Supplementary Material

The supplementary material is structured as follows:

- More details about the dataset construction in 1.
- Surface question-answer pairs generation prompt template in 2
- Answer extraction from predicted captions prompt template in 3
- Video reasoning-based question-answering generation prompt template in 2
- More details about the dataset statistics in 4.
- More details about the evaluation strategies in 5.
- Surface question-answering judgement prompt template in 6.
- Reasoning-based question-answering judgement prompt template in 7.
- More analysis about the model size and overall performance in 8.
- More analysis about the visual token reduction in 9.

1. More Details About Video-MMLU Construction

For video caption, we first use Aria [7] to capture the temporal motion of the video from a global perspective. We set the sampling rate to 1 frame per second, with a resolution of 980. We present the framework of **Video-MMLU** construction process as shown in Figure A1 .

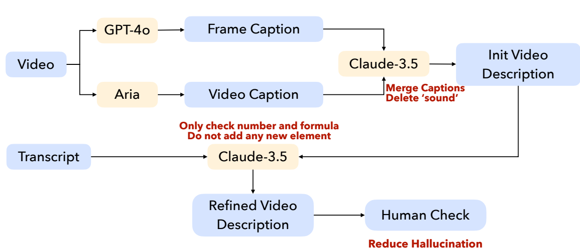


Figure A1. **Video-MMLU** construction pipeline.

2. Question-answer Pairs Generation Prompt Template of Video-MMLU

To decompose the ground-truth structured detailed captions in **Video-MMLU**, we utilize `Claude-3.5-sonnet` as the LLM assistant to generate numerous short question-answer pairs for subsequent evaluation. The full prompt and example cases are presented as followings:

Type	**Prompt**
SYSTEM	You are an intelligent chatbot designed for generating 20 question-answer pairs given a detailed description of a video or image. You are describing the video. Here's how you can accomplish the task: INSTRUCTIONS: - Cover the main objects and actions in the video or image. - The questions should be open-ended and start with 'What', 'Who', 'Where', 'When', 'Why', 'How', etc. - The answer should be a short sentence or phrase. - Generate 20 question-answer pairs.
User	Please generate 15 question-answer pairs given a detailed description of a video or image: detailed description: In this comprehensive mathematical exploration, the lecture delves into the fundamental nature of the exponential function e^t by examining its defining properties and unique characteristics. The presentation begins with a crucial questioning approach, shifting from "What is e^t?ïo the more profound inquiry "What properties define e^t?" This methodological shift proves instrumental in understanding the function's essence . . . Please generate the response in the form of a Python list of tuple with the question and the corresponding answer. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python list of tuple. For example, your response should look like this: [(the question 1, the answer 1), (the question 2, the answer 2), . . .].
`Claude`	["question": "What background color is used in the visual presentation?", "answer": "Dark background", "question": "What is the initial condition of e^t mentioned in the lecture?", "answer": "$e^0 = 1$", "question": "What is the approximate value of $e^{-}2$ shown in the presentation?", "answer": "0.135", "question": "What is the value of e^0.675 shown in the lecture?", "answer": "314.19", "question": "What is the approximate value of e^1.27?", "answer": "3.58", "question": "What is the approximate value of e^3.07?", "answer": "21.47", "question": "What is the approximate value of e^2.17?", "answer": "8.73", "question": "What are the strategic marker positions shown on the number line?", "answer": "25, 50, and 75", "question": "What is the approximate value of e^4.86?", "answer": "128.81", "question": "What is the approximate value of e^3.96?", "answer": "52.37", "question": "What is the approximate value of e^0.37?", "answer": "1.45", "question": "What is the typical span of the number line shown?", "answer": "0 to 9, sometimes extending to 800", "question": "What label is given to the horizontal number line?", "answer": "Position", "question": "What visual elements are used to show relationships on the number line?", "answer": "Arrows pointing to positions and velocities", "question": "What is the derivative of e^2t?", "answer": "$2e^2t$", . . .]
User	[Manually check and refine]

3. Predicted Answer Extraction Prompt Template

Given the question-answer pairs based on the ground truth caption, we utilize `Claude-3.5-sonnet` to extract predicted answers based on the generated caption by our designed prompt template. The complete prompt is shown as followings:

Type	Prompt
SYSTEM	You are an intelligent chatbot designed for providing accurate answers to questions related to the content based on a detailed description of a video or image. Here's how you can accomplish the task:" —— ##INSTRUCTIONS: - Read the detailed description carefully. - Answer the question only based on the detailed description. - The answer should be a short sentence or phrase.
User	Please provide accurate answers to questions related to the content based on a detailed description of a video or image: detailed description: This detailed mathematics tutorial video provides comprehensive instruction on applying the product rule for derivatives, featuring a consistent bright green background and an engaging male instructor positioned in the lower right corner, dressed in a dark jacket. The instructor maintains an enthusiastic and approachable teaching style throughout the presentation, making complex calculus concepts more accessible to viewers. question: What color is the video background? DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide short but accurate answer.
`Claude`	Bright green.

4. More details about Video-MMLU Statistics

We collect videos from the open-source platform YouTube, primarily sourced from ten video creators. Their channel homepage links are as follows:

- brain-station-videos
- cpmrana
- AndyMath
- aleksandr-physics
- 3blue1brown
- LittleSaigonWack
- ProfessorDaveExplains
- DrTrefor
- MathVisualProofs
- TheOrganicChemistryTutor

Figure D3 and Figure indicates the word distribution of the video detailed captions and reasoning question-answering in **Video-MMLU**.

