A. Benchmark

We choose papers from 5 fields:Mathematics, Molecular Biology, Geology, Machine Learning, and Climate Science. The paper list is shown below:

Mathematics: *"Solutions of irreflexive relations"* by Richardson M. (1953), *"On the existence of hermitian-yang-mills connections in stable vector bundles"* by Uhlenbeck,K. and Yau,S. T. (1985), *"Higher algebraic structures and quantization"* by Freed D S. (1994), *" Understanding qualitative calculus:A structural synthesis of learning research"* by Stroup W M. (2002), *"Oscillation theorems in the complex domain"* by Hille E. (1922), *"Fourier integral operators."* by Hörmander L. (1971), *"Solutions of irreflexive relations"* by Richardson M. (1953), *"Internal set theory:a new approach to nonstandard analysis"* by Nelson E. (1977), *"Topological vector spaces of continuous functions"* by Nachbin L. (1954), *"Oscillation theory"* by Kreith K. (1973)

Molecular Biology: *"Molecular structure of nucleic acids:a structure for deoxyribose nucleic acid"* by Watson J D, Crick F H C (1953), *"DNA sequencing at 40:past, present and future"* by Shendure J, Balasubramanian S, Church G M, et al. (2017), *"Specific enzymatic amplification of DNA in vitro:the polymerase chain reaction"* by Mullis K, Faloona F, Scharf S, et al. (1986), *"C2c1-sgRNA complex structure reveals RNA-guided DNA cleavage mechanism"* by Liu L, Chen P, Wang M, et al. (2017), *"C2c1-sgRNA complex structure reveals RNA-guided DNA cleavage mechanism"* by Liu L, Chen P, Wang M, et al. (2017), *"An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference"* by Peng W, Feng M, Feng X, et al. (2015), *"Meningioma:International Consortium on Meningiomas consensus review on scientific advances and treatment paradigms for clinicians, researchers, and patients"* by Wang J Z, Landry A P, Raleigh D R, et al. (2024), *"Imaging live-cell dynamics and structure at the single-molecule level"* by Liu Z, Lavis L D, Betzig E. (2015), *"Genetic regulatory mechanisms in the synthesis of proteins"* by Jacob F, Monod J. (1961), *"Aquaporins in plants"* by Maurel C, Boursiac Y, Luu D T, et al. (2015)

Geology: *"Effects of cattle husbandry on abundance and activity of methanogenic archaea in upland soils"* by Radl V, Gattinger A, Chroňáková A, et al. (2007), *"Proterozoic to Phanerozoic case studies of laser ablation microanalysis for microbial carbonate U–Pb geochronology"* by Jiang Y, Hohl S V, Huang X, et al. (2024), *"Zircon $U–Pb$ ages and $O–H\,f$ isotopes of Quaternary trachytes from the East Sea: Implications for the genesis of low-δ18O magmas"* by Choi H O, Oh J, Kim C H, et al. (2024), *"Fluoride contamination in groundwater: A global review of the status, processes, challenges, and remedial measures"* by Shaji, E., et al. (2024), *"Metagenomic insights into soil microbial communities under climate change"* by Elhottova, D., et al. (2023), *"Molecular biomarkers in ancient sediments:Tracking microbial evolution through deep time"* by Huang, X., et al. (2022), *"Meningioma:International Consortium consensus review on molecular classification"* by Wang, J.Z., et al. (2024), *"Real-time visualization of single-molecule transcription dynamics"* by Hoskins, A.A., et al. (2024), *"Nanoscale mineral-fluid interactions:Implications for carbon sequestration"* by Kim,C.H, et al. (2023), *"Cosmochemical constraints on the sulfur content in the Earth's core"* by Dreibus G, Palme H. (1996), *"A reinforced lunar dynamo recorded by Chang'e-6 farside basalt"* by Shuhui Cai, et al. (2024),

