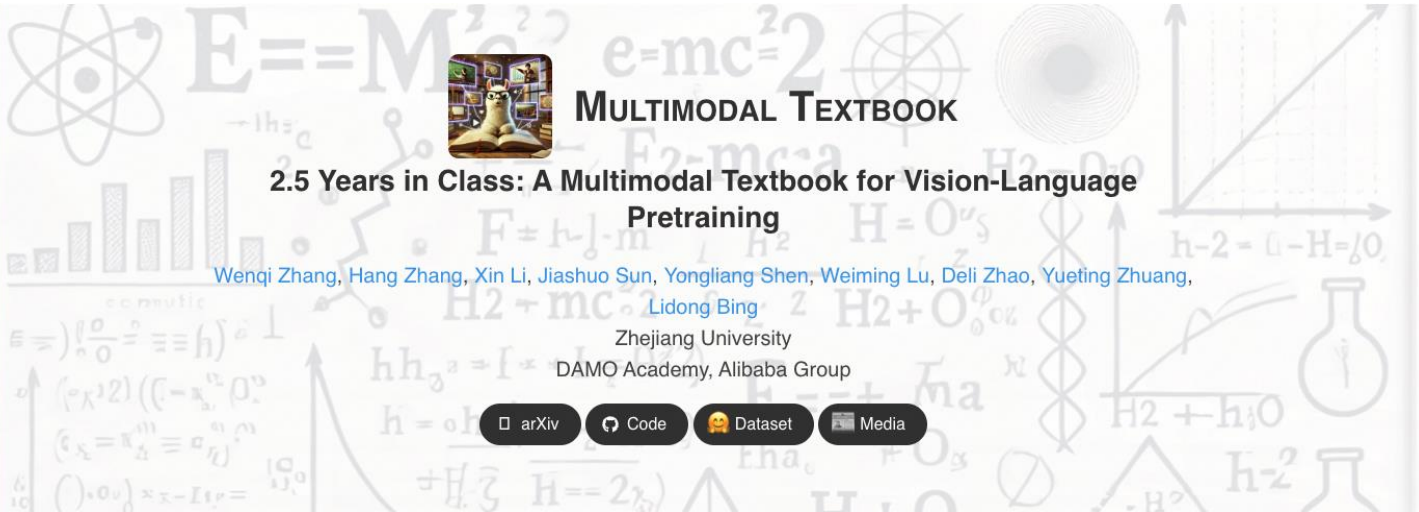
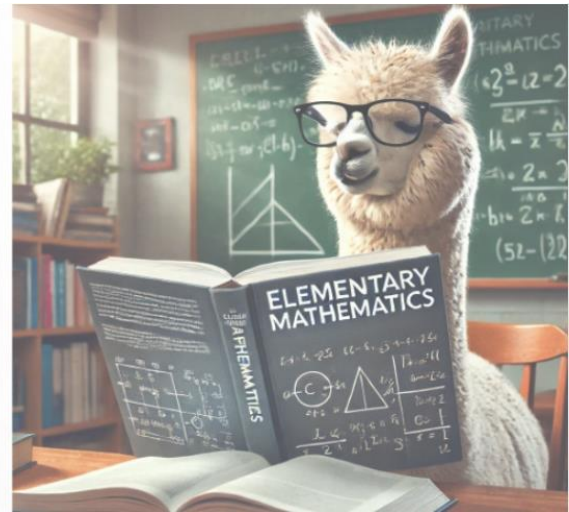
2.5 Years in Class: A Multimodal Textbook for Vision-Language Pretraining

ICCV 2025 Highlight

zhangwenqi@zju.edu.cn





Online Instructional Videos

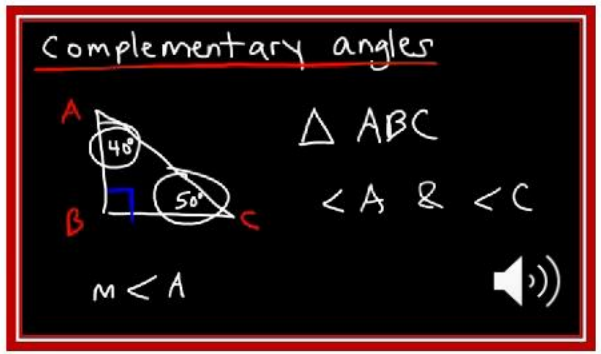
- **159K** instructional videos
- **22000 class hours** (2.5 years)

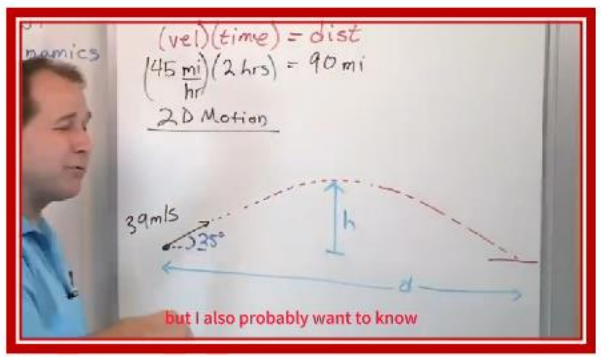
High-quality Corpus

- Image-text **Interleaved For Pretraining**
- Textbooks for **six fundamental subjects**

Vision-Language Models

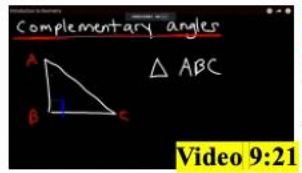
- **6.5M Keyframes**
- **0.75B Tutorial Texts**



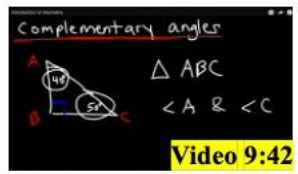


Video-to-Textbook

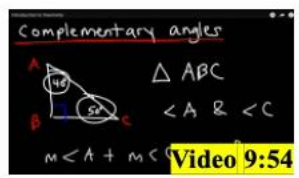
《Textbook: Mathematics》



Tutorial Text Extract From Video: The next term in Geometry is complementary angles. So, what are Complementary Angles? Complementary Angles are two angles whose measures add up to 90…..