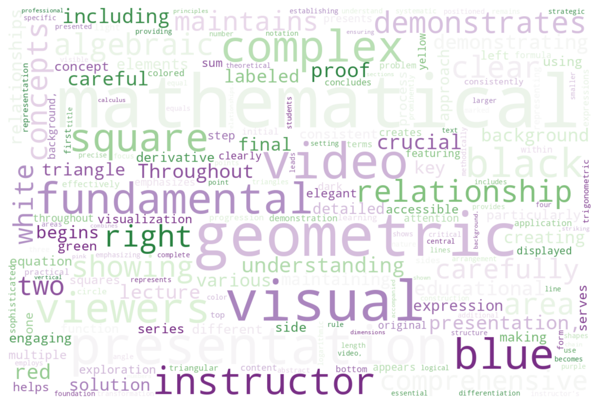


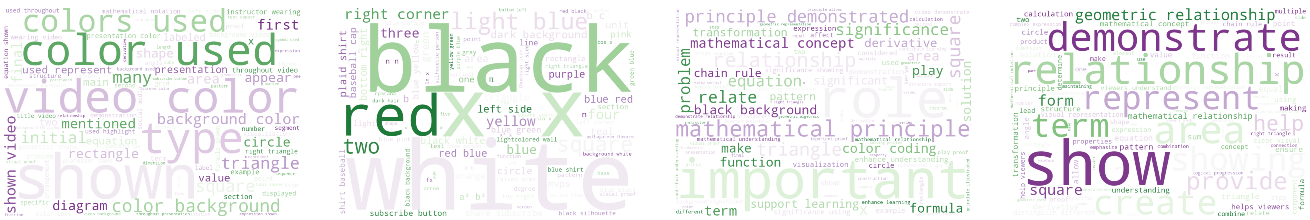
Figure D2. Word cloud of lecture detailed captions in **Video-MMLU**, showing the diversity.

5. More Details about the Evaluation Strategies

The latest versions of the evaluated properietary models before March 2025 were `Gemini-1.5-Flash-002`, `GPT-4o-2024-05-13`, and `Claude-3.5-sonnet-20241022`.

For the visual QA track, we require models to provide concise responses via the system prompt 'Answer briefly and directly in one sentence.' and limit `max_new_tokens` to 64. For the video captioning track, we allow `max_new_tokens` to follow each model's default setting and prompt with the designed instruction in different length randomly as follow:

- The images are given containing equally spaced video frames. Please imagine the video based on the sequence of frames, and provide a faithfully detailed description of this video in more than three sentences.
- You are given a sequence of equally spaced video frames. Based on these frames, imagine the full video and provide a detailed description of what is happening in more than three sentences.
- The following set contains equally spaced video frames. Imagine the video from which these frames were taken and describe it in detail in at least three sentences.
- Below are equally spaced frames from a video. Use these frames to visualize the entire video and provide a detailed description in more than three sentences.
- A sequence of equally spaced video frames is presented. Please imagine the full video and write a faithfully detailed description of the events in more than three sentences.
- The images provided include equally spaced frames from a video. Based on these frames, imagine the video and describe it comprehensively in at least three sentences.
- You are given equally spaced frames from a video. Use these frames to envision the entire video and provide a



(a) Surface Question (b) Surface Answer (c) Deeper Question (d) Deeper Answer

Figure D3. Word cloud of different question-answering pairs in **Video-MMLU**, showing the diversity.

detailed description of the events in more than three sentences.

- The sequence includes equally spaced frames from a video. Imagine the full video based on these frames and provide a detailed description in more than three sentences.
- The provided images contain equally spaced frames from a video. Visualize the video from these frames and describe it in detail in more than three sentences.
- Here are equally spaced frames from a video. Based on these frames, imagine the video and provide a detailed, faithful description of it in more than three sentences.
- The set of images includes equally spaced video frames. Please imagine the video these frames come from and describe it comprehensively in at least three sentences.
- Describe the video based on these frames in a few sentences.
- Explain the video using these frames.
- Imagine the video from these frames and describe it in detail in a few sentences.
- Based on these frames, provide a narrative of the video in more than three sentences.
- Describe the events in the video shown by these frames in at least three sentences.
- Visualize the video from these frames and explain what is happening in more than three sentences.
- Describe the sequence of events in the video depicted by these frames in a detailed manner.
- Given these equally spaced frames, imagine the entire video and provide a detailed description of the events, including the setting, characters, and actions, in more than three sentences.
- Visualize the video based on these frames and write a comprehensive description of what happens, describing the beginning, middle, and end in at least three sentences.
- Using these frames as a reference, imagine the full video and provide a thorough description of the plot, including key details and actions, in more than three sentences.
- Based on the sequence of these frames, describe the entire video in detail, mentioning important aspects such as the context, movements, and transitions in more than three sentences.
- Imagine the video that corresponds to these frames and provide an elaborate description, covering the storyline, visual elements, and any notable features in at least three sentences.

We use Qwen2.5-72B [15] as the LLM evaluation assistant and accelerate with LMDeploy [5].

6. Correctness Evaluation for Detailed Captioning Prompt Template

Following [11], we evaluate the correctness and score of the predicted answers with the assistant of Qwen2.5-72B [16]. Given the question, correct answer, and predicted answer from the generated caption, `Qwen2.5-72B [16]` should return the *True* or *False* judgement and relative score (0 to 5). We specially design a strict prompt for OCR-related question-answering evaluation. The complete prompt is shown as followings:

Type	Prompt
SYSTEM	You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. The evaluation criteria differ based on the type of question: —— ##INSTRUCTIONS: 1. **For OCR-related questions**: - Perform a strict letter-by-letter comparison. - Any difference in characters (including case, punctuation, or letter substitution) must result in 'no'. - Minor spelling errors or missing characters should not be accepted. 2. For **non-OCR-related questions**: - Focus on the meaningful match between the predicted answer and the correct answer. - Synonyms or paraphrases can be considered valid matches. - Minor spelling differences or alternative expressions should not be penalized.
User	Please evaluate the following video-based question-answer pair:

	Question: What specific DNA sequence is shown in the presentation? Correct Answer: TCCGTGCAGTAAATGC Predicted Answer: TTCCGTAATACGACTGCGC Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {'pred': 'yes', 'score': 4.8}.
Qwen2.5	{'pred': 'no', 'score': 0}

