why we chose it is that it epitomizes a scenario where a data-rich narrative is conveyed exclusively through the audio, accompanied by thematically related but quantitatively inert visual footage. Compared to purely explanatory images, the audio track is the only repository of the clip's quantitative information. The narrator provides a series of precise and comparative data points, such as top speed vs. actual performance, average speed, international comparison.

VC_D_Finance: This clip, taken from a CNN news interview, was selected as an exemplary case of complex, multi-point quantitative data being delivered exclusively through conversational audio. The video captures a political candidate discussing the financial specifics of his policy proposals. In the video vlip, the visual stream is entirely static and contextual, depicting a standard studio interview between a host and a guest. The visuals provide no data. They simply show two individuals engaged in a discussion. Conversely, the audio track is exceptionally data-dense, containing a rapid-fire series of financial figures and policy metrics articulated by the speaker. The audio is the only source for understanding the quantitative aspects of the discussion, which include: policy costs, budgetary context, Revenue Generation. This clip was chosen precisely because of this profound informational asymmetry.

The video filtering process, involving both initial corpus selection and subsequent refinement, demonstrates that videos containing data have two key characteristics: their data features and how information lines up between audio and visuals.

In terms of data features, these videos have single or multiple data points. For example, the diving clip includes both qualitative comments (like judging a technical flaw) and biomechanical details, while finance clips have various financial figures. The data can be real-time (such as live commentary during a dive) or pre-prepared (like scripted financial analyses). It might flow continuously (like steady commentary in a dive) or come seperately (like specific tech stats in product launches). Some data ties to what's happening on screen (like describing a diver's movements), while others are abstract (like ecological numbers in documentaries). They also cover different types: numbers (speeds, scores, revenues) and descriptive judgments (assessing technical issues, expert opinions).

Regarding information mismatch, these videos exhibit varied alignment patterns between video and audio stream. Mismatches are evident when audio carries critical data absent from visuals. For instance, the visuals of all the three selected finance clips are contextually illustrative but data-free, while their audio tracks are densely packed with quantifiable information; similarly, the diving clip's audio provides explanatory data (e.g., technical flaws, score justification) not inherently apparent in the visuals. Additionally, we also observed partial alignment in the diving clip, where audio and visuals synchronize during the action but audio adds interpretive layers that quantify and contextualize the visual content. Collectively, these patterns highlight scenarios where audio often serves as the primary or exclusive carrier of data, with visuals functioning as a backdrop rather than a source of data.

This section details the implementation of this practice: extracting structured data from the refined video clips using generative artificial intelligence (GenAI), specifically Gemini 2.5 Pro. We designed a rigorous, domain-adapted workflow to ensure systematic data extraction across modalities, with comprehensive documentation of prompts (some involving multi-round iterations) and the model's reasoning and results. Notably, not all prompts required multi-round interactions; follow-up queries were only used when initial extractions failed to capture all necessary information. The specific figures included prompts, chains of thought, and data tables for each stimulus can be found in the Supplemental Materials section.

B.3 Stimuli and Processing Sequence

For each refined video clip—VC_A_Diving, VC_B_Finance, VC_C_Finance, and VC_D_Finance—we processed three distinct stimuli to isolate and compare data extractability across modalities.

The first stimulus, VC_x_Audio, refers to the isolated human voice track containing expert commentary (for diving) or financial narration (for finance clips). The second stimulus, VC_x_SilentVideo, is the video clip with audio removed, preserving only visual elements such as athlete movements, scoreboards, or scene contexts. The third stimulus, VC_x, denotes the original video clip with synchronized audio and visual content, which enables cross-modal validation of extracted data.

For each video, we processed these stimuli in a fixed sequence: first VC_x_Audio, then VC_x_SilentVideo, and finally VC_x.

B.4 Prompt Design and Data Extraction

To extract precise, structured data, we developed modality-specific prompts for each stimulus of a video, with focuses tailored to the unique characteristics of audio, visual, or combined modalities. Some prompts yielded complete data in a single query, while other required iterative refinement.We labeled them as “1st round” and “2nd round” in the figure(e.g., Figure 2).

B.4.1 Prompts for Audio-Only Stimuli (VC_x_Audio)

For audio-only stimuli (e.g., VC_A_Diving_Audio), prompts focused exclusively on information carried in the commentary, with no reference to visuals. The core goal was to extract domain-specific auditory data.

For diving audio (VC_A_Diving_Audio), the prompt explicitly requested extraction of ”diving move names, judges' scores and deduction reasons, and key technical points emphasized in commentary (e.g., 'Entry Angle' 'Body Posture')”. It required presentation in a table with timestamps, data points, and audio references, alongside a step-by-step explanation of reasoning. This prompt yielded complete data in a single round.

For financial audio (e.g., VC_B_Finance_Audio, VC_C_Finance_Audio, VC_D_Finance_Audio), prompts targeted ”all numerical information (e.g., dates, amounts, percentages) with their category labels (e.g., 'revenue' 'growth rate')”.

B.4.2 Prompts for Silent Video Stimuli (VC_x_SilentVideo)

For silent video stimuli (e.g., VC_A_Diving_SilentVideo), prompts relied solely on visual information, excluding any audio-related content. The focus was on extracting observable visual features.

For diving silent video (VC_A_Diving_SilentVideo), the 1st round prompt requested extraction of ”the athlete's body posture at takeoff (e.g., 'straight' or 'bent'), diving platform type and height, splash size (small/medium/large), and other relevant visual details”. Initial results omitted wrist guard information between 00:10–00:28, prompting a 2nd round query: ”Check the 00:10–00:28 segment for unextracted data points”. A 3rd round prompt further ensured completeness: ”Review the video again to confirm no visual data points are missed”.

For financial silent video (e.g., VC_B_Finance_SilentVideo, VC_C_Finance_SilentVideo, VC_D_Finance_SilentVideo), prompts focused on ”scene details, architectural features, and in-scene physical text”, explicitly excluding subtitles and logos. VC_B_Finance_SilentVideo required a 2nd round correction after initial results included subtitles: ”Please ignore the subtitles above or at the bottom. They are not considered.”. A 3rd round prompt clarified: ”Also ignore the CNBC logo. Please just analyze the video frame”. In contrast, some financial silent videos (e.g., VC_D_Finance_SilentVideo)—with static interview scenes—were fully processed in a single round.

B.4.3 Prompts for Combined Audio-Visual Stimuli (VC_x)

For combined audio-visual stimuli (e.g., VC_A_Diving), prompts integrated information from both modalities, focusing on their correspondence and consistency.

For the diving combined video (VC_A_Diving), the 1st round prompt requested "the complete process of the athlete's entire set of movements (e.g., 'running-up → jumping → somersault → entering water') and the relationship between commentary and visuals (e.g., specific timestamps where commentary mentions 'perfect entry' or deduction reasons)". Initial results lacked clarity on audio-visual sync, so a 2nd round prompt specified: "For example, the audio at 00:04 and the video at 00:10 are in sync. Please let me know and write it down in the data table".

For financial combined videos (e.g., VC_B_Finance, VC_D_Finance, VC_C_Finance), prompts focused on "linking numerical audio data to visual scenes (e.g., '$34 billion revenue' to footage of crowded parks)" and validating consistency.