Let's consider a right triangle, and we will label it as triangle ABC. The symbol for this triangle is as follows: triangle ABC



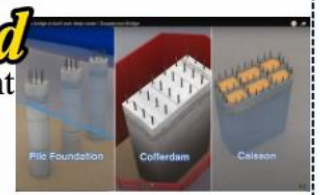
angle A measures 40 degrees and angle C measures 50 degrees. In this case, we can say that angle A and angle C are complementary, because the sum of their measures equals 90 degrees

So, the fundamental concept behind Complementary Angles is that the measure of angle A plus the measure of angle C is equal to 90 degrees….

《Textbook: Engineering》

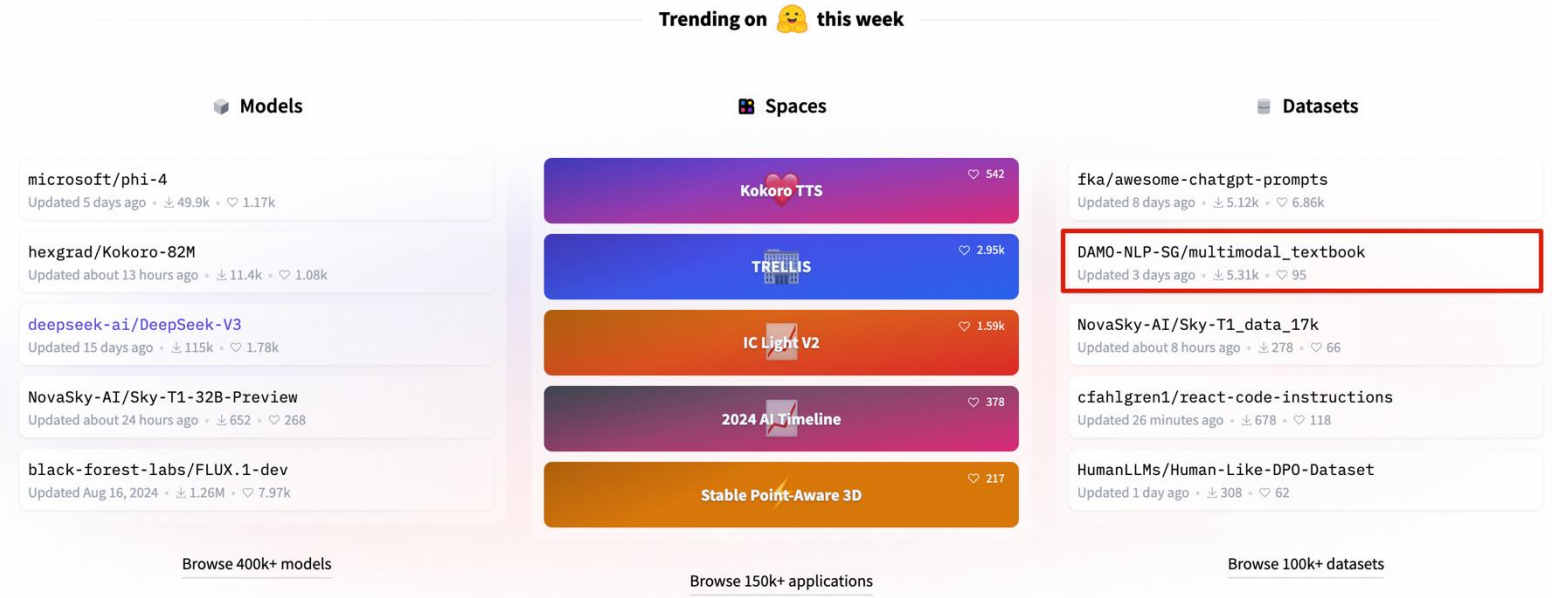
In this video, I'm using a suspension bridge as an example. Let's first discuss how it works. The parts of a suspension bridge include towers, anchors, main cables, hangers or suspenders, and the bridge deck. When vehicles drive across the deck and exert weight, the deck is…..

…g…
equal force exerted on both sides. Finally, the weight is transferred to the towers. On both sides of the bridge, there are anchors that hold the main cables.