7. Correctness Evaluation for Reasoning QA Prompt Template

Following [11], we evaluate the correctness and score of the predicted answers with the assistant of Qwen2.5-72B [16]. Given the question, correct answer, and predicted answer from the generated caption, Qwen2.5-72B [16] should return the *True* or *False* judgement and relative score (0 to 5). We ask the LLM assistan focus on the evaluation of the reasoning process. The complete prompt is shown as followings:

Type	Prompt
SYSTEM	You are an intelligent chatbot designed for evaluating the correctness of generative outputs for reasoning-based question-answer pairs. Your task is to compare the predicted answer with the correct answer based on the following rules: —— ##INSTRUCTIONS: 1. Evaluate Reasoning Tasks Strictly: - The predicted answer must capture all critical concepts and details mentioned in the correct answer. - If the correct answer mentions specific concepts or examples (e.g., 'odd numbers accumulate to form perfect squares'), the predicted answer must include these concepts or examples. - Even if the phrasing differs, the key meaning and concepts must be preserved. However, omitting or altering key concepts or examples is not acceptable. - **Example 1**: If the correct answer is 'The construction method shows how odd numbers accumulate to form perfect squares,' the predicted answer must include 'odd numbers' and 'perfect squares.' - **Example 2:** If the correct answer is 'To eliminate HBr and form an alkene,' the predicted answer must address the elimination of HBr as well. - Minor differences in phrasing are acceptable as long as the key information is retained. - **Critical Detail:** If any essential element (e.g., key terms, concepts, or examples) is missing from the predicted answer, the answer is considered incorrect. - Do **not** introduce new, unrelated information in the predicted answer.
User	Please evaluate the following video-based question-answer pair: Question: What role does RNA polymerase play in the lac operon system? Correct Answer: It initiates transcription of the structural genes when allowed access to the promoter region Predicted Answer: RNA polymerase binds to the promoter gene and transcribes the structural genes when lac operon is active. Provide your evaluation only as a yes/no and score where the score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. Please generate the response in the form of a Python dictionary string with keys 'pred' and 'score', where value of 'pred' is a string of 'yes' or 'no' and value of 'score' is in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. For example, your response should look like this: {'pred': 'yes', 'score': 4.8}.
Qwen2.5	{'pred': 'yes', 'score': 5}

8. More Analysis About the Model Size and Overall Performance

As shown in Figure 8, we visualize the relationship between model size and the average performance across video detailed captioning and question-answering tasks. Generally, larger models (over 20B) tend to achieve better performance, but the scaling trend is not strictly linear. While some models, such as InternVL2.5 [3], show consistent improvements as size increases, others [8, 12] exhibit fluctuating gains. Notably, certain mid-sized models (e.g., around 8B–13B parameters) might outperform larger ones, suggesting that beyond model size, architecture and training strategies play a crucial role in lecture understanding. Additionally, proprietary models like `Gemini-1.5-Flash`, `GPT-4o`, and `Claude-3.5-sonnet` significantly outperform open-source models, highlighting a substantial gap in LMM capabilities.

9. More Analysis about the Visual token Reduction

We supplement Table I1 with the performance of various visual-token reduction models on video detailed captioning and reasoning-based QA tasks in Video-MMLU. As a core design of AuroaCap [2], token merging plays a crucial role in reducing visual token redundancy. Therefore, we systematically test AuroraCap [2] under different visual token kept ratios to analyze how varying compression levels affect video detailed captioning capability to understand its

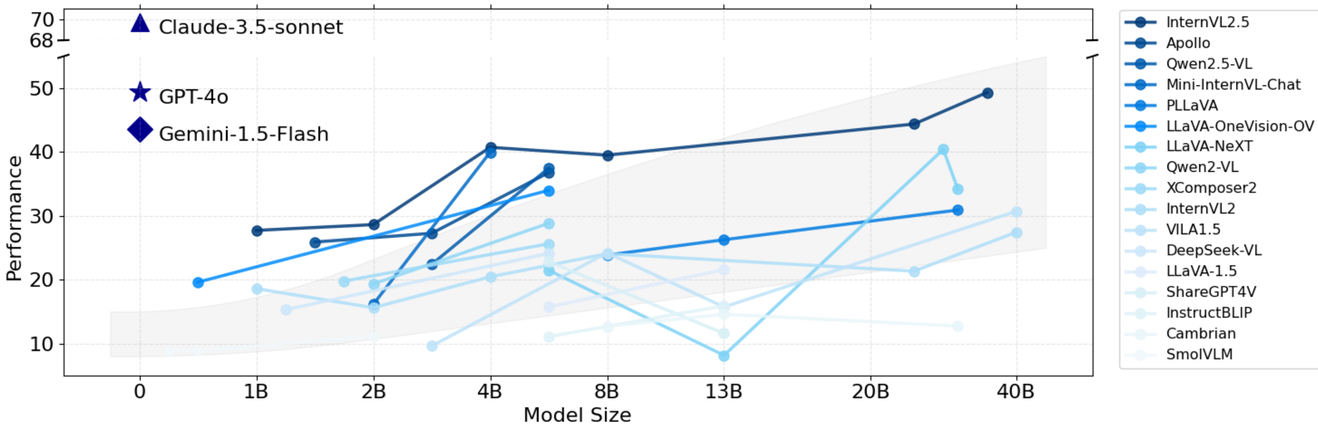


Figure H4. **Relationship between model size and the average video captioning and question-answering performance.** The shaded region shows the confidence interval, with darker colors indicating better performance.

Table I1. Results of visual token compression models on **Video-MMLU**.