All prompts required Gemini 2.5 Pro to output structured tables and detailed chain-of-thought reasoning, explaining how data points were identified (e.g., "Wrist guards observed in 00:13 slow-motion frames" or "'34 billion' extracted from audio at 00:03"). For multi-round prompts, reasoning also included explanations of refinement logic (e.g., "Initial extraction missed temporal context; 2nd round clarified via contextual inference").

B.5 Post-Extraction Validation

To prevent cross-contamination of results between stimuli, we implemented a strict memory-clearing protocol after processing each stimulus for a video. Following data extraction from VC_x_Audio, VC_x_SilentVideo, or VC_x, we instructed Gemini 2.5 Pro: "Forget our entire conversation". We verified memory clearance by asking: "What was the content of the video stimulus you just processed?" This step was repeated until the model confirmed no recollection of prior stimuli, ensuring independence between extraction tasks.

B.6 Result Documentation

We systematically documented all outputs, including single-round/multi-round prompts, the model's chain-of-thought, and final data tables, to ensure traceability. For each video, three result sets (one per stimulus) were compiled.

The first component, Prompts, involved archiving all original prompts and follow-ups, distinguishing single-round from multi-round to highlight design logic (e.g., "Excluding audio-reliant info for silent videos"). The second component, Chain-of-Thought, recorded the model's analytical process, including how "over-rotation" was identified in audio or how splash size was judged from video frames. For multi-round prompts, this included explanations of corrections (e.g., "Wrist guards added in 2nd round after slow-motion frame review"). The third component, Data Tables, was customized to modality—audio tables focused on auditory data (e.g., "40.0 score" with audio quotes), silent video tables on visual data (e.g., "large splash" with frame descriptions), and combined tables emphasized audio-visual correlations (e.g., "Audio 00:30 mentions score; video 00:28–00:32 shows scoreboard").

These results were categorized by video (VC_A to VC_D) and formatted as supplementary figures, clearly illustrating differences in modality-specific prompt design and the need for iterative refinement where applicable.

In summary, in this section, we employed modality-specific prompt design—distinguishing between single-round and multi-round approaches—to achieve precise data extraction from audio, visual, and combined stimuli.

VC_A_Diving_Audio

Q: Please extract the following information from this diving commentary audio: the name of the diving move (such as 'Forward Tumble Three and a Half Turns'); the judges' scores and the reasons for deductions; the key points emphasized in the commentary (such as 'Entry Angle' 'Body Posture'). Then summarize and present these data in a concise and orderly manner, Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references.and do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in english.

A: Initial Analysis: I first listened to the entire audio clip to get a general understanding of the content. The clip features two commentators discussing a diver's performance in a final.
Data Point Identification: I re-listened to the audio, pausing at specific timestamps to identify and extract the required data points:
Diving Move: I listened for the technical name of the dive, such as "Forward 2 1/2 Somersaults Tuck". The commentators did not explicitly name the dive. They did, however, mention key components like "over-rotations" and the "tuck" position.
Scores and Deductions: I listened for any numerical scores and the specific reasons given by the commentators for why points were lost. I identified a score of "40.0" and several reasons for deductions, including "over-rotations" and a poor "entry" because the diver "ran out of air" (i.e., opened up too early).
Key Commentary Points: I identified the main technical aspects the commentators focused on. These were the "over-rotation", the diver's use of "wrist support", the quality of the "tuck", and the critical "entry" into the water.
Data Structuring: I organized the extracted information chronologically based on when it appeared in the audio. Each distinct piece of information was assigned a row in the table.
Table Population: I created a table with the specified headings: Timestamp, Data Points, and Data References. I then filled each row with the corresponding timestamp, the identified data point (categorized for clarity, e.g., "Score", "Reason for Deduction"), and the exact quote from the audio as the Data Reference. This ensures the data is accurate and directly traceable to the source audio.

Here is the data table:

VC_A_Diving_Audio

TimeStamp	Data Points	Data References
00:00 - 00:02	**Key Point / Deduction Reason:** Over-rotation	"Another one of those over-rotations that we've seen quite a bit..."
00:06 - 00:08	**Key Point:** Wrist Support	"...using lots of support on those wrists..."
00:18 - 00:20	**Key Point / Deduction Reason:** Poor Entry	"...just losing it on the entry."
00:22 - 00:23	**Key Point:** Body Posture (Tuck)	"Nice and tight on the tuck."
00:25 - 00:28	**Deduction Reason:** Opened Up Too Early	"Just ran out of air really before hitting the water."
00:30 - 00:32	**Judges' Score:** 40.0	"...and there you can see the 40.0 scored."

VC_A_Diving_SilentVideo

Q: (1st round) Analyze this diving silent video, identify and record: the body posture of the athlete at the moment of diving (such as 'straight body' or 'bent body'); the type and height of the diving board / platform; the size and effect of the water splash (described as 'small', 'medium', or 'large'); the score board numbers shown in the picture, and other related information. Then summarize and present these data in a concise and orderly manner, output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references, and do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in english.

(2nd round) Was there no data point between 00:10 and 00:28? If there was one, please add it.

(3rd round) Please review this video again carefully to ensure that no data points are missed. Then output the data table and your chain of thought.

A: (3rd round) Systematic Review: I conducted a multi-pass review of the video. The first pass focused on the overall sequence of events. Subsequent passes focused on specific details: the environment, the athlete's equipment, the technical execution of the dive (broken down into take-off, flight, and entry), and the final results and reactions.

Initial Context (Coaches' Reaction): I began at 00:00. The video opens with the German coaches' reaction. Their body language and hand gestures, mimicking an entry, are not celebratory but analytical, immediately suggesting a flaw in the dive's finish. This is an important contextual data point.

Dive Execution & Equipment: I analyzed the multiple replay angles to extract key technical details.

- Platform: The replay at 00:04 clearly shows a high, fixed 10-meter platform.
- Equipment: The close-up slow-motion at 00:13 reveals the athlete is wearing black wrist guards. This is a significant detail, often used for support after injury or to protect against the high impact of platform diving.
- Body Posture: The replays confirm the athlete maintains a compact "tuck" position throughout the dive's rotations (00:07, 00:17).

Performance Analysis (Splash & Error): The primary performance indicators were the splash and the technical reason for it.

- Splash: The entry at 00:10 creates a very large splash, which is a clear visual indicator of a sub-optimal dive.
- Technical Flaw: The crucial slow-motion replay at 00:22 provides the definitive reason for the large splash. It clearly shows the athlete's body has rotated *past the vertical axis* upon water entry. This is a classic "over-rotated" entry, a significant deduction for judges.