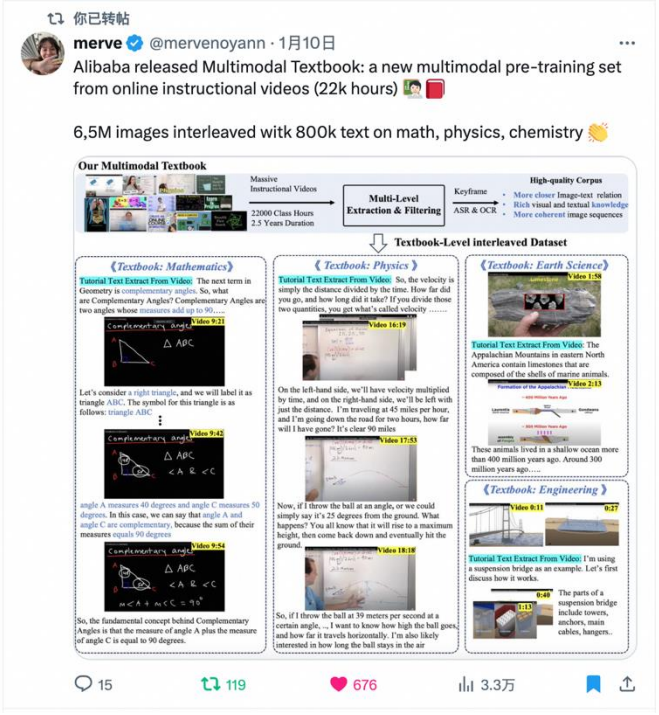


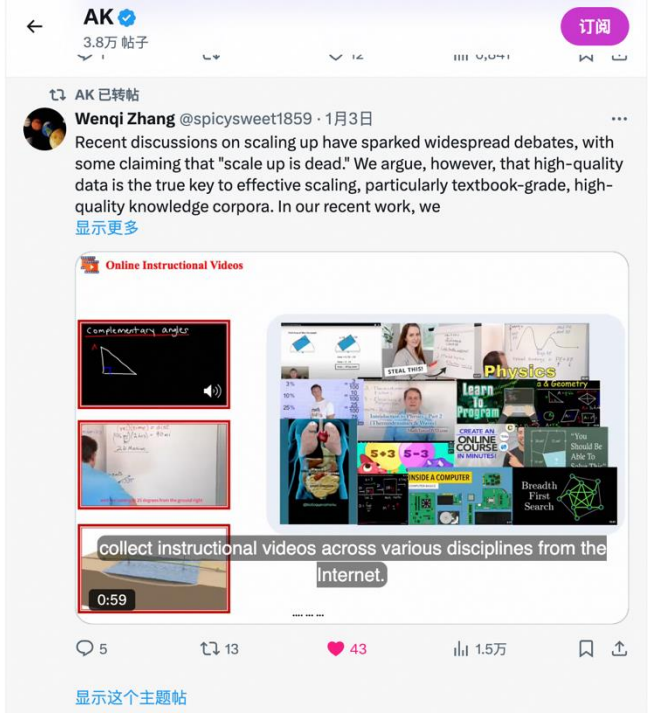
…… … …

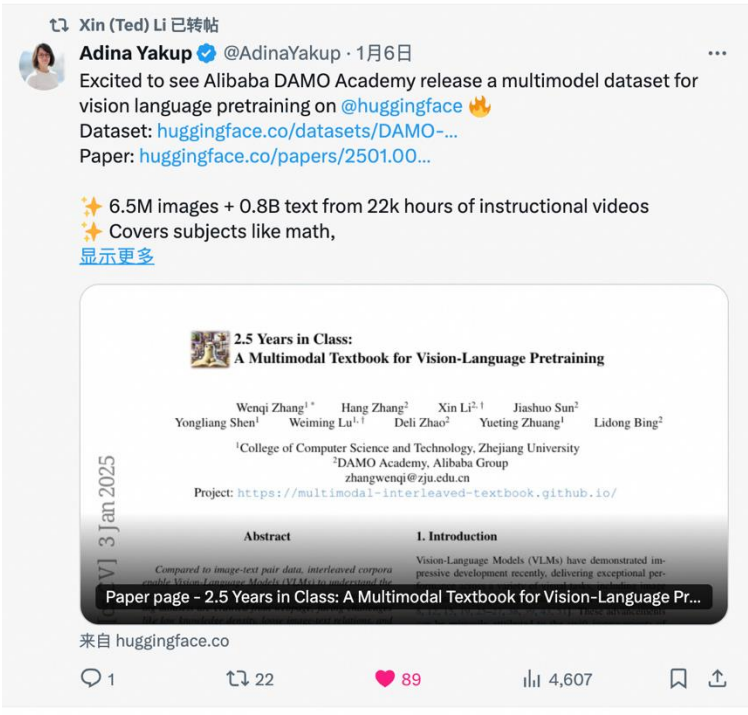
these richly contextualized and professionally curated



- **Rank #2** on Hugging Face Trend
- Over **30,000 downloads** in the community
- Received considerable attention on Twitter, exceeding **50,000 views**

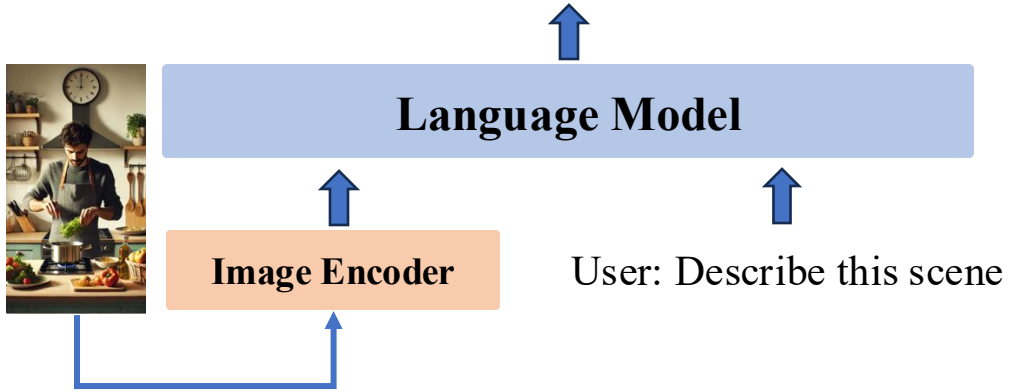






Pretraining Corpora

Vision Language Model (VLM):



- Most multimodal models are pretrained on image-text paired data
- Two different Pretraining Corpora

1. Two different pre-training corpora

Image-text Pair Data

Query: Please Describe this scene in detail ....
Caption: A man is cooking in the kitchen…

<Image, Caption>

Image-text Interleaved Data

A man is cooking. First, he cuts some vegetables and puts them into the pot.

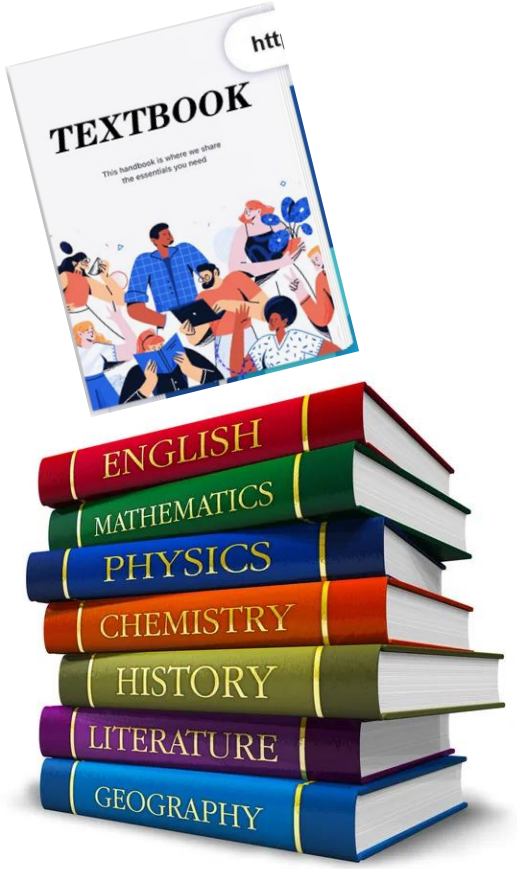
Then he took a shovel and stirred it in the hot pot, watching ….