Models	LLM	# Tokens Per Frame	Overall	Notebook Avg.	Notebook Math	Notebook Physics	Notebook Chemistry	Quiz Avg.	Quiz Math	Quiz Physics	Quiz Chemistry
Chat-UniVi [6]	Vicuna [4]-7B	112	24.82	21.73	16.68	26.97	21.55	27.91	23.85	25.35	34.55
Chat-UniVi-7B-v1.5 [6]	Vicuna [4]-7B	112	18.66	13.62	10.30	16.84	13.74	23.70	21.21	23.92	25.98
LLaMA-VID [10]	Vicuna [4]-7B	2	19.07	13.87	8.11	14.16	19.33	24.27	12.32	20.24	40.25
	Vicuna [4]-13B	2	21.72	13.53	10.08	12.50	18.00	33.35	28.08	26.63	45.34
AuroraCap [2]	Vicuna [4]-7B	48	21.45	18.36	16.75	15.00	23.33	23.65	19.46	12.24	39.26
		79	22.19	19.55	16.49	17.50	24.66	24.84	21.23	13.34	39.97
		110	26.61	20.18	16.06	19.16	25.33	33.09	21.08	38.08	40.11
		172	23.31	19.62	17.35	17.50	24.00	27.00	19.17	26.67	35.17
		265	22.90	19.55	16.74	26.66	15.25	25.14	25.34	20.12	29.96
		327	27.53	17.84	15.64	22.56	15.33	37.21	28.08	33.31	50.23
		389	26.21	18.97	17.00	22.57	17.35	33.44	23.28	26.93	50.12
		451	25.86	19.31	13.93	21.34	22.65	31.44	23.27	20.73	50.33
		544	21.60	19.17	16.32	20.83	20.35	26.32	20.54	13.43	45.00
		606	21.87	18.78	15.21	20.83	20.31	23.95	23.54	12.97	35.35
		668	17.10	14.13	12.56	15.83	14.00	20.07	22.61	6.57	31.03
		730	20.83	16.78	14.95	21.33	14.07	25.22	23.30	6.64	45.71
VideoChat-Flash [9]	Qwen2.5 [15]-2B	16	25.45	27.03	25.09	25.15	30.86	25.53	21.64	9.93	45.02
	Qwen2 [13]-7B	16	27.71	30.58	30.17	33.43	28.15	24.83	17.82	13.39	43.29
InternVideo2.5 [14]	InternLM2.5 [1]-7B	16	32.29	33.40	29.74	30.05	40.42	31.18	24.65	33.27	35.62
PVC [17]	InternLM2.5 [1]-7B	64	30.00	33.70	27.43	38.33	35.33	26.29	28.07	20.53	30.29

impact, where the number of remaining visual tokens per frame varies from 49 to 730.

Figure I5 illustrates the impact of visual token reduction on performance across video detailed captioning and reasoning-based QA tasks in different disciplines. Overall, most models retain over 80% of their peak performance even when keeping only 20%–40% of visual tokens, suggesting that significant token compression is feasible without severe degradation. However, performance does not follow a strict monotonic trend, indicating that the sequential nature of videos introduces additional complexity in token merging, leading to non-trivial effects on performance.

When comparing captioning and QA tasks, we observe that captioning performance remains relatively stable across different token kept ratios, particularly in mathematics and chemistry. This suggests that structural elements like formulas and static visual cues are more resilient to compression. In contrast, QA performance, especially in physics and chemistry, exhibits sharp fluctuations, highlighting the greater sensitivity of reasoning-based tasks to token reduction. The varying performance across disciplines further reinforces that subjects relying on dynamic visual elements, such as physics and chemistry, require a higher number of retained tokens to maintain accuracy.

Interestingly, optimal performance does not always occur at the highest kept ratio but rather at a mid-range level. This indicates that moderate token merging can improve efficiency without significantly compromising performance, but excessive compression leads to information loss, particularly in reasoning-heavy tasks. Since AuroraCap [2]