VC_A_Diving_SilentVideo

A: Here is the data table:

TimeStamp	Data Points	Data References
00:00	**Coaches' Reaction:** Analytical / Disappointed	The German coaches are shown with analytical expressions, using hand gestures that mimic the dive's entry, indicating a focus on a flaw.
00:04	**Diving Platform:** 10-Meter Platform	The athlete is shown diving from a high, fixed, concrete platform.
00:07	**Body Posture:** Tuck Position	The athlete performs the dive with her body bent at the waist and knees, holding her shins in a compact tuck.
00:10	**Water Splash:** Large	A large and turbulent splash is created upon entry into the water, indicating a non-vertical entry.
00:13	**Athlete Equipment:** Wrist Guards	The slow-motion replay shows the athlete is wearing black guards on both wrists for support.
00:22	**Key Point / Deduction Reason:** Over-rotated Entry	The slow-motion replay shows the athlete's body entering the water angled forward, past the vertical line, confirming an over-rotation.

VC_A_Diving

Q: (1st round) By integrating the video footage and the commentary audio, comprehensively extract: the complete process of the athlete's entire set of movements (such as 'running-up → jumping → somersault → entering water'); the corresponding relationship between the commentary and the visuals (such as 'the specific time point of the commentary mentioning 'perfect entry into water' or the reasons for insufficient performance resulting in deductions'), and other related information. Then summarize and present these data in a concise and orderly manner, output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references, and most importantly, Is the content of the audio consistent with the content of the video? If there is no match, please inform me of the timestamp and method that have a higher degree of matching between the two. Do not omit any of useful datapoint. Explain your reasoning step by step and show me your thought process. Answer my question in English.

(2rd round) What I mean is, for example, the audio at 00:04 and the video at 00:10 are in sync. Please let me know and write it down in the data table (you can have other headings as well). Please modify the data table and tell me your thought process so that the data can be presented more intuitively.

A: (2rd round) My previous analyses focused on what was happening simultaneously in the audio and video. Your request prompted me to shift to a referential analysis model. The goal is no longer to just see what happens at 00:10, but to answer the question: "When the commentator says 'X', what is the best visual proof of 'X' in the video, and what is the relationship between them?"
This led to the following thought process:
Isolate the Commentary: First, I identified each distinct analytical comment made by the audio commentators and noted its exact timestamp.
Locate the Best Visual Proof: For each comment, I scanned the entire video to find the single best visual shot that illustrates or proves the commentator's point. This could be a live shot, a slow-motion replay, or a close-up. I noted the timestamp of this "best visual match."

VC_A_Diving

A: Analyze the Sync: I then analyzed the relationship between the audio comment and its visual proof. The key was to describe *how* they are linked. Is the commentary happening in real-time with the action? Is it a diagnosis that is later confirmed by a replay? Is it providing context for something we are seeing?

Design an Intuitive Table: To present this clearly, I designed a new table with headings that capture this referential link:

- Commentary (Audio): The specific quote from the announcer.
- Audio Timestamp: When the comment was spoken.
- Best Visual Match (Video): A description of the best visual evidence.
- Video Timestamp: When the best visual evidence appears.
- Analysis of Sync: A brief explanation of the relationship, clarifying if it's simultaneous, a diagnosis-then-proof, or contextual.

Here is the data table:

VC_A_Diving

Commentary (Audio)	Audio Timestamp	Best Visual Match (Video)	Video Timestamp	Analysis of Sync
"Another one of those over-rotations..."	00:00	The slow-motion replay clearly shows the athlete's body entering the water past the vertical axis.	00:22	**Diagnosis > Proof.** The commentator diagnoses the flaw in real-time as the dive ends. The broadcast provides definitive visual proof in a replay 22 seconds later.
"...using lots of support on those wrists..."	00:06	A close-up, slow-motion shot reveals the black guards the athlete is wearing on both wrists.	00:13	**Context for Visual Detail.** The commentary explains the *purpose* of the equipment that becomes clearly visible in the replay 7 seconds later.
"...hitting the water at about 30, 35 miles an hour."	00:11	The initial shot of the water entry shows a very large, forceful splash.	00:10	**Simultaneous Context.** The commentary provides scientific context for the forceful impact and large splash that is seen on screen at virtually the same time.
"...that wasn't one of her best, unfortunately, just losing it on the entry."	00:19	The slow-motion replay shows the over-rotated entry.	00:22	**Diagnosis > Proof.** Similar to the first point, this is a direct critique of the entry, which is best illustrated by the replay that follows shortly after.
"Nice and tight on the tuck."	00:22	The replay shows the athlete holding a very compact and well-formed tuck position in mid-air.	00:17 - 00:22	**Simultaneous Analysis.** This is a perfect sync where the commentator praises a specific technical element at the exact moment the video replay is highlighting it.
"Just ran out of air really before hitting the water."	00:25	The flawed, over-rotated entry.	00:22	**Explanation of a Prior Event.** The commentary provides the technical *reason* (losing control) for the flawed entry that was shown visually 3 seconds earlier.
"...and there you can see the 40.0 scored."	00:30	The official scoreboard graphic appears on screen, displaying "Round 4 40.00".	00:28 - 00:32	**Simultaneous Confirmation.** The commentary reads the score at the exact moment it is displayed for the audience, confirming the result.

VC_B_Finance_Audio

Q: Extract all numerical information from this financial explanation audio and label their meanings, such as dates, prices, names, etc. as key values and other valid data. Then summarize and present these data in a concise and orderly manner, Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings. Do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in English.

A: My process for extracting and presenting this information involves four key steps:
Initial Analysis & Keyword Identification: I first listened to the entire audio clip to understand its context—a financial report on Disney's theme park performance. I identified the types of numerical data to listen for: years (e.g., "2024"), currency amounts (e.g., "billion dollars"), percentages (e.g., "5%"), and fractions (e.g., "two-thirds").
Granular Data Extraction & Timestamping: I re-listened to the audio, pausing at each mention of a number. I recorded the exact timestamp, the numerical value, and the direct quote as a reference. This ensures accuracy and traceability.
Categorization and Inference: For each numerical data point, I assigned a clear Category to define its meaning (e.g., "Financial Year," "Revenue," "Announced Investment"). This fulfills the requirement to "label their meanings." Crucially, for the phrase "Last year" at timestamp 00:21, I made a logical inference. Since the report begins by referencing the year "2024," "last year" logically refers to 2023. This is a critical piece of inferred data.
Table Design and Population: To present the data most intuitively, I designed a table with five columns:

- Timestamp: To locate the data in the audio.
- Category: To clearly state the meaning of the number.
- Value: To isolate the specific numerical figure.
- Entity / Segment: To specify what the number refers to (e.g., Disney's Theme Park Segment, Disney Company).
- Data Reference: To provide the exact quote from the audio as evidence.

Here is the data table:

VC_B_Finance_Audio

Timestamp	Category	Value	Entity / Segment	Data Reference
00:00	Financial Year	2024	Disney's Theme Park Segment	"In **2024**..."
00:03	Revenue	>$34 billion	Disney's Theme Park Segment	"...raked in a record revenue of more than **34 billion dollars**."
00:06	Revenue Growth	5%	Disney's Theme Park Segment	"...a **5%** climb from the year prior."
00:13	Profit Share	~⅔ (approx. 66.7%)	Disney Company (Experiences vs. other segments)	"...so it's about **two-thirds** of the company's operating profits..."
00:20	Announced Investment	$30 billion	Disney (Experiences Division)	"...its announced **30 billion dollar** investment."
00:21	Financial Year	2023	Central Florida Tourism Oversight District	"**Last year**..." (Inferred from being the year prior to 2024)
00:24	Deal Quantity	A historic deal (1)	Walt Disney World & Oversight District	"...approved a **historic deal** with Walt Disney World."