Machine Learning: *"Syntactic representations of semantic merging operations"* by Meyer T, Ghose A, Chopra S. (2002), *"Artificial intelligence in medicine"* by Holmes J, Sacchi L, Bellazzi R. (2004), *"Adding conditional control to text-to-image diffusion models"* by Zhang L, Rao A, Agrawala M. (2023), *"Evaluation of the use of combined artificial intelligence and pathologist assessment to review and grade prostate biopsies"* by Steiner D F, Nagpal K, Sayres R, et al. (2000), *"Erasing concepts from diffusion models"* by Gandikota R, Materzynska J, Fiotto-Kaufman J, et al. (2023), *"Explainable AI over the Internet of Things (IoT):Overview, state-of-the-art and future directions"* by Jagatheesaperumal S K, Pham Q V, Ruby R, et al. (2022), *" Exploring the potential of GPT-4 in biomedical engineering:the dawn of a new era"* by Cheng K, Guo Q, He Y, et al. (2023), *"Blockchain technology in financial sector and its legal implications"* by Divyashree K S, Mishra A. (2022), *"Artificial intelligence in education"* by Moturu V R, Nethi S D. (2022), *"Automated haggling:Building artificial negotiators"* by Jennings N R, Faratin P, Lomuscio A R, et al. (2000),

Climate Science: *"Global warming in the pipeline"* by Hansen J E, Sato M, Simons L, et al. (2023), *"Possible seismogenic-trigger mechanism of methane emission, glacier destruction and climate warming in the Arctic and Antarctic"* by Lobkovsky L I, Baranov A A, Ramazanov M M, et al. (2023), *"Temperature, crime, and violence:A systematic review and meta-analysis[J]. Environmental Health Perspectives"* by Choi H M, Heo S, Foo D, et al. (2024), *"The missing risks of climate change"* byRising J, Tedesco M, Piontek F, et al. (2022), *"Climate change 2022:Mitigation of climate change"* by Shukla P R, Skea J, Slade R, et al. (2022), *"Informing decisions in a changing climate"* by National Research Council, Division of Behavioral, Social Sciences, et al. (2009), *"The impact of the permafrost carbon feedback on global climate"* by Schaefer K, Lantuit H, Romanovsky V E, et al. (2014), *"Permafrost carbon-climate feedbacks accelerate global warming"* by Koven C D, Ringeval B, Friedlingstein P, et al. (2011), *"Global climate change:the potential effects on health"* by McMichael

A J, Haines A. (1997), *"The human imperative of stabilizing global climate change at 1.5 C"* by Hoegh-Guldberg O, Jacob D, Taylor M, et al. (2019),

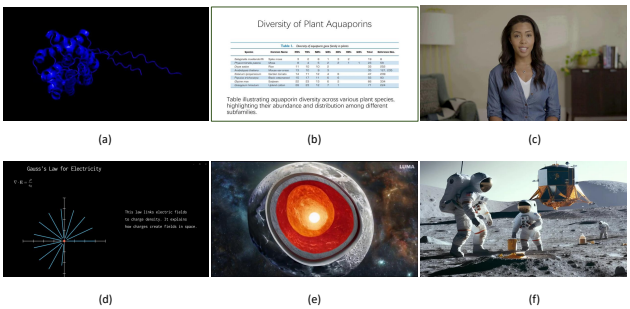

Figure 7. Preacher can generate video abstracts with six different video styles. (a) Molecular Visualization (b) Slides (c) Talking Heads, (d) Mathematics, (e) General Style (f) Static concept.

B. More Details of Preacher

B.1. Different Styles of Video Abstract

A video abstract is a specialized format of video content in which a single video abstract can integrate multiple distinct styles concurrently [19]. Preacher currently supports the generation of six such styles: "talking heads," "general style," "static concept," "molecular visualization," "slides," and "mathematics.", as shown in Fig. 7. Below is a detailed introduction and implementation methodology for each video style:

Molecular Visualization Biological and chemical research papers often include discussions on specific molecular structures. Visualizing these structures enhances reader comprehension by providing a clear representation of complex molecular configurations. We employ pymol=3.0.0 to generate three-dimensional molecular structures from multiple perspectives, which are then compiled into a cohesive video. Given the access restrictions on repositories such as AlphaFold [27], our approach requires users to manually download the PDB (Protein Data Bank) files specified by name in the key scene for processing.