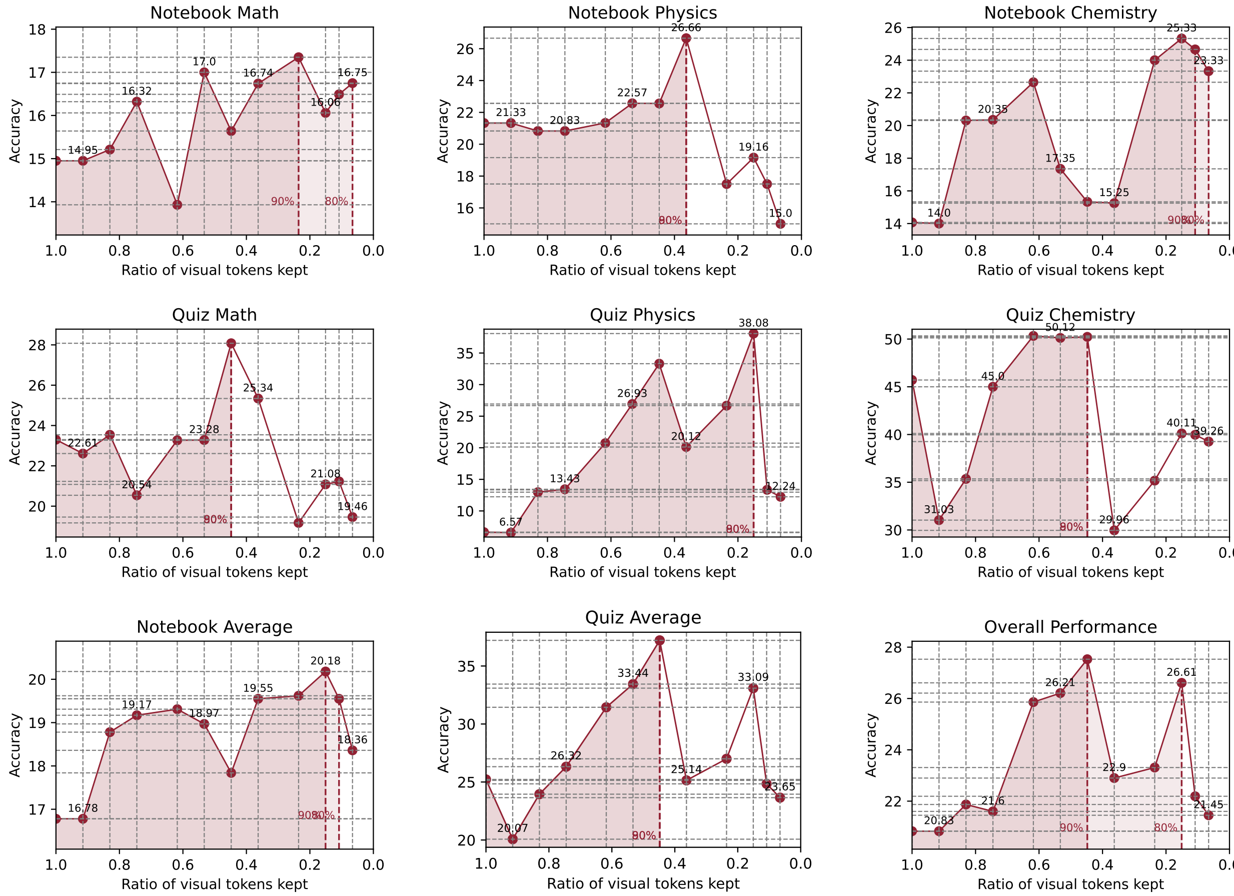


Figure I5. Ablation study of token merging in AuroraCap [2] on **Video-MMLU**.

primarily focuses on spatial token merging without explicitly addressing temporal dependencies, these results suggest that additional strategies are needed to optimize token merging for time-dependent visual reasoning. The results also imply that different tasks and disciplines may benefit from customized token compression strategies rather than a one-size-fits-all approach.

We also compare its effectiveness in video detailed captioning and reasoning-based QA tasks across different desciplines in Figure I6. In the overall performance plot, QA scores (purple line) generally exceed captioning scores (green line) across most token kept ratios, with the largest gap (gray bars) appearing at mid-range ratios (0.4–0.6). Although AuroraCap [2] is designed as a video captioning model, its captioning performance on **Video-MMLU** lags behind its QA performance. We believe that the results is relative to its token merging strategy, which is based on token similarity. In lecture videos, crucial text patches containing key information may be merged due to embedding similarity, leading to degraded caption quality. Reasoning tasks initially benefit from retaining more visual tokens but do not necessarily improve at the highest kept ratios. However, when extreme compression is applied (ratios below 0.2), both QA and captioning performance drop sharply, reinforcing the importance of maintaining a sufficient number of tokens. Across different disciplines, mathematics shows a relatively small gap between QA and captioning across all token ratios, suggesting that both tasks require a similar level of visual information. In physics, QA performance starts significantly lower than captioning at high token counts but surpasses it as fewer tokens are retained, indicating that sequential reasoning in physics may require a more refined token selection strategy. Chemistry consistently shows the widest gap favoring QA, particularly at mid-range token ratios, suggesting that chemistry reasoning benefits more from retaining structured visual elements and textual annotations.

To provide a more intuitive comparison, we present AuroraCap [2]'s video captioning results at different token kept ratios on a video about the Arithmetic Mean-Root Mean Square Inequality demonstration. We observe that as the visual token kept ratio increases (retaining more tokens), the

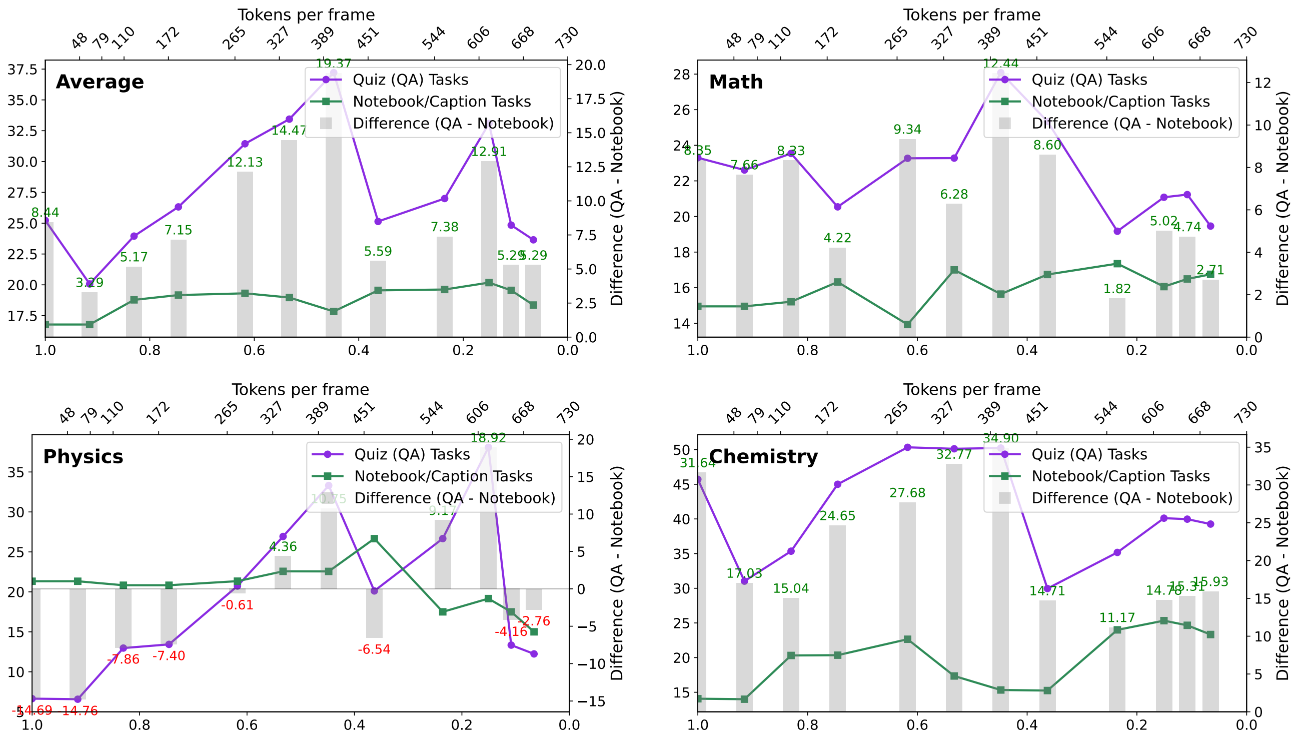


Figure I6. Performance comparsion of token merging in AuroraCap [2] on **Video-MMLU**across different discipline.

generated captions become progressively shorter.