VC_B_Finance_SilentVideo

Q: (1st round) Please analyze the images in the silent video, extract all the relevant and data from the images except for the large visualized titles and numbers displayed on the screen(such as the scene and other details), then summarize and present these data in a concise and orderly manner, Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings. Do not omit any of them.Explain your reasoning step by step and show me your chain of thought. Answer my question in English.

(2nd round) please ignore the subtitles above or at the bottom. They are not considered. Output the chain of thought and rewrite the data table.

(3rd round) Also ignore the CNBC logo. Please just analyze the video frame

A: Step 1: Deconstructing the User's Cumulative Instructions

My first action is to synthesize all the instructions I have received throughout our conversation into a single, definitive set of rules.

Core Task: Analyze the visual content of the provided silent video.

Primary Output: Present the analysis in a structured data table with specific columns (Timestamp, Key Visual Information, Visual Context / Location, Additional Details / Observations).

Rule #1 (from initial request): Exclude large, graphically highlighted titles and numbers. This refers to the prominent yellow data box (e.g., "$34b").

Rule #2 (from second request): Exclude all text overlays and subtitles. This broadens Rule #1 to include the smaller text captions at the top and bottom of the screen (e.g., "our guests in 2024," "a 5% climb from the year").

Rule #3 (from third request): Exclude the CNBC logo. This applies to the persistent network logo in the corner and any instance of the logo within the scene itself (like on the mug), as it is part of the broadcast's branding, not the raw scene's content.

VC_B_Finance_SilentVideo

A: The final, combined goal is to act as a pure visual analyst, describing *only* what is physically present in the scene as if I were looking at raw, unedited footage. The only text I am permitted to analyze is text that is physically part of the environment itself (e.g., a sign on a building).

Step 2: Establishing the Analysis Framework for Each Video Segment

For each distinct scene in the video, I will follow this mental checklist:

Identify the Scene: Watch the segment and get a general sense of what is happening. Where is it? Who is there? What are they doing?

Apply the "Exclusion Filter": Mentally erase all forbidden elements based on the rules from Step 1. This means I actively ignore the corner logo and all text that is not physically part of the background.

Extract Core Visual Data: With the filtered view in mind, I describe the essential elements:

- Subjects: People, crowds, a person on a screen.
- Setting: The architecture, landscape, and specific environmental features (e.g., greenhouse, spaceship, entrance arch).
- In-Scene Text: I carefully read any text that is part of a physical sign (e.g., "The Majestic Theater," "Walt Disney World"). This is crucial data as it helps identify the location.

Identify Contextual Details: I observe secondary details that add depth to the description. This includes the time of day (daylight, dusk, sunset), lighting conditions (natural vs. artificial), and indicators of activity (crowd density, moving vehicles).

Determine Location: Based on the combination of architecture, in-scene text, and iconic landmarks, I identify the specific, real-world location (e.g., a particular theme park or attraction).

Populate the Table: I translate these observations into concise, factual statements and enter them into the appropriate columns of the data table. I will process each scene chronologically.

Step 3: Executing the Analysis, Segment by Segment

VC_B_Finance_SilentVideo

A: Segment @ 00:00:

Observation: A street scene with many people. Stylized buildings. It's getting dark.

Filtering: I ignore the "our guests in 2024" caption and the CNBC logo.

Extraction: I see crowds of people. I read the marquee on the yellow building: "THE MAJESTIC THEATER." I read the sign on the blue building: "CARTHAY CIRCLE CLUB."

Context: The lighting is artificial and warm, indicating dusk.

Location ID: Hollywood-style architecture + Carthay Circle reference = Disney's Hollywood Studios.

Table Entry: Fill the first row with this information.

Segment @ 00:02:

Observation: People on a boat ride, inside a large structure with plants.

Filtering: I ignore the "$34b" box, the "a 5% climb" text, the "Source: Disney" text, and the CNBC logo.

Extraction: The subjects are guests on a green boat. The environment is a greenhouse with a visible geodesic dome ceiling. The ride is on water.

Context: The setting is indoors and well-lit.

Location ID: Greenhouse boat ride + geodesic dome = "Living with the Land" at Epcot.

Table Entry: Fill the second row.

Segment @ 00:07:

Observation: A crowded outdoor area with unique, otherworldly architecture and a large, famous spaceship.

Filtering: I ignore the "park company that also makes" text and the CNBC logo.

Extraction: The key landmark is a life-sized replica of the Millennium Falcon. The architecture is distinctive rockwork. The area is filled with many guests, including families with strollers.

Context: It is daytime with bright sunlight. The high crowd level indicates a popular location.

Location ID: Millennium Falcon + specific architecture = Star Wars: Galaxy's Edge.

Table Entry: Fill the third row.

VC_B_Finance_SilentVideo

A: Segment @ 00:11:

Observation: A close-up on a laptop showing a man in a video call. There is a mug and a plant on the desk.

Filtering: I ignore the "profits are coming from" text and the main CNBC logo. Crucially, I also ignore the logo on the mug, treating it as just a white mug to adhere strictly to Rule #3.

Extraction: The scene shows a man on a laptop screen, a white mug, and a potted plant.

Context: The setup implies a remote interview or video conference.

Location ID: The setting is generic and identified as a "Remote Interview / Home Office setting."

Table Entry: Fill the fourth row.

Segment @ 00:18:

Observation: A very large archway over a road at sunset.

Filtering: I ignore the "Tourism Oversight District" text and the CNBC logo.

Extraction: I read the physical text on the arch: "Walt Disney World" and "THE MOST MAGICAL PLACE ON EARTH." I see a crest with the number "50" on the right pillar and smaller directional signs below.

Context: The beautiful, colorful sky clearly indicates sunset.

Location ID: The text on the arch explicitly identifies it as the entrance to Walt Disney World.

Table Entry: Fill the fifth and final row.

Step 4: Final Review

I perform a final read-through of the generated table to ensure it is 100% compliant with the synthesized rules from Step 1. I check that no forbidden text or logos are mentioned, and that the descriptions are based solely on the permissible visual evidence. The chain of thought is now complete and the final table is ready.