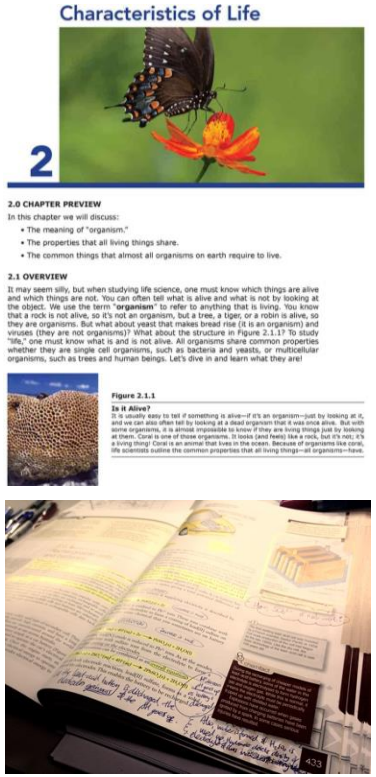
< Text, Image, Text, Image, Text, …. >

- **Easy** to collect
- **Simple** descriptive text
- Only **statically** describe images

- **Difficult** to collect
- **Dynamically** describe **continuous** actions and **complex** processes

What should the human learning process be like?





- Experts carefully design **richly illustrated textbooks** for each subject and course
- Learn knowledge and the underlying logic more profoundly through **textbooks with illustrations and text**
- Learn the knowledge of various disciplines, progressing **from easy to difficult**
- Consolidate knowledge through **after-class exercises**

What should the human learning process be like?

- From image-caption pair data to image-text interleaved data

2. Most previous interleaved data is crawled from web pages

➢ **Image-text Relation is Loose and Noise**

If 'eclectic' to you is when Green Day change their guitar tone or McDonald's puts two burgers in one bun, then steer clear…..

If however you take your pepperoni pizza with extra cream and can stomach the idea of an album with ….

➢ **Lacking Connection Between Images**

Firearm Licensing and Registration Act would establish licensing requirements to posses a firearm and ammunition, including a psychological evaluation ….

Individuals hospitalized with a mental illness ….

➢ **Low Knowledge Density**

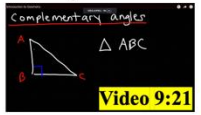
Dedicated to mince, peel and cut with delicacy, the slicing knives are precision tools that you have to choose with care …

The high zirconium oxide content of the ceramic blade of these TB knives makes it a premium tool …

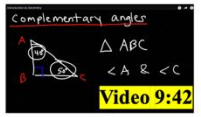
3. Our Multimodal Textbook From instructional videos

《Textbook: Mathematics》

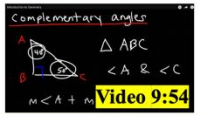
Tutorial Text Extract From Video: The next term in Geometry is complementary angles. So, what are Complementary Angles? Complementary Angles are two angles whose measures add up to 90…..



Let's consider a right triangle, and we will label it as triangle ABC. The symbol for this triangle is as follows: triangle ABC ….

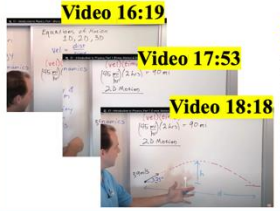


angle A measures 40 degrees and angle C measures 50 degrees. In this case, we can say that angle A …..



So, the concept behind Complementary Angles is that the measure of angle A …

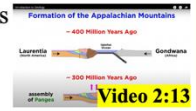
《Textbook: Physics》



<Image>So, the velocity is simply the distance divided by the time. How far did you go, and how long did it take? If you divide those two quantities, you get what's called velocity …….<image>…..<image>

《Textbook: Earth Science》

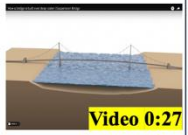
The Appalachian Mountains in eastern North America contain limestones that are composed of shells of ….



《Textbook: Engineering》

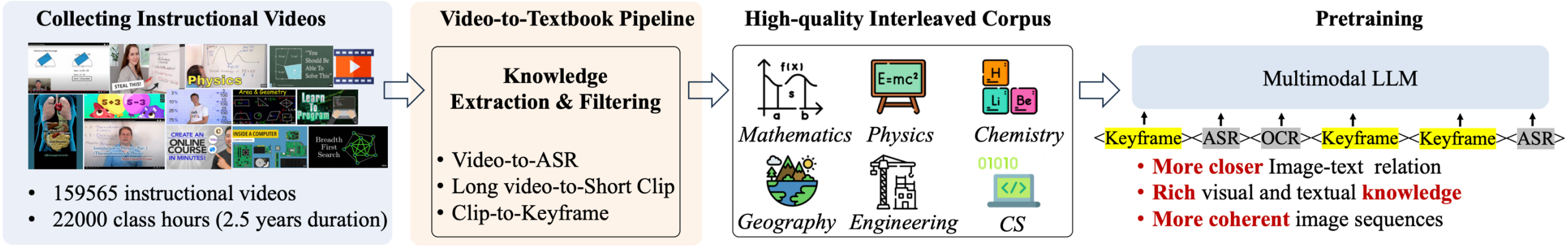
I'm using a suspension bridge as an example…

The parts of a suspension bridge include towers ….