Mathematics Many scientific papers contain critical formulas and theorems. Animating these mathematical principles facilitates deeper conceptual understanding. We utilize Manim = 0.18.1 for generating mathematical demonstrations. Notably, this process is highly susceptible to execution errors caused by code interruptions. To ensure robustness, we predefine a complete script and limit the AI agent's modifications strictly to animation-related functions. Updates to the code are applied through text replacement rather than full script regeneration, significantly improving execution stability and accuracy.

Static Concept Certain papers introduce fundamental real-world concepts that lack intrinsic dynamism. For these, we employ wan-2.1-t2i-turbo [61] to generate high-fidelity static visualizations, which are subsequently expanded into video sequences to enhance engagement.

General Style: Some abstract or complex concepts in academic literature are best conveyed through dynamic visual representation. We utilize Luma [35] to generate high-quality, contextually relevant visual content for these concepts.

Slides Key research findings, such as structured tables and framework diagrams, are indispensable elements of academic papers. We present these elements using a structured slide-based format with python-pptx = 1.0.2, leveraging a standardized slide template where textual and visual content is dynamically replaced to match the specific paper. To extract relevant figures, our agent identifies coordinates based on "source" components within the key scene and captures screenshots from the designated page accordingly. However, this automated process may be prone to errors due to planning errors from LLMs. To mitigate potential failures, users can manually capture and place the necessary images in the designated location to facilitate seamless content generation by Preacher.

Talking Head The Talking Head style facilitates the articulate presentation of key insights through human narration, enhancing audience engagement and comprehension. The Talking Head style allows for the delivery of key insights through human narration. For enhanced presentation quality, we leverage a professional talking head generation API, Tavus [51], which enables the synthesis of high-fidelity facial animations synchronized precisely with the provided audio. This ensures realistic and articulate video outputs with accurate lip-syncing.

B.2. Design of Rule-Based Reflection Agents

Many existing multi-agent frameworks incorporate reflection mechanisms to prevent agents from generating erroneous outcomes [39, 49, 69]. Preacher includes two reflection agents:the Text Reflection Agent $\mathcal{A}_{tref}$ and the Video Reflection Agent $\mathcal{A}_{vref}$. Both agents are equipped with MLLMs but serve distinct tasks. $\mathcal{A}_{tref}$ is responsible for reviewing the input paper and providing a reflection on the current plan. While $\mathcal{A}_{vref}$ evaluates keyframes in the video and compares them with the corresponding part of the paper, offering a reflection on the visual content.

Unlike other work [39, 49, 69], the task faced by Preacher is fixed. Therefore we have build specific rules to guide $\mathcal{A}_{tref}$ and $\mathcal{A}_{vref}$ in conducting targeted checks on the current content. Depending on the different planning stages and video styles, $\mathcal{A}_{tref}$ and $\mathcal{A}_{vref}$ will ask questions with different parts of the rules. These questions primarily focus on the correctness, professionalism, and consistency with the paper. We continuously update these rules throughout the practical process.