Here is the data table:

VC_B_Finance_SilentVideo

Timestamp	Key Visual Information	Visual Context / Location	Additional Details / Observations
00:00 - 00:02	Crowds of guests walk down a street lined with stylized buildings. A prominent marquee for "THE MAJESTIC THEATER" is visible, along with a blue building signed "CARTHAY CIRCLE CLUB".	Disney's Hollywood Studios (re-creation of Hollywood Boulevard)	The scene takes place at dusk, with artificial lights illuminating the buildings and street.
00:02 - 00:07	Guests are seated in a green boat on an indoor water ride that moves through a greenhouse. Various plants and a geodesic dome ceiling are visible.	"Living with the Land" attraction, The Land Pavilion, Epcot	The ride is well-lit and appears to be in motion. The setting combines natural-looking flora with a large architectural structure.
00:07 - 00:11	A wide shot of numerous guests in a highly themed area with unique rockwork architecture. A life-sized replica of the Millennium Falcon spaceship is docked in the background.	Star Wars: Galaxy's Edge (likely at Disney's Hollywood Studios)	The scene is during the day with bright, natural light. Many guests and strollers are visible, indicating a busy park environment.
00:11 - 00:18	A man is shown on a laptop screen during a video interview. Next to the laptop is a white mug and a green potted plant.	Remote Interview / Home Office setting	The setting appears to be a remote interview from a home or office, focused on the individual speaking.
00:18 - 00:26	The illuminated main entrance archway to a resort. The sign reads "Walt Disney World" and "THE MOST MAGICAL PLACE ON EARTH".	Main vehicle entrance to Walt Disney World, Orlando, Florida	The scene is at sunset, with a colorful sky. A "50" crest is on the right pillar. Directional signs for various areas are visible below the main arch.

VC_B_Finance

Q: (1st round) By integrating audio and video information, extract: the corresponding relationship between the numbers mentioned in the commentary and the images; the causal logic of financial events; the rapidly flashing digital information in the images, and other such valuable information. Emphasize the cross-validation of audio and video information, for example, by correlating the logical structure of the commentary with the data in the images. Then summarize and present these data in a concise and orderly manner. Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings, and most importantly, Is the content of the audio consistent with the content of the video? If there is no match, please inform me of the timestamp and method that have a higher degree of matching between the two, for example, if the audio at 00:04 and the video at 00:10 are in sync, please let me know and write it down in the data table. Do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in English.

(2nd round) In the data table, the numerical data should be separated out and placed in a separate column, so as to make the data in the table more intuitive.

A: Deconstruct the Revised Request: The user's core instruction is to make the table more intuitive by isolating all numerical data into a new, separate column. This means I need to modify the table structure from my previous analysis.
Redefine Table Structure: I determined the most logical way to implement this change was to create a new column specifically for numbers. The original Data Points column, which mixed narrative descriptions with figures, will be split into two more focused columns:

Key Statement / Event: This column will describe the main narrative point or event from the audio (e.g., "Disney's theme park segment achieved record revenue").

Numerical Data: This new column will contain only the specific numbers, percentages, or fractions mentioned in the audio (e.g., "$34 billion," "5%").

Reprocess and Repopulate: I re-examined each segment of the video, applying this new structure:

For each segment, I first identified the core statement or event being described and placed it in the Key Statement / Event column.

A: Next, I extracted the associated numerical figure(s) from the audio and placed them in the Numerical Data column.

The Data References and Causal Logic column was then updated to explain how both the statement and the numerical data are supported by the video's visuals (on-screen graphics, contextual scenes) and to clarify the underlying business logic.

The final column assessing audio-video consistency remained unchanged in its function.

Final Review: I reviewed the newly formatted table to ensure it directly addresses the user's request. The separation of text-based statements from numerical figures makes the key financial data much easier to identify at a glance, fulfilling the goal of a more "intuitive" layout.

Here is the data table:

VC_B_Finance

Timestamp	Key Statement / Event	Numerical Data	Data References and Causal Logic
00:00 - 00:07	Disney's theme park segment achieved a record revenue in 2024, showing significant growth over the previous year.	• $34 billion (revenue) • 5% (year-over-year growth)	Data Reference: At 00:03, a graphic explicitly displays "$34b" and "Disney theme park revenue in 2024," directly validating the audio. Causal Logic: The visuals of busy parks (Hollywood Studios, Epcot) illustrate the high guest traffic that drives this record revenue.
00:07 - 00:18	The "experiences" division (theme parks) is the primary source of The Walt Disney Company's operating profits.	Two-thirds (2/3) (share of operating profits)	Data Reference: The analyst states this fraction while footage of a popular "experience" (Star Wars: Galaxy's Edge) is shown. Causal Logic: This highlights that despite being a large media company, Disney's financial core relies heavily on the profitability of its physical parks.
00:18 - 00:26	The Central Florida Tourism Oversight District approved a major, long-term investment deal with Walt Disney World.	$30 billion (investment plan)	Data Reference: The narrator references a previously announced investment figure. At 00:23, a text overlay "Tourism Oversight District" appears, validating the entity named in the audio. Causal Logic: To ensure future growth and maintain its profit leadership, Disney is securing approval for massive capital investments with its local governing body.

VC_C_Finance_Audio

Q: (1ST round) Extract all numerical information from this financial explanation audio and label their meanings, such as dates, prices, names, etc. as key values and other valid data. Then summarize and present these data in a concise and orderly manner, Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings. Do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in English.

(2nd round) Please modify the headings, let them be Timestamp, Category, Value, Entity / Segment, Data Reference (and other heading if it is necessary)

A: Deconstruction of the Request: My primary goal is to extract all numerical information from the audio clip. The request specifies that these numbers should be labeled with their meanings and presented in a table with three columns: timestamp, data points, and a column for context/references.

Data Extraction Strategy: I will listen to the audio multiple times.

- The first pass will be to understand the overall narrative: The audio contrasts the advertised top speed of Amtrak's Acela with its actual average speed and then compares it to China's high-speed rail system.
- The second pass will be a meticulous data extraction process. I will pause at each mention of a number and note the exact timestamp, the numerical value, and the unit (e.g., mph, miles).

Data Labeling and Structuring: For the data points column, I will not just list the number. As requested, I will create a clear "key-value" pair (e.g., "Top Speed: 150 mph") to explicitly label what each number represents. This makes the data immediately understandable.

Contextual Analysis for the Final Column: The "data references and other useful and necessary headings" column is crucial for adding value. For each piece of data, I will provide:

- Direct Context: A brief explanation of what the number describes, as stated in the audio.

A: Relational Context: How the number relates to other data points mentioned in the clip. For example, contrasting the Acela's top speed with its average speed, or comparing the distance traveled at top speed to the total length of the track.

Significance: The underlying meaning or implication of the data. This involves explaining *why* the narrator presents these numbers in this order—to build an argument about the limitations of US high-speed rail.

Table Population and Final Review: I will populate the table row by row, ensuring each entry is accurate and the context is clearly articulated. The final step is to review the entire table to ensure it is concise, orderly, and directly answers all aspects of the user's prompt.

Here is the data table:

Timestamp	**Category**	**Value**	**Entity / Segment**	**Data Reference**
00:02	Top Speed	150 mph	Acela	The maximum advertised speed the train can achieve, presented as the service's headline capability.
00:04	Distance (at Top Speed)	34 miles	Acela	The limited portion of the total track where the Acela can actually achieve its top speed.
00:06	Total Distance / Track Span	457 miles	Acela	The total length of the Acela line, used to show that the high-speed portion is only a small fraction (~7.4%) of the whole.
00:10	Average Speed	~65 mph	Acela (New York to Boston segment)	The practical, real-world travel speed on a key route, highlighting the significant difference between top speed and actual performance.
00:18	Regular Operating Speed	>200 mph	China High-Speed Rail System	An international benchmark used to create a "stark contrast" with the Acela's performance and highlight the US system's limitations.