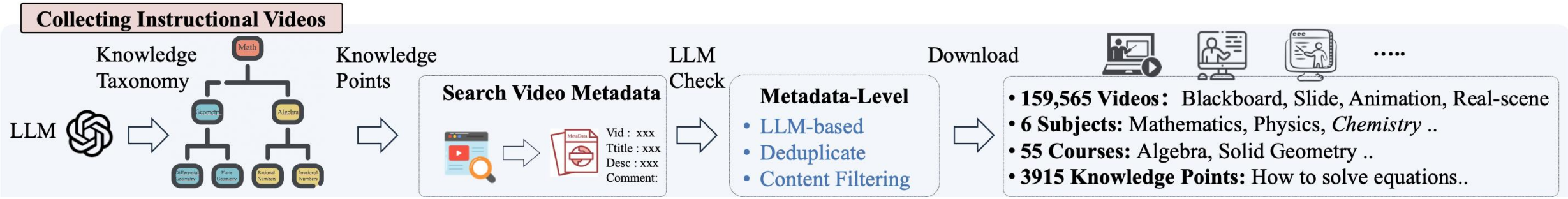


Multimodal Textbook: Image-text interleaved corpora from instructional videos

- There is a vast amount of **instructional videos** on the Internet, including knowledge teaching and action teaching videos (such as cooking and yoga teaching).
- These video provides a wealth of knowledge in **various forms**: images, voice (explanatory dubbing), text (text in the frame), as well as viewpoints from multiple perspectives: the author's self-description, viewers' bullet comments, reviews, ratings, etc.
- Videos inherently display **dynamic processes** and are very suitable for learning knowledge/concepts/actions.



Construct Knowledge Taxonomy



- **Knowledge taxonomy synthesized with LLM agent**

Subject → Course → Sub-course → Knowledge Point

- **Search for corresponding video metadata based on the knowledge taxonomy**

Collected metadata of 159k instructional videos and labeled each video with its corresponding knowledge point

- **Metadata filtering and video crawling**

Step1: Use LLM to review the theme, introduction, comments, and ratings of each video, and filter out low-quality instructional videos

Step 2: Crawl the corresponding videos and store them according to the tree structure

Collect Instructional Videos

Subject	#Video	Duration (h)	#Topic	#Video Clip	#Keyframe	#ASR Token	#OCR Token	#Sample
Mathematics	21.7k	4,423	725	809k	1.67M	72.5M	145M	123k
Physics	11k	3,511	530	822k	0.95M	36.7M	73.4M	119k
Chemistry	4.5k	2,643	410	234k	0.49M	15M	30M	32k
Earth Science	12k	3,670	520	640k	1.03M	40M	80M	88k
Engineering	13k	4,096	810	713k	1.15M	43.3M	86.6M	98k
Computer Science	12.8k	4,354	820	782k	1.21M	42.8M	85.5M	150k
All	**75k**	**22,697**	**3,915**	**4M**	**6.58M**	**258M**	**500M**	**610k**

Subject: Mathematics

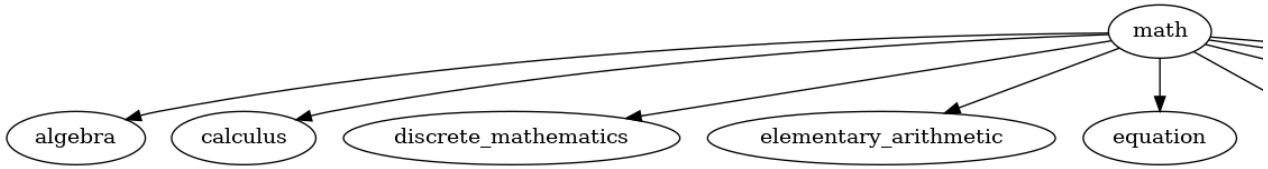
Course: Algebra

Sub-course: Multivariate equations

Knowledge Point:

- Definition of equations: video_{1}
- Application of multivariate equations: $\text{video}_{2\text{-}4}$
- How to solve linear equations: $\text{video}_{5\text{-}7}$

…..

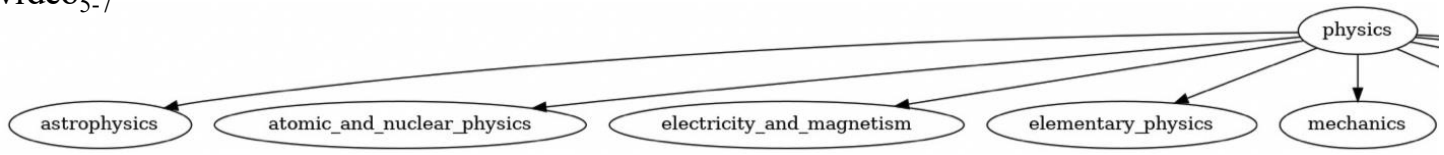


Sub-course: Functions and Equations

Knowledge Point:

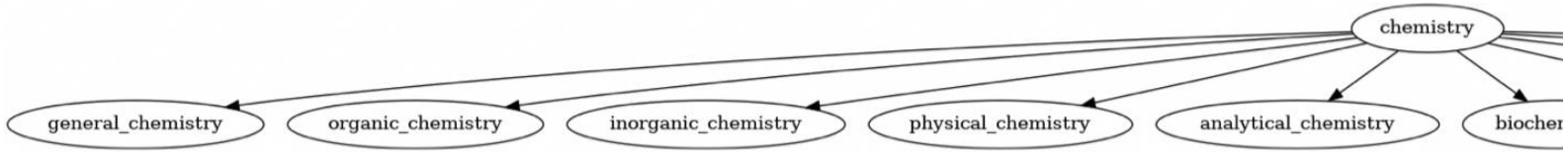
- Functions and equations: video_{8}

…..



Sub-course: …..

Course: …..



Video-to-Textbook Pipeline

Extracting coherent visual and text knowledge from the video:

1. The video has low knowledge content.
2. The video's frame has severe occlusion.
3. The teacher's explanation is too colloquial.
4. There is a lot of redundancy in video's frame.

Video-to-Textbook Pipeline

Video-Level

Video-to-ASR

Extracting

- Extract Audio
- Transcribe into text
- ASR Refining

Filtering

- Rule-based filtering
- Scoring ASR by LLM
- Discard non-instructional videos using ASR score.