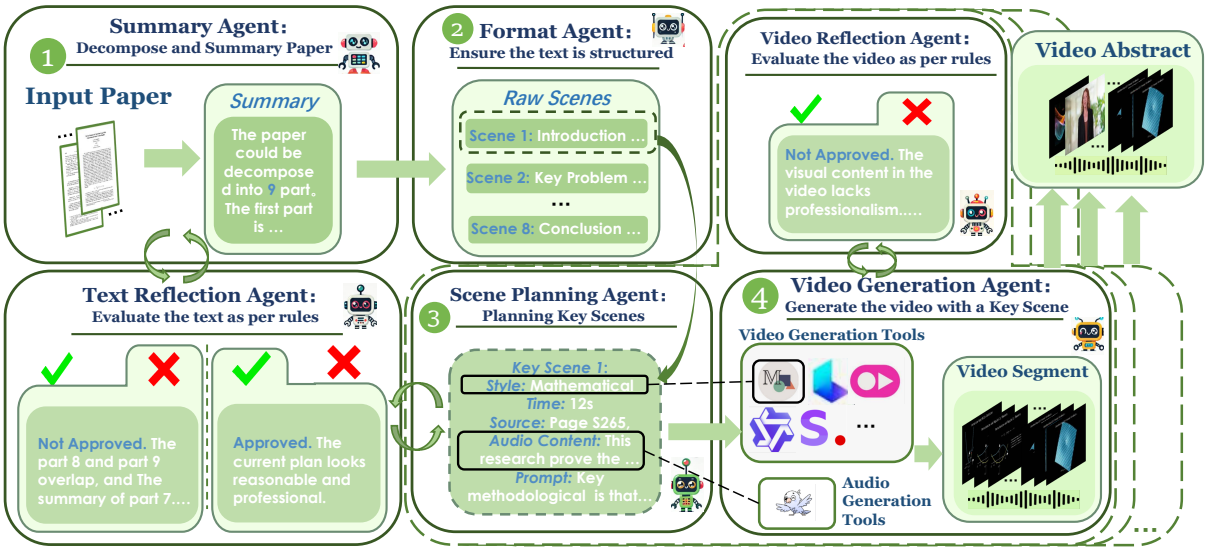

Figure 8. The workflow of Preacher. The parts within the dashed box correspond to the process of generating a video segment. If the number of key scenes is N scenes, the process within the dashed box will repeat N times until all video segments are generated. Finally, all the video segments are concatenated as the video abstract.

B.3. Complete Workflow

Fig. 8 illustrates Preacher's workflow, with the video generation process detailed in Fig. 3. After generating video segments, the corresponding audio track is synthesized based on the "audio content" in the key scenes.

Since the video duration, audio length, and predefined "time cost" in key scenes may not align, we adjust video playback speed—either slowing down or accelerating—to match the designated time cost. However, modifying audio speed introduces distortion. To mitigate this, we constrain audio script length during planning and introduce intentional pauses when the audio is shorter than required.

C. Model Selection

Table 3. Performance comparisons on key scenes on four metrics. We report mean values and standard error. The best is in bold.

Method	Accuracy↑	Professionalism↑	Compatibility↑	Aliment↑
Gemini-2.0-flash	4.70(0.35)	4.63(0.34)	4.38(0.66)	**4.50**(0.31)
GPT-4o [72]	4.55(0.41)	4.30(0.61)	4.45(0.41)	4.40(0.61)
OpenAI-o3-mini [38]	**4.80**(0.12)	4.61(0.38)	**4.50**(0.36)	**4.50**(0.43)
DeepSeek-R1[9]	**4.80**(0.14)	**4.70** (0.19)	3.90(1.12)	4.08(0.91)

The integration of Gemini [52] within Preacher is primarily driven by its API support for direct PDF uploads. To isolate this factor, we conducted experiments where PDFs were manually uploaded via a web interface, and prompts were entered manually. This enabled the replacement of Gemini with alternative MLLMs, allowing a systematic evaluation of their key scene planning performance. As shown in Tab. 3, OpenAI-o3-mini exhibited the most consistent and effective performance. In contrast, DeepSeek generated highly precise yet overly intricate outputs, often deviating from the core content of the input paper. Given that Preacher is model-agnostic, its capabilities are expected to scale with advancements in MLLM performance and interoperability. To further assess Preacher's effectiveness, we directly input key scenes into state-of-the-art (SOTA) video generation models, including Wan-2.1-14B [61] (open-source) and Sora [37] (closed-source). Fig. 9 compares the video segments generated by Preacher, Wan-2.1-14B, and Sora.

General video generation models like Sora prioritize visual continuity and aesthetic appeal but lack the ability to visualize specialized professional content. In Fig. 9(a), existing methods fail to represent the concept of "stable bundles" and omit critical mathematical proofs. In Fig. 9(b), while Sora and Wan attempt to depict "molecular mechanisms," they generate nonexistent cellular structures, misrepresenting the original research. Such inaccuracies significantly diminish the scholarly value of the video abstract. By incorporating multidimensional planning, multi-stage reflective mechanisms, and diverse video generation tools, Preacher ensures an accurate and domain-specific representation of the input paper, preventing the propagation of erroneous knowledge in generated videos.