VC_C_Finance_SilentVideo

Q: Please analyze the images in the silent video, extract all the relevant and data from the images except for the large visualized titles and numbers displayed on the screen(such as the scene and other details), then summarize and present these data in a concise and orderly manner, Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings. Do not omit any of them.Explain your reasoning step by step and show me your chain of thought. Answer my question in English. please ignore the subtitles above or at the bottom. They are not considered. Output the chain of thought and rewrite the data table. Also ignore the CNBC logo. Please just analyze the video frame

A: Step 1: Synthesizing All User Instructions into a Final Rule Set
First, I will consolidate all the instructions from our conversation into a single, definitive set of rules for my analysis.
Initial Request: Analyze the video, but exclude large visualized titles and numbers (e.g., "65 mph").
Second Request: Ignore *all* subtitles, both at the top and bottom. This broadens the exclusion rule.
Third Request: Ignore the CNBC logo. This completes the filtering of all non-diegetic (not part of the scene's world) elements.
Final Goal: The cumulative instruction is to perform a pure visual analysis of the raw video footage. I must describe only what is physically present within the scene's environment. The only text I am permitted to analyze is text that is an integral part of the scene itself, such as a physical sign.
Step 2: Establishing the Visual Analysis Framework
For each distinct scene, I will follow a strict, systematic process:
Scene Identification: I will watch the segment to understand the location, subjects, and actions.
Apply the "Exclusion Filter": I will actively and mentally erase all forbidden elements defined in Step 1. This means the corner logo and all superimposed captions/numbers are completely disregarded. I am looking at the scene as if it were raw, unedited footage.

VC_C_Finance_SilentVideo

A: Core Visual Data Extraction: With the filtered view, I will identify and describe the essential elements:

- Subjects: A silver and red train, an individual on a platform.
- Environment: An interior train car, a steel railway bridge, an outdoor train station platform, a rural landscape with a viaduct.
- Diegetic Text (In-Scene Text): I will carefully read any text that is part of a physical object within the scene. In this video, I will look for signs on the train or in the station.
- Action/Movement: I will note the motion of the train (fast motion blur, pulling into a station) and the perspective of the camera (interior, trackside, aerial).

Contextual Detail Identification: I will observe secondary details like the type of landscape (industrial, rural), infrastructure (catenary wires, steel bridge construction), and weather (daylight, overcast sky).

Table Population: I will translate these filtered observations into concise, factual statements and populate the data table. The data points column will be the primary description, and the data references column will provide additional context and observations.

Step 3: Executing the Analysis, Segment by Segment

Segment @ 00:00:

- Filtering: Ignore subtitles and the CNBC logo.
- Visuals: Interior view of a train car, looking at a door with a window. The exterior scenery is blurred due to high speed. A red "EXIT" sign is visible above the door.
- Table Entry: Populate the first row.

Segment @ 00:02:

- Filtering: Ignore subtitles and the CNBC logo.
- Visuals: A low-angle, trackside view of a silver train with a red stripe crossing a steel railway bridge. The train is moving very fast, causing motion blur.
- Table Entry: Populate the second row.

A: Segment @ 00:06:

Filtering: Ignore the large "65 mph" graphic and other subtitles. Ignore the CNBC logo.

Visuals: A view from a train station platform. Another train is pulling into the station on an adjacent track. A person is visible in the distance on the platform. A blue sign with "Caution Low Clearance" is visible. The number "2004" can be seen on the front of the arriving train.

Table Entry: Populate the third row.

Segment @ 00:12:

Filtering: Ignore the large "200 mph" graphic and other subtitles. Ignore the CNBC logo.

Visuals: An aerial drone shot of a white high-speed train traveling on a long, elevated track (a viaduct) across a lush, green rural landscape. The landscape includes fields, patches of water, and distant mountains.

Table Entry: Populate the fourth row.

Step 4: Final Review

I will perform a final check of the completed table to ensure it is 100% compliant with the synthesized rule set from Step 1. All descriptions are based purely on the permitted visual evidence, and all non-diegetic elements have been successfully ignored.

Here is the data table:

VC_C_Finance_SilentVideo

Timestamp	Data Points	Data References and Other Useful and Necessary Headings
00:00 - 00:02	An interior view from a moving train car, looking at a door with a window. A red "EXIT" sign is above the door.	**Observation:** The scenery seen through the window is heavily blurred, indicating the train is traveling at high speed.
00:02 - 00:06	A low-angle, trackside view of a silver train with a red stripe crossing a steel railway bridge.	**Observation:** The train is in fast motion, causing significant motion blur. The shot emphasizes the train's speed and the industrial nature of the bridge.
00:06 - 00:12	A view from an outdoor train station platform. A silver train with the number "2004" on its front is seen arriving on an adjacent track.	**Observation:** A blue sign reading "Caution Low Clearance" is visible hanging from the station's canopy. Catenary wires are visible above the tracks.
00:12 - 00:22	An aerial shot of a white high-speed train traveling on an elevated track (viaduct) over a vast, rural landscape.	**Observation:** The landscape is green and lush, featuring fields, bodies of water, and distant, low mountains. The long, straight viaduct is a prominent piece of infrastructure.

VC_C_Finance

Q: By integrating audio and video information, extract: the corresponding relationship between the numbers mentioned in the commentary and the images; the causal logic of financial events; the rapidly flashing digital information in the images, and other such valuable information. Emphasize the cross-validation of audio and video information, for example, by correlating the logical structure of the commentary with the data in the images. Then summarize and present these data in a concise and orderly manner. Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings, and most importantly, Is the content of the audio consistent with the content of the video? If there is no match, please inform me of the timestamp and method that have a higher degree of matching between the two, for example, if the audio at 00:04 and the video at 00:10 are in sync, please let me know and write it down in the data table. Do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in English.
In the data table, the numerical data should be separated out and placed in a separate column, so as to make the data in the table more intuitive.

A: Step 1: Deconstructing the Cumulative Request
My primary goal is to synthesize all instructions into a single, cohesive analysis plan.
Core Task: Integrate audio and video information to extract data, relationships, and causal logic.
Audio Analysis: Extract all numerical data points (speeds, distances) and the narrative context.
Video Analysis: Analyze the visual scenes, focusing on how they illustrate or correspond to the audio.
Cross-Validation: Directly link audio statements to visual elements. This is key. For example, when the audio mentions a speed, I must check if a graphic displaying that speed appears on screen.
Causal Logic: Explain *why* the information is presented in a specific sequence. For example, explain that the limited high-speed track *causes* the average speed to be much lower than the top speed.
Consistency Check: For each segment, determine if the audio and video are synchronized and logically connected.

VC_C_Finance

A: Table Structure: The output must be a data table with specific, well-defined headings. The most important structural change requested is to create a separate column exclusively for Numerical Data to make the information more intuitive. The final headings will be: Timestamp, Key Statement / Event, Numerical Data, Data References and Causal Logic, and Is the content of the audio consistent with the content of the video?.

Step 2: Sequential Analysis and Cross-Referencing

I will process the video chronologically, breaking it down into segments based on the audio narrative and corresponding visuals.