ASR

Clip-Level

Long video-to-Short Clip

Extracting

- Merge incomplete ASR Segments
- Split long video based on ASR's timestamps

Filtering

- Caption each video clip
- Calculate the similarity between clip's caption and ASR
- Discard clip unrelated to ASR

Video Clip

Keyframe-Level

Clip-to-Keyframe, OCR

Extracting

- Detecting inter-frame changes
- Extracting text, symbols, and formulas using OCR.

Filtering

- Discard keyframes with object occlusion.
- Remove OCR that are identical to previous frames or lack useful info.

OCR

Multimodal Textbook

$\langle \text{frame}_1^{k_1}, \text{frame}_1^{k_2}, \text{ocr}_1, \text{asr}_1 \rangle, \text{asr}_2, \text{asr}_3, \langle \text{frame}_4^{k_1}, \text{ocr}_4, \text{asr}_4 \rangle, ..$

$\langle \text{frame}_1^{k_1}, \text{ocr}_1, \text{asr}_1 \rangle, \langle \text{frame}_2^{k_1}, \text{ocr}_2, \text{asr}_2 \rangle, \text{asr}_3, \text{asr}_4, ..$

- 75,000 instructional videos
- 2.5 years video duration (22697 hours)
- 4M video clips and 6.5M keyframes
- 259M ASR tokens and 500M OCR tokens
- 610K image-text interleaved samples

Long video: 30s~2h

Advertising video

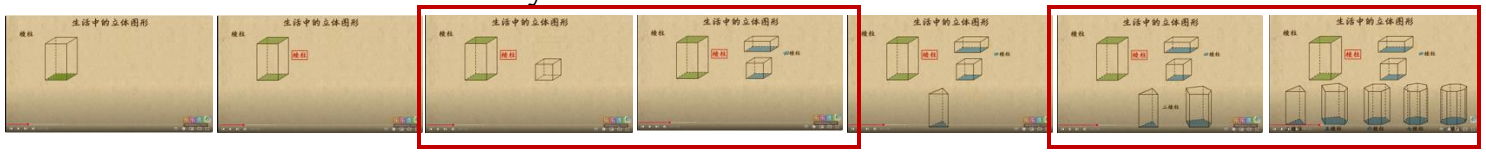
Non-teaching videos

Video clip: 15s ~ 50s

human face is too large

Too little visual knowledge

Keyframe detection and extraction



Keyframe Keyframe Keyframe Keyframe

Original ASR:

Hmm, it can be done this way. First, draw an auxiliary line here, connecting points AB. Then you can see that these two angles are equal...

ASR after polishing by LLM:

First, we can add an auxiliary line between vertices A and B, that is, we connect points A and B. At this time, we can see that the two angles are equal because...

Characteristics of Our Multimodal Textbook

Image-text relation is loose

If 'eclectic' to you is when Green Day change their guitar tone or McDonald's puts two burgers in one bun, then steer clear of this album. If however you take your pepperoni pizza with extra cream and can stomach the idea of an album with something other than one song reworked ten times, then you should buy Stimmung now." (so says some English writer type, and he ought to know…)

Lacking connection between images

The Firearm Licensing and Registration Act would establish licensing requirements to posses a firearm and ammunition, including a psychological evaluation and insurance policy. Individuals hospitalized with a mental illness would be denied a license. File photo\n\nOCEANSIDE ……

Low Knowledge Density

Dedicated to mince, peel and cut with delicacy, the slicing knives are precision tools that you have to choose with care

The high zirconium oxide content of the ceramic blade of these TB knives makes it a premium tool.

With optimum durability and everlasting sharp edge that hardly ever need sharpening, the ceramic blade of these slicing knives signed Tarrerias-Bonjean is as efficient as resistant.

The average similarity between all images in the sample

Dataset	#Image Min.	#Image Max.	#Image Avg.	#Text Token Min.	#Text Token Max.	#Text Token Avg.	In-sample Image SIM^L ↑ L=4	L=5	L=6	L=7	L=8	Avg.	Source
Image-text Paired Dataset													
COYO-700M	1	1	1	1	811	16	-	-	-	-	-	-	Common Crawl
LAION-5B	1	1	1	6	683	27	-	-	-	-	-	-	Common Crawl
Image-text Interleaved Dataset													
MMC4	0	117	5.7	4	16715	417	0.363	0.348	0.310	0.298	0.276	0.319	Common Crawl
MMC4-core-ff	0	15	4.1	15	16715	329	0.431	0.406	0.404	0.403	0.396	0.407	Common Crawl
OBELICS	1	30	2.5	12	10717	816	0.366	0.351	0.339	0.337	0.336	0.345	Common Crawl
OmniCorpus*	1	16	3.9	14	6893	574	0.358	0.329	0.310	0.305	0.301	0.321	Multi-sources
Ours	**2**	45	**10.7**	11	**34174**	**1927**	**0.687**	**0.697**	**0.698**	**0.688**	**0.662**	**0.686**	Video Website

Number of both images and texts in our corpus is larger.

Connection between images is closer.

- **Previous Interleaved dataset (left):**
  - Mostly crawled from websites, such as Wikipedia
  - A small number of images
  - Low connections between images and texts
  - Lack of logical relations between images
  - Low knowledge density
- Our textbook-6.5M:
  - Derived from teaching videos
  - Each sample contains a greater number of image and text tokens
  - The connections between images are closer
  - High knowledge density and can be organized by category

Pre-training on our textbook dataset, from easy to difficult

Multimodal Large Language Models

<keyframe1> <ASR1> <OCR1> <keyframe2> <ASR2> <OCR2> <keyframe3> <ASR3> <OCR3> <Comment> ...

From easy to difficult

Action Teaching

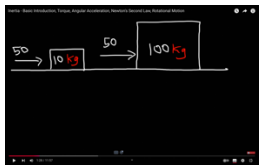
A man is cooking. First, he cuts some vegetables and puts them into the pot.

Then he took a shovel and stirred it in the hot pot, watching as the pot gradually began to emit hot steam.

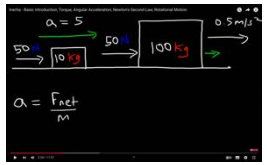
After that, he took some seasonings and added them to the hot pot.