Token Describe this image in detail. (Figure ?? Row 1)

GT The educational video opens with a distinctive title screen displaying **"MVPs"** and **"The Arithmetic Mean-Root Mean Square Inequality"** against a dark background, accompanied by a geometric logo in blue and purple tones and a subscribe button with a thumbs-up icon.

The mathematical presentation establishes its fundamental premise by introducing two positive real numbers, a and b, displayed against a black background. The geometric construction begins with two adjacent squares: a blue square with side length a and a maroon square with side length b. This fundamental construction serves as the foundation for demonstrating a profound relationship between arithmetic and quadratic means. Each square is methodically divided by diagonal lines intersecting at their respective centers, creating four congruent triangular sections. This division is crucial as it creates isosceles right triangles within each square, with legs measuring $a/2$ in the blue square and $b/2$ in the maroon square.

The geometric visualization advances by connecting the center points of the two squares, forming a right triangle with legs measuring $a/\sqrt{2}$ and $b/\sqrt{2}$. These measurements arise from the fact that these segments are hypotenuses of the isosceles right triangles formed within each square. Through the application of the Pythagorean theorem to this connecting triangle, the hypotenuse measures $\sqrt{a^2/2 + b^2/2}$. The construction creates a trapezoid formed by two isosceles right triangles and the connecting triangle, which proves fundamental to establishing the inequality: $\frac{a}{2} + \frac{b}{2} \leq \sqrt{\frac{a^2}{2} + \frac{b^2}{2}}$.

The proof, credited to Juan-Bosco Romero Márquez, elegantly demonstrates that equality occurs if and only if $a = b$. This condition becomes geometrically evident as the squares would be identical in this case, creating a symmetric configuration. Throughout the presentation, the squares maintain their distinct blue and maroon/purple coloring, with their dimensions consistently labeled. When the squares overlap, the resulting purple-shaded intersection area provides additional visual insight into the mathematical relationship.

The visualization effectively combines algebraic expressions with geometric representations, demonstrating how the arithmetic mean (represented by $(a/2 + b/2)$) relates to the root mean square ($\sqrt{a^2/2 + b^2/2}$) through the careful arrangement and interaction of colored squares and triangles. This geometric approach provides a compelling visual proof of the arithmetic mean-root mean square inequality, showing that the arithmetic mean of two positive numbers is always less than or equal to their root mean square, with equality occurring only when the numbers are identical.

Multiple variations of the core geometric arrangement

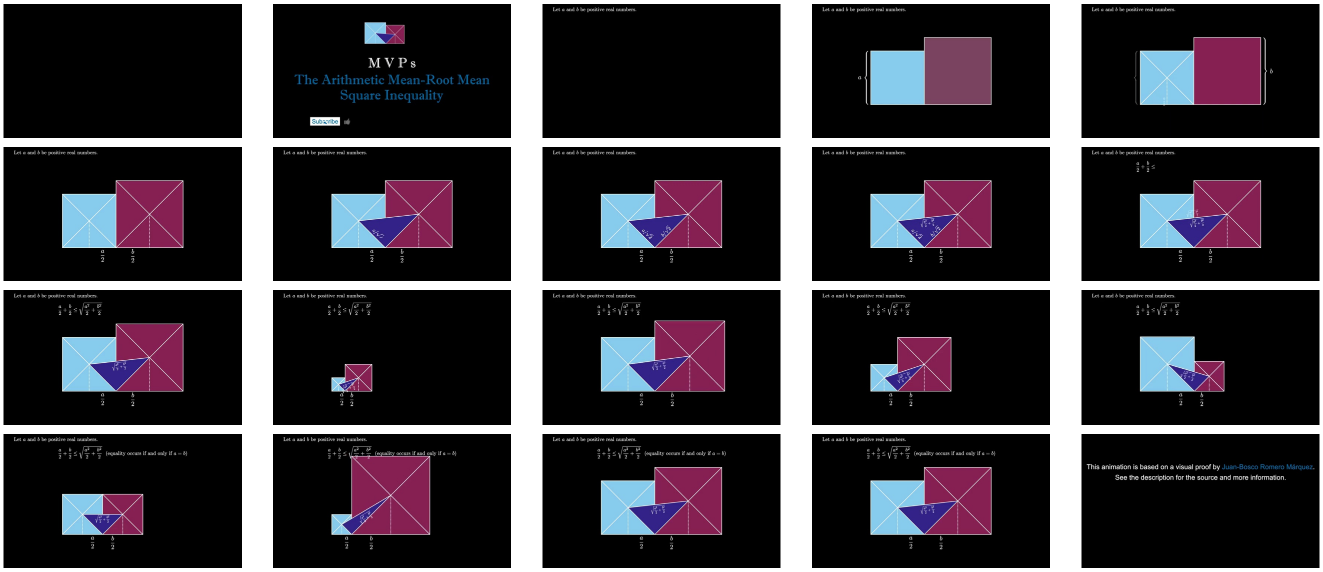


Figure I7. Example video frames in **Video-MMLU**. The video YouTubeID is $9Qyba_swKTI$, which is focus on the demonstration of the Arithmetic Mean-Root Mean Square Inequality, provided by *Mathemativs Visual Proofs*.

are presented, each emphasizing different aspects of the inequality. The black background provides strong contrast for the blue, maroon, and purple geometric elements, ensuring the mathematical relationships remain clear and visually striking. Every frame builds upon the previous understanding, with careful attention to labeling, color coding, and the precise geometric relationships between the squares and triangles, maintaining consistent use of variables throughout while demonstrating their relationships both algebraically and geometrically.