Segment 1 (00:00 - 00:07): Acela Top Speed vs. Reality

- Audio Data: Acela can reach "150 mph," but only for "34 miles" of its "457-mile" span.
- Visual Data: The video shows fast-moving footage of the Acela train, first from the interior and then a trackside view of it crossing a bridge. This establishes the subject matter.
- Cross-Reference: The audio provides the numbers, and the video provides the visual identity of the train being discussed. There are no on-screen graphics for these specific numbers.
- Causal Logic: The low distance at top speed (34 miles) relative to the total span (457 miles) is the *cause* of the discrepancy between advertised potential and real-world performance.
- Consistency: The high-speed visuals are consistent with the discussion of the train's top speed capabilities.

Segment 2 (00:08 - 00:12): Acela Average Speed

- Audio Data: The average speed between New York and Boston is "about 65 mph."
- Visual Data: The video cuts to an Acela train pulling into a station. A large graphic "65 mph" is superimposed over the scene.
- Cross-Reference: This is a perfect match. The audio states the average speed, and a large, clear graphic simultaneously validates that exact number.
- Causal Logic: This average speed is the logical *effect* of the limited high-speed track mentioned in the previous segment.
- Consistency: The visual of a train at a station (a place of slowing down) provides good context for a discussion about average speed, which includes stops and slower sections. The direct graphic overlay makes the consistency perfect.

A: Segment 3 (00:12 - 00:19): International Comparison

Audio Data: The narrator introduces a "stark contrast" with China's high-speed rail, which regularly travels at "over 200 mph."

Visual Data: The scene shifts dramatically to a sweeping aerial shot of a white high-speed train in China traveling on a long viaduct. A large graphic "200 mph" is superimposed over this scene.

Cross-Reference: Another perfect match. The audio introduces the comparison subject (China) and its speed, and the visuals immediately confirm both with a new train model, a different landscape, and a matching numerical graphic.

Causal Logic: This comparison serves to highlight the performance gap and frame the Acela's speed as insufficient by international standards. The dedicated infrastructure (the viaduct) shown is the visual *cause* for China's high, regular speeds.

Consistency: The visual shift is deliberate and highly effective. The contrast between the older-looking US station and the modern, sweeping landscape in China visually reinforces the "stark contrast" mentioned in the audio.

Step 3: Populating the Final Table

I will now organize all these findings into the structured table, ensuring each column is populated accurately according to the refined rules. The separation of numerical data will be prioritized to meet the user's final request.

Here is the data table:

VC_C_Finance

Timestamp	Data	Category	Entity / Segment	Data Reference and Causal Logic
00:02	150 mph	Top Speed	Acela	The maximum advertised speed the train can achieve, presented as the service's headline capability.
00:04	34 miles	Distance at Top Speed	Acela	The limited portion of the total track where the Acela can actually achieve its top speed. This is the *cause* of its lower average speed.
00:06	457 miles	Total Track Span	Acela	The total length of the Acela line, used as a reference to show how small the high-speed portion is.
00:10	~65 mph	Average Speed	Acela (New York to Boston)	The practical, real-world speed. This is validated by a large **"65 mph"** graphic on-screen and is the *effect* of the limited high-speed track.
00:18	>200 mph	Regular Operating Speed	China High-Speed Rail	An international benchmark used for contrast. This is validated by a large **"200 mph"** graphic on-screen, accompanied by visuals of modern infrastructure.

VC_D_Finance_Audio

Q: Extract all numerical information from this financial explanation audio and label their meanings, such as dates, prices, names, etc. as key values and other valid data. Then summarize and present these data in a concise and orderly manner, Output in the form of a data table. The table's headings are Timestamp, Category, data points (the specific data points are determined by the audio), Entity / Segment, Data Reference (and other heading if it is necessary). Do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in English.

A: Initial Audio Analysis: The first step was to carefully listen to the audio clip and simultaneously read the transcript to identify every instance of a numerical value. This included dollar amounts, percentages, and simple counts.
Contextual Data Extraction: For each number identified, I analyzed the surrounding words to understand its meaning and context. This is crucial for accurate labeling.

When the speaker says "...cost around ** 700million∗∗...",
700 millionand its context as thecostof a proposed program (making buses free`).
For "...about $5 or $6 billion...", I noted this was a range and also a cost associated with universal child care.
For "...a $113 billion municipal budget...", I extracted the value and linked it to the Municipal Budget as an existing financial figure.
The mention of "...two major revenue streams..." was identified as a simple count related to the structure of the Proposed Plan.
The phrase "...that raises $5 billion..." was clearly linked to Revenue Generation from the proposed corporate tax increase.
The mention of the "...top 1% of New Yorkers..." was identified as a Population Segment targeted by the plan.
The qualifier "...who make a million dollars or more a year..." was extracted as the specific Income Threshold defining that segment.
Finally, "...by 2%..." was clearly identified as the proposed Tax Increase for that income group.

A: Categorization and Column Assignment: Once the data points and their contexts were understood, I assigned them to the specific columns requested in the prompt:

- Timestamp: The start and end time in the audio where the information was spoken.
- Category: A high-level classification was created for each data point (e.g., "Cost / Proposed Spending", "Budget Size", "Revenue Generation", "Tax Increase"). This helps group related information.
- Data Points: The specific numerical value itself was placed here (e.g., $113 billion, 2%).
- Entity / Segment: This column specifies *what* the number is describing (e.g., "Universal child care", "New York City Municipal Budget", "Top 1%").
- Data Reference: This column provides crucial context about the nature of the data—whether it is part of a Proposed Plan or an Existing Context/Budget used for comparison.

Table Population and Refinement: With all the data extracted and categorized, I populated the table in a clear and orderly manner. I reviewed the table against the audio transcript one final time to ensure no numerical data was omitted and that all labels were accurate and concise. This final review ensures the output is both complete and easy to understand.

Here is the data table:

VC_D_Finance_Audio

Timestamp	Category	Data Points	Entity / Segment	Data Reference
00:05 - 00:06	Cost / Proposed Spending	~$700 million	Making buses free	Proposed Plan
00:07 - 00:08	Cost / Proposed Spending	~$5 billion to $6 billion	Universal child care	Proposed Plan
00:12 - 00:13	Budget Size	$113 billion	New York City Municipal Budget	Existing Context
00:15 - 00:16	Budget Size	$252 billion	New York State Budget	Existing Context
00:22 - 00:23	Plan Structure	2	Major revenue streams	Proposed Plan
00:26 - 00:27	Revenue Generation	$5 billion	Increased top corporate tax rate	Proposed Plan
00:28 - 00:30	Population Segment	Top 1%	High-income earners in New York	Tax Proposal Target
00:30 - 00:31	Income Threshold	$1 million+ per year	The top 1% of New Yorkers	Tax Proposal Target
00:32 - 00:32	Tax Increase	2%	Income tax for the top 1%	Proposed Plan

VC_D_Finance_SilentVideo

Q: Please analyze the images in the silent video, extract all the relevant and data from the images except for the large visualized titles and numbers displayed on the screen(such as the scene and other details), then summarize and present these data in a concise and orderly manner, Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings. Do not omit any of them.Explain your reasoning step by step and show me your chain of thought. Answer my question in English. please ignore the subtitles above or at the bottom. They are not considered. Output the chain of thought and rewrite the data table. Also ignore the CNN logo. Please just analyze the video frame

A: Initial Frame Analysis: The first step was to examine the video frames to identify sources of non-spoken data. The most prominent source is the lower-third on-screen graphic, also known as a chyron, which remains static throughout the clip.
Instruction Interpretation: I carefully considered the instructions to "extract all the relevant and data from the images except for the large visualized titles and numbers displayed on the screen" and to "ignore the subtitles."