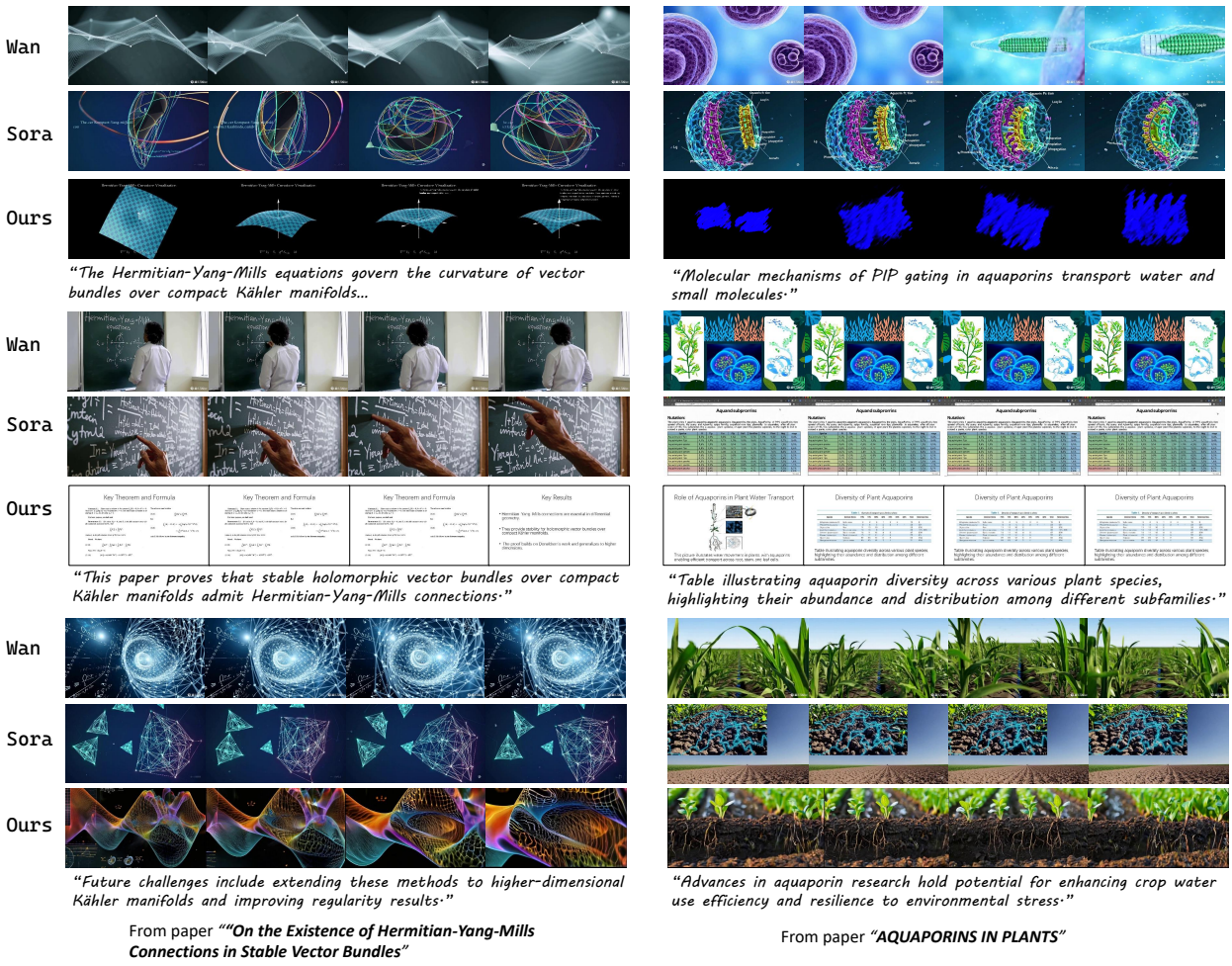

Figure 9. The workflow of Preacher. The parts within the dashed box correspond to the process of generating a video segment. If the number of key scenes is N scenes, the process within the dashed box will repeat N times until all video segments are generated. Finally, all the video segments are concatenated as the video abstract.