.....

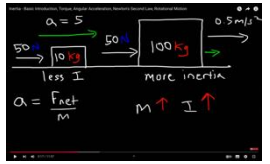
Physics. Newtonian mechanics



To illustrate the concept of inertia... The mass of the first object is 10 kilograms, while the mass of the second object is 100 kilograms...



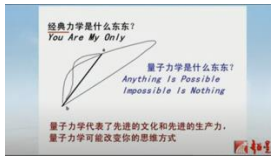
According to Newton's second law, the resultant force acting on an object is equal to the product of its mass and acceleration...



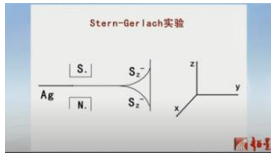
You can see in the figure that lighter objects have much greater acceleration, while heavier objects have much smaller acceleration…..

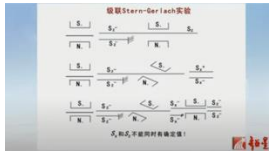
Like!! The junior high school physics class taught by this teacher is very clear, and it helped me master the concept of acceleration....

Physics. Rigid Body Mechanics

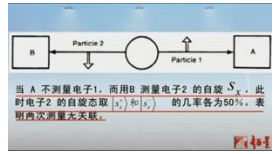


First, let's introduce the differences between classical mechanics ….



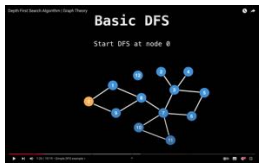


.....



Physics. Quantum Mechanics

Computer. Algorithm

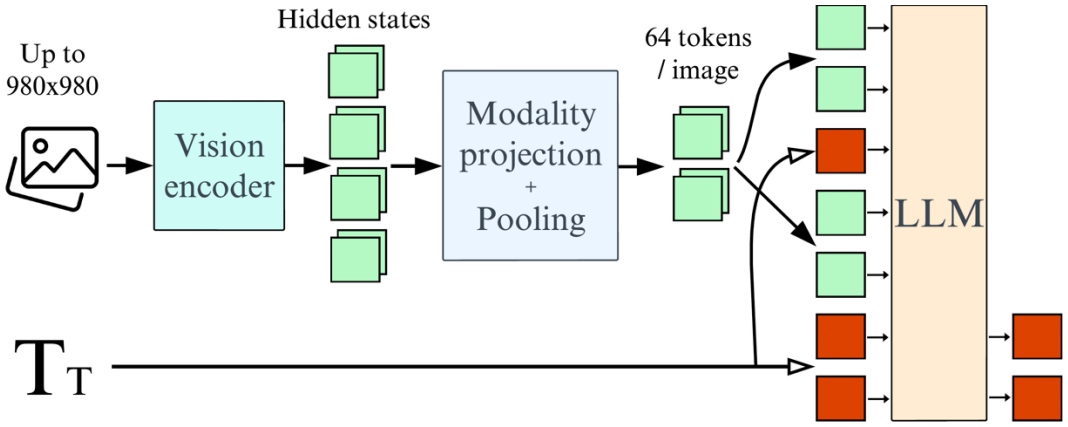


Depth-First Search (DFS) works by selecting the next node to explore until it can no longer proceed, at which point …..

At node 8, we arbitrarily choose an edge and proceed to node 7. At node 7, there are multiple edges to….

n = number of nodes in the graph\ng = adjacency list representing graph\nvisited = [false, …, false] # size n\n\nfunction dfs(at):\n if visited[at]: return\n visited[at] = true\n\n neighbours = graph[at]\n for next in neighbours:\n dfs(next)\n\n# Start DFS at node zero\nstart_node = 0\ndfs(start_node)"

Pre-training on our textbook dataset, from easy to difficult



Zero-shot Evaluation:

A baby wants to know what's inside the cabinet. (From the ScienceQA)
Question: What type of force should the baby's hand use to open the cabinet door?
Options: A. Pull B. Push

Qwen2-VL-base: The image shows a boy holding the cabinet door with one hand. To open the cabinet door, the little boy should push the cabinet hard. I should choose B. Push
After pre-training with Textbook-6.5M: The image shows .... If he wants to open the cabinet door, he should pull it outward …. the cabinet door will slowly open with an increasingly larger angle, and…. My choice is A. Pull

Continual pre-training

- Continual pre-training on our textbook-6.5M dataset
- Training process is organized according to the difficulty level of knowledge points: from easy to difficult

Evaluation: Few-shot settings

- 2 general VQA test benchmarks (TextVQA, OKVQA)
- 3 multimodal reasoning benchmarks (MathVista, MathVision, MathVision)
- 1 multimodal knowledge benchmark (ScienceQA)