48 The video presents a vibrant and engaging exploration of mathematical concepts, focusing on the properties of triangles. It begins with a bold statement, Ä triangle is a real number,s̈etting the stage for a deep dive into the world of geometric shapes. The scene transitions to a black background, where a large, brightly colored triangle dominates the frame. The triangle is divided into two distinct sections, each filled with a different color, creating a striking visual contrast.
As the camera pans across the triangle, it reveals a series of mathematical equations and expressions that highlight the relationship between the lengths of the sides and the angles of the triangle. The text, Ä triangle is a real number. (equivalently, if f is a function, then f(a) is a real number),äppears, further emphasizing the connection between mathematical concepts and real-world applications. The camera then zooms in on the triangle, allowing viewers to appreciate the intricate details of the shapes and the precise placement of the mathematical notations.
The video continues to explore the triangle's properties, showcasing various angles and their corresponding expressions. The camera pans slowly, allowing viewers to absorb the information and appreciate the beauty of the geometric shapes. The background transitions from black to a gradient of blue and purple, enhancing the visual appeal and drawing attention to the mathematical content.
As the video progresses, it delves deeper into the triangle's attributes, including its area and the sum of its angles. The camera captures the angles with precision, highlighting their significance in the triangle's structure. The text, Ä triangle is a real number. (equivalently, if f is a function, then f(a) is a real number),r̈eappears, reinforcing the connection between mathematical concepts and real-world applications.
The video concludes with a thoughtful reflection on the triangle's significance, emphasizing its role as a fundamental shape in mathematics. The camera zooms in on the triangle, allowing viewers to appreciate the intricate details of the shapes and the precise placement of the mathematical notations. The background transitions to a gradient of blue and purple, enhancing the visual appeal and drawing attention to the mathematical content.
Throughout the video, the camera work is smooth and deliberate, allowing viewers to absorb the information and appreciate the beauty of the geometric shapes. The use of color, lighting, and camera angles creates a visually engaging experience that highlights the importance of mathematical concepts in our daily lives.

79 The video presents a vibrant and dynamic scene set in a lively urban environment. The main focus is on a bustling street filled with a diverse crowd of people, each engaged in various activities. Some are walking briskly, while others are seen enjoying leisurely strolls, creating

a lively atmosphere. The street is adorned with colorful buildings, each showcasing unique architectural styles, and the atmosphere is enhanced by the presence of street vendors and parked cars.
As the camera pans through the scene, it captures the essence of city life, highlighting the interplay of movement and stillness. The sound of chatter, laughter, and the occasional car horn fills the air, enhancing the sense of realism. The camera work is smooth, transitioning seamlessly between different angles and perspectives, allowing viewers to appreciate the depth and complexity of the urban landscape.
The video transitions to a cozy indoor setting, featuring a warmly lit living room where a group of friends are gathered, engaged in animated conversation. The atmosphere is inviting, with soft lighting and comfortable furniture, creating a sense of intimacy and camaraderie. The friends exhibit a range of emotions, from laughter to contemplation, reflecting the dynamic nature of human interaction.
As the video progresses, it shifts focus to a serene outdoor setting, where a person is seen enjoying a peaceful moment by the water. The tranquil environment contrasts with the earlier bustling street scene, providing a calming visual experience. The camera captures the beauty of the natural surroundings, with gentle ripples on the water and lush greenery in the background, evoking a sense of calm and reflection.
The video concludes with a return to the urban environment, showcasing a vibrant street scene filled with people, vehicles, and colorful buildings. The camera work is dynamic, capturing the essence of city life with a mix of close-up shots and wider angles, allowing viewers to appreciate the intricate details of the urban landscape. The sound of traffic, chatter, and laughter fills the air, encapsulating the essence of a lively, dynamic city.

110 The video presents a series of mathematical diagrams and explanations, focusing on the concept of real numbers. The initial frame displays a large, bold text that reads, Ïn this video, we explore the basics of real numbers.T̈his sets the stage for an educational series that aims to delve into the intricacies of mathematical concepts.
As the video progresses, the visuals transition to a vibrant display of geometric shapes and lines, creating a striking contrast against a dark background. The shapes are primarily triangles, with one large triangle in a deep purple color and a smaller one in a lighter shade, both outlined in white. The triangles are interconnected by lines, forming a complex network that suggests a visual representation of mathematical relationships.
The narrative is enhanced by a clear, informative text overlay that explains the significance of the shapes and their connections. The text is presented in a clean, white font that stands out against the dark backdrop, ensuring readability. The combination of visual elements and text creates a dynamic learning experience, making the subject matter accessible and engaging.
Throughout the video, the camera work is smooth, with a steady focus on the shapes and text, allowing viewers to absorb the information without distraction. The transitions between frames are seamless, maintaining a cohesive flow that keeps the viewer's attention on the educational content.
As the video nears its conclusion, the text shifts to emphasize the importance of understanding the source of real numbers, as it is the foundation of all mathematical concepts. The final frame features a concluding statement that reinforces the significance of the topic, leaving viewers with a deeper appreciation for the intricate world of real numbers.

172 The video presents a series of mathematical diagrams and text, focusing on the concept of real numbers and their properties. The initial frame features a large, bold text that reads, T̈his is a visual proof by Juan-Borrego of the real numbers.Ërelow this text, a geometric figure is displayed, consisting of a large square with a smaller square inside it, creating a visual representation of the concept of a real number line. The figure is color-coded, with the larger square in blue and the smaller square in purple, with a dotted line indicating the real number line.
As the video progresses, the focus shifts to a more complex geometric figure, where the same square and line are now enclosed within a larger square, creating a three-dimensional perspective. The text below this frame reads, T̈he description for the source information.T̈his frame is followed by a mathematical expression, ä = b = 2,ẅhich is highlighted in blue, indicating the equality of the variables 'a' and 'b' with the number '2'. The next frame introduces a new element, a right-angled triangle with the hypotenuse labeled 'a' and the legs labeled 'b' and 'c'. The text below this frame states, T̈he description for the source information.
The video then transitions to a frame where the triangle is rotated, revealing a new perspective. The text below this frame reads, T̈he description for the source information.T̈he final frame presents a concluding statement, J̈uan-Borrego's proof of the real numbers.T̈he text is accompanied by a mathematical expression, ä = b = c = 2,ẅhich is highlighted in blue, emphasizing the equality of the variables 'a', 'b', and 'c' with the number '2'. The video concludes with a frame that features a large, bold text stating, T̈his is a visual proof by Juan-Borrego of the real numbers.
Throughout the video, the background is a solid black, providing a stark contrast to the vibrant colors of the geometric figures and text. The text is presented in a clear, sans-serif font, ensuring readability. The overall layout is organized and methodical, guiding the viewer through the mathematical concepts being presented.