- The text "BREAKING NEWS" was identified as a large, visualized title and was therefore excluded as requested.
- The CNN logo was also ignored as per the instructions.
- The lower-third graphic is not a subtitle (a transcription of dialogue) but a data overlay providing context about the segment, host, and guest. Therefore, it was the primary target for data extraction.

Data Deconstruction and Categorization: I broke down the text within the chyron into distinct, meaningful pieces of information and assigned them to logical categories.

- Program Information: The text "LIVE ON OUTFRONT" and the name "ERIN BURNETT OUTFRONT" clearly identifies the program.
- Host Information: "Erin Burnett" is identified as the host.
- Guest Information: The guest's name, "Zohran Mamdani," is listed twice. His political party, "(D)" for Democrat, and his role, "New York City Mayoral Candidate," are explicitly stated. An additional piece of context, "Who upset Cuomo in NYC race," is provided as a descriptor.

A: Table Structure and Population: Based on the extracted data, I created the specified table structure.

- Timestamp: Since the graphic is visible and unchanged for the entire duration of the clip, the timestamp 00:00 - 00:37 was used for all entries.
- Category: I used descriptive categories like "Program Information," "Host Information," and specific sub-categories for the guest to organize the data logically.
- Data Points: The specific text extracted from the chyron was placed in this column. For clarity, I separated the guest's name, descriptor, political affiliation, and title into different rows.
- Data Reference: The source for all extracted information was consistently the "On-Screen Graphic (Chyron)."

Final Review: The table was reviewed against the video frames one last time to ensure all relevant data points from the chyron were accurately captured and correctly categorized, adhering strictly to the user's instructions.

Here is the data table:

VC_D_Finance_SilentVideo

Timestamp	Category	Data Points	Data Reference	Timestamp
00:00 - 00:37	Program Information	LIVE ON OUTFRONT	On-Screen Graphic (Chyron)	00:00 - 00:37
00:00 - 00:37	Host Information	Erin Burnett	On-Screen Graphic (Chyron)	00:00 - 00:37
00:00 - 00:37	Guest Name	Zohran Mamdani	On-Screen Graphic (Chyron)	00:00 - 00:37
00:00 - 00:37	Guest Descriptor	Who upset Cuomo in NYC race	On-Screen Graphic (Chyron)	00:00 - 00:37
00:00 - 00:37	Guest Political Affiliation	(D) - Democrat	On-Screen Graphic (Chyron)	00:00 - 00:37
00:00 - 00:37	Guest Title / Role	New York City Mayoral Candidate	On-Screen Graphic (Chyron)	00:00 - 00:37

VC_D_Finance

Q: By integrating audio and video information, extract: the corresponding relationship between the numbers mentioned in the commentary and the images; the causal logic of financial events; the rapidly flashing digital information in the images, and other such valuable information. Emphasize the cross-validation of audio and video information, for example, by correlating the logical structure of the commentary with the data in the images. Then summarize and present these data in a concise and orderly manner. Output in the form of a data table. The table's headings are timestamp, data points (the specific data points are determined by the audio), and data references and other useful and necessary headings, and most importantly, Is the content of the audio consistent with the content of the video? If there is no match, please inform me of the timestamp and method that have a higher degree of matching between the two, for example, if the audio at 00:04 and the video at 00:10 are in sync, please let me know and write it down in the data table. Do not omit any of them. Explain your reasoning step by step and show me your chain of thought. Answer my question in English.
In the data table, the numerical data should be separated out and placed in a separate column, so as to make the data in the table more intuitive.

A: Initial Analysis: The video displays a CNN interview. The on-screen text (chyron) identifies the guest as Zohran Mamdani, a New York City Mayoral Candidate. The audio consists of Mamdani discussing his financial plans for the city, including costs for social programs and proposed revenue streams from tax increases. The interviewer is Erin Burnett.
Information Extraction (Audio): I will listen to the audio and pull out all specific numerical figures and the policies they relate to.

- Cost of "making buses free": This is likely part of the "700 million dollars" figure.
- Cost of "universal childcare": This is likely the "5 or 6 billion dollars" figure.
- NYC Municipal Budget: "$113 billion dollar".
- NY State Budget: "$252 billion dollar".
- Revenue from corporate tax increase: "$5 billion dollars".
- Target group for income tax increase: "top one percent".
- Income threshold for the tax increase: "million dollars or more a year".
- Proposed income tax increase rate: "two percent".

VC_D_Finance

A: Information Extraction (Video): I will examine the visual information, primarily the chyron.

- Program Name: ERIN BURNETT OUTFRONT.
- Breaking News Headline: "LIVE ON OUTFRONT: ZOHRAN MAMDANI, WHO UPSET CUOMO IN NYC RACE".
- Guest Identification: "Zohran Mamdani | (D) New York City Mayoral Candidate".
- The video shows Mamdani speaking and gesturing, and the interviewer listening and taking notes. The background is a digital representation of the New York City skyline. There are no graphics or text displaying the financial data Mamdani mentions.

Cross-Validation and Consistency Check:

- The core question is whether the audio and video content are consistent. The video identifies the speaker and the context (a political candidate in an interview). The audio consists of that speaker discussing his political platform. In this regard, the audio and video are perfectly consistent and complementary.
- However, the *specific numerical data* mentioned in the audio is not visually represented in the video. The chyron provides context but does not repeat or display the figures. Therefore, there is no direct data match or contradiction.
- The video provides the "who," and the audio provides the "what."

Table Construction: I will structure the data into a table as requested, with columns for timestamp, topic, numerical data, unit/context, source, and consistency. I will list each piece of data from both the audio and video to create a comprehensive log.

Here is the data table:

VC_D_Finance

Timestamp	Topic/Concept	Numerical Data	Unit/Context	Data Source (Audio/Video)
00:00 - 00:37	Speaker Identification	Zohran Mamdani	Name, (D) NYC Mayoral Candidate	Video (Chyron)
00:04 - 00:07	Cost of Free Buses & Universal Childcare	700	Million Dollars	Audio
00:07 - 00:08	Cost of Free Buses & Universal Childcare	5-6	Billion Dollars	Audio
00:12 - 00:14	NYC Municipal Budget	113	Billion Dollars	Audio
00:15 - 00:17	NY State Budget	252	Billion Dollars	Audio
00:23 - 00:28	Proposed Revenue from Corporate Tax Increase	5	Billion Dollars	Audio
00:28 - 00:30	Target for Income Tax Increase	1	Percent (Top earners)	Audio
00:30 - 00:31	Income Threshold for Tax Increase	1	Million Dollars (or more a year)	Audio
00:31 - 00:32	Proposed Income Tax Increase	2	Percent	Audio