Experiment: Pre-training Performance

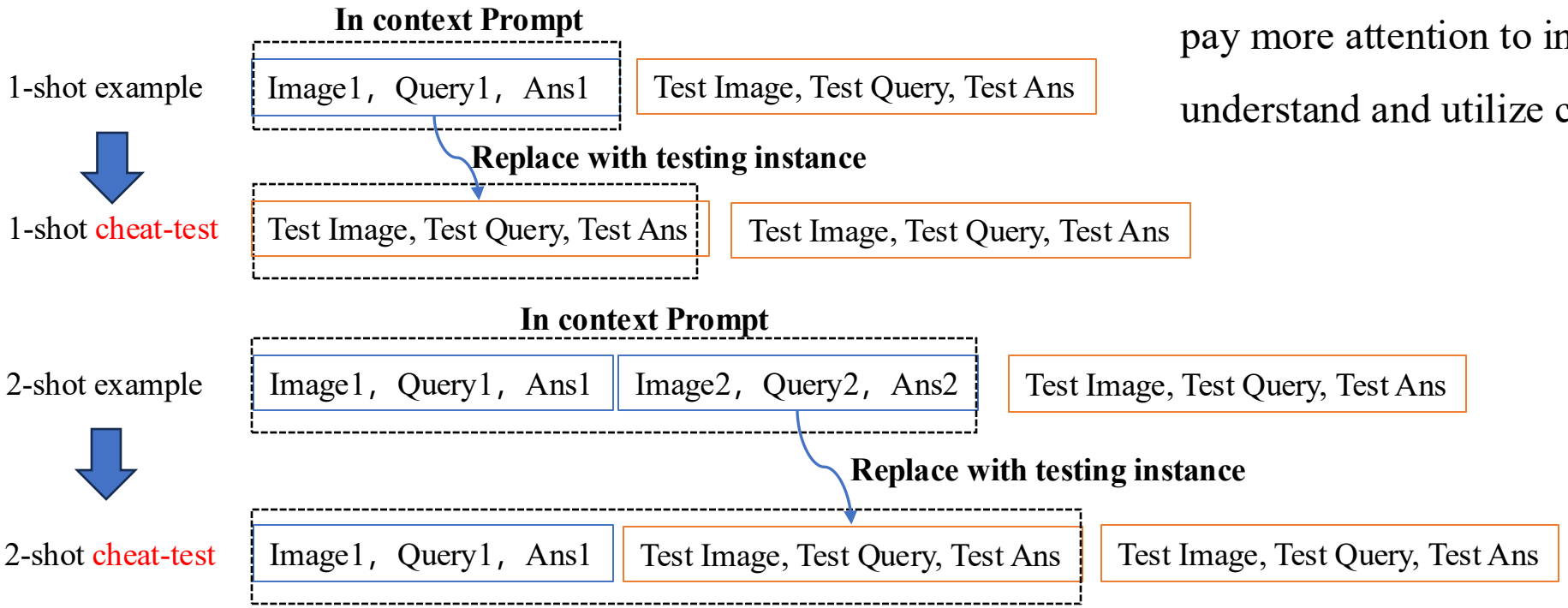
#Shot	0	1	2	4	0	1	2	4	0	1	2	4	0	1	2	4
Dataset	ScienceQA^{IMG}				OKVQA				TextVQA				TextVQA^{ocr}			
MMC4	-	1.6	3.9	11.6	8.6	23.6	21.5	28.7	12.1	16.2	16.8	20.9	14.5	23.9	29.9	34.7
MMC4-Core-ff	-	2.1	10.1	10.2	11.8	21.2	25.3	30.4	**13.6**	18.7	18.8	22.1	**16.1**	26.6	28.7	33.1
OBELICS	-	2.8	3.0	16.4	**13.0**	**31.7**	35.7	37.5	9.2	26.5	30.2	32.2	11	30.7	36.3	41
Textbook-6.5M	**26.3**	**29.4**	**25.1**	**37.3**	10.2	31.2	**36.8**	**39.9**	11.8	**26.7**	**32.1**	**33.5**	14.1	**33.1**	**36.4**	**42.8**
Dataset	MathVista				MathVision				MathVerse				Avg.			
MMC4	20.4	30	27.9	26	12.2	21.3	15.5	16.1	8.6	19.4	21.2	**15.9**	10.9	19.4	19.5	21.9
MMC4-Core-ff	22.5	33.0	29.2	27.8	13.7	23.4	16.3	17.7	**8.6**	19.9	**21.8**	15.2	12.3	20.7	21.4	22.3
OBELICS	21.6	28.5	31.1	27.6	13.4	20.1	16.8	14.9	6.9	19.4	20.7	14	10.7	22.8	24.8	26.2
Textbook-6.5M	**24.3**	**43.4**	**33.2**	**29.2**	**14.5**	**25.6**	**18.2**	**18.1**	7.7	**28.5**	19.8	14.6	**15.5**	**31.1**	**28.8**	**30.8**

- Our textbook has significantly improved the performance on **knowledge and reasoning-oriented test benchmarks.**
- The pretraining that interweaves images and text enhances the **in-context learning ability** of multimodal models.
- Textbook corpora with coherent contexts enable VLMs to pay more attention to the input multimodal context and better **understand and utilize the clues** in the input context.

Analysis Experiment

Dataset	OKVQA	TextVQA	Mathvista	Mathvision	Mathverse
1-shot Cheat: Example:$\{I_t, q_t, a_t\}$ + Test-case: I_t, q_t					
MMC4-cf	69.0	41.0	72.6	69.3	55.7
OBELICS	71.5	43.8	67.7	66.5	62.8
Ours	**79.2**	**51.9**	**94.1**	**98.4**	**76.8**
2-shot Cheat: Example:$\{I_t, q_t, a_t\}$, $\{I_e, q_e, a_e\}$+Test-case: I_t, q_t					
MMC4-Cf	53.5	39.2	55.7	51.9	40.8
OBELICS	71.3	42.8	56.7	39.9	39.5
Ours	**84.3**	**49.4**	**77.1**	**70.7**	**63.1**

- We designed a cheat test to test whether VLMs can truly pay attention to the interleaved context
- Cheat-test: We replace one of the examples in the few-shot examples with a test sample, and in this case, the theoretical accuracy of VLMs should be close to 100%.
- Cheat-test shows that VLMs pretraining from ours can pay more attention to input multimodal context and better understand and utilize clues in the prompt.



Future Works

- **Textbook-level multimodal corpora** →  **Create pretraining corpora that include basic knowledge of various disciplines**

Pretraining

- **Pre-training from easy to difficult, with interleaved images and text** →  **Similar to humans, learn basic knowledge in a way that combines pictures and texts**

SFT

- **Synthesize diverse exercise instruction and fine-tune** →  **Design practice questions around knowledge points to consolidate knowledge**

Build multimodal disciplinary large models

Build a unified LLMs for generation and understanding of any modality

Design a better world model

- Design multimodal disciplinary corpora to pre-train VLMs, enabling them to learn professional knowledge in a natural and image-text interleaved manner.
- Collecting massive online educational videos and converting them into a dataset where key image frames and textual explanations are interleaved, this textbook provides a more coherent and interconnected learning context, supplementing the traditional image-text alignment methods.
- After pre-training on multimodal textbooks, VLMs have enhanced their context awareness and disciplinary reasoning abilities.

Paper **Code** **Dataset** **WeChat**