265 The video presents a mathematical explanation focusing on the concept of real numbers, as described by the visual elements of a geometric figure. The figure is a large square divided into smaller squares, with a diagonal line creating a right-angled triangle. The triangle is highlighted in blue, with its hypotenuse labeled as 'a' and

the legs as ’b’ and ’c’. The square is labeled ’a’ and ’b’, while the smaller squares are labeled ’a’ and ’b’ as well. The background is a deep blue, providing a contrast that emphasizes the geometric shapes.
At the top of the video, there is a text overlay in white that reads, T̈his is based on a visual proof by Juan-Bosco Rodriguez.B̈elow this, in a larger font, the text states, T̈he real numbers are based on a visual proof by Juan-Bosco Rodriguez.T̈he text is clear and legible, set against the dark blue background. The overall layout is clean and organized, with the geometric figure and text providing a clear visual representation of the mathematical concept being explained.

327 The video presents a mathematical concept, specifically focusing on the relationship between the number of real numbers and their representation. It features a geometric diagram with a large purple square and a smaller blue square, both sharing a common side. The purple square is divided into four smaller triangles, each labeled with a number from 1 to 4, while the blue square is divided into two triangles, labeled with ’a’ and ’b’. The triangles are arranged in a way that suggests a visual proof of a mathematical concept, with the purple square representing the set of real numbers, while the blue square represents a specific subset of these numbers. The video is set against a black background, emphasizing the vibrant colors of the geometric shapes. The text overlay, in white font, provides context to the visual, stating that the number of real numbers is based on a visual proof by Juan-Bosco Rojas, and it describes the source of information as more informative.

389 The video presents a mathematical concept, focusing on the relationship between the number of real numbers and their representation. It features a geometric diagram with a large purple square and a smaller blue square, both sharing a common side. The purple square is divided into smaller triangles, each labeled with a mathematical expression. The expressions include the square root of a number, a fraction, and a variable ’b’. The blue square is a smaller representation of the purple square, with a similar division into triangles, each labeled with a different expression. The expressions are mathematical in nature, involving the square root of ’a’ and ’b’, a fraction ’a/b’, and a variable ’b’. The video is educational, aiming to explain the concept of real numbers and their visual representation.

451 The video presents a series of geometric shapes and mathematical expressions, primarily focusing on the concept of real numbers. It begins with a large blue triangle, labeled ’a’, which is divided into two smaller triangles, ’b’ and ’c’, creating a visual representation of the number line. The number ’2’ is placed at the bottom of ’a’, while ’1’ is at the top, indicating the scale of the number line. The shapes are set against a black background, enhancing their visibility.
A mathematical expression, ’(2a + b) = a + b’, is displayed above the triangles, suggesting a relationship between the lengths of the sides. The expression is highlighted in white text, contrasting with the dark background. The video transitions to a purple square, ’d’, with a smaller square ’e’ inside it, creating a visual representation of the square root symbol. The expression ’2̆21a(2a) = a’ is shown, indicating the square root property. The purple square is labeled ’d’, while ’e’ is the square root symbol. The video concludes with a text overlay in white, stating, J̈uan-Bosco Roque based on a visual proof by Jian-Bosco Roque,ẅhich credits the creator of the visual representation.

544 The video presents a mathematical concept, focusing on the properties of triangles and their relationships with real numbers. It begins with a title that reads, J̈uan-Bosco Ro.B̈elow the title, a statement explains that the video is based on a visual proof by Juan-Bosco Ro. The main visual element is a geometric diagram featuring a large triangle with a smaller triangle inside it, both sharing a common side. The larger triangle is colored in shades of blue, while the smaller one is in purple. The shared side is highlighted with a dotted line, indicating its significance.
The video then transitions to a black background where the text continues, stating that the description is based on the source and more information can be found. The text is white, providing a clear contrast against the dark backdrop. The overall layout is clean and organized, with the text and diagram clearly separated, allowing for easy comprehension of the mathematical content.

606 The video presents a mathematical concept, focusing on the relationship between the number of real numbers and their representation. It features a geometric diagram with a large purple square and a smaller blue square, both sharing a common side. The purple square is divided into four smaller squares, each labeled with a number from 1 to 4. The blue square is also divided into four smaller squares, with the top left square labeled ’a’ and the bottom left ’b’. The video explains that the number of real numbers is based on a visual proof by Juan-Bosco, which is described as a source for more information. The text is overlaid on a black background, enhancing readability. The video is educational, aiming to convey the mathematical concept of real numbers through visual representation.

668 The video presents a vibrant and dynamic scene featuring a bustling cityscape at dusk. The sky is painted with warm hues of orange and pink, transitioning into a deep blue as the sun sets. Skyscrapers with illuminated windows rise against the backdrop of the fading daylight, creating a striking contrast. The streets below are alive with the movement of people, vehicles, and the glow of streetlights. The atmosphere is further enhanced by the presence of a river reflecting the city lights, adding a serene touch to the lively urban environment.

References

[1] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 5

[2] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024. 4, 5, 6, 7

[3] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 4

[4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 5

[5] LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. `https://github.com/InternLM/lmdeploy`, 2023. 3

[6] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 5

[7] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 1

[8] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 4

[9] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 5

[10] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 5

[11] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 3, 4

[12] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 4

[13] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 5

[14] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025. 5

[15] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 3, 5

[16] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3, 4

[17] Chenyu Yang, Xuan Dong, Xizhou Zhu, Weijie Su, Jiahao Wang, Hao Tian, Zhe Chen, Wenhai Wang, Lewei Lu, , and Jifeng Dai. Pvc: Progressive visual token compression for unified image and video processing in large vision-language models. *arXiv preprint arXiv:2412.09613*, 2024. 5