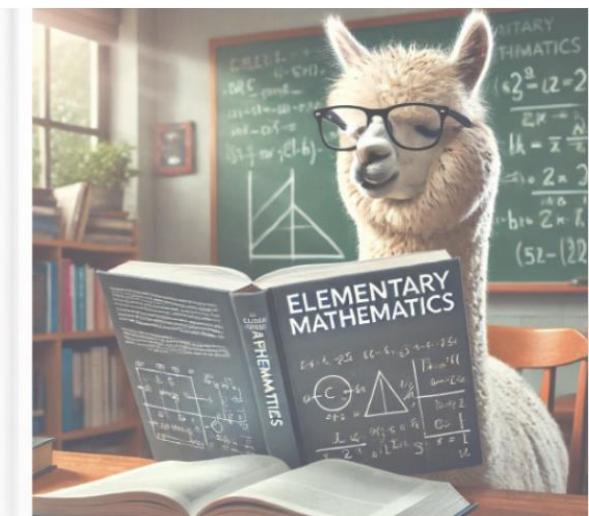
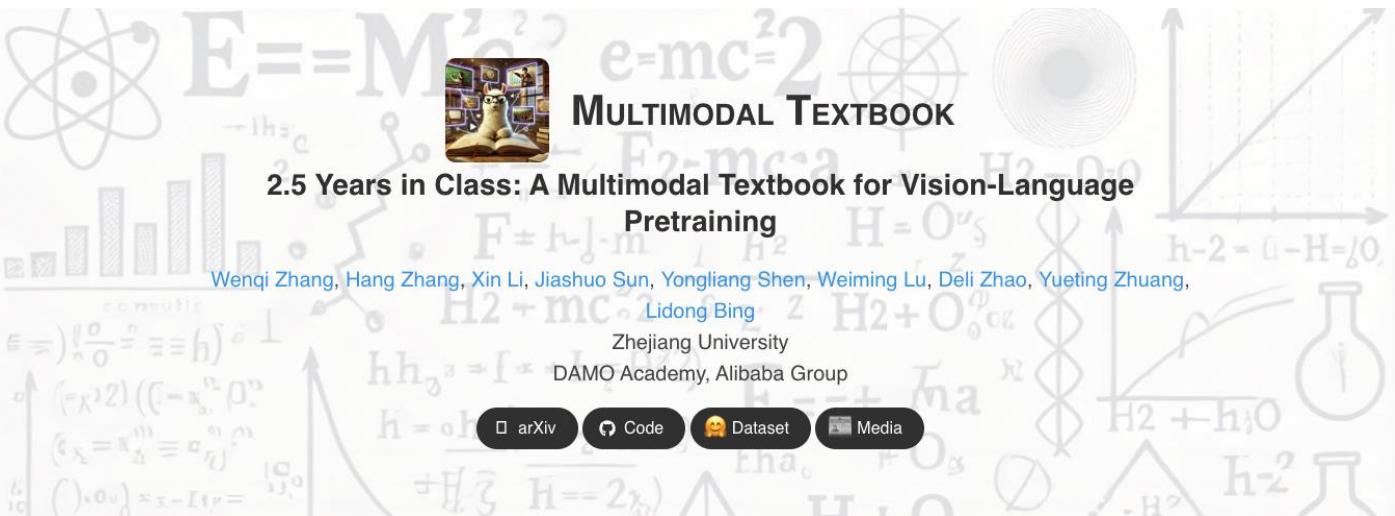
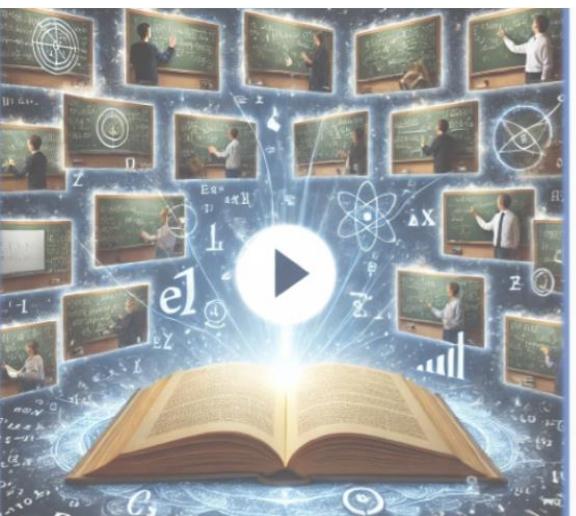


2.5 Years in Class: A Multimodal Textbook for Vision-Language Pretraining

ICCV 2025 Highlight

zhangwenqi@zju.edu.cn





Online Instructional Videos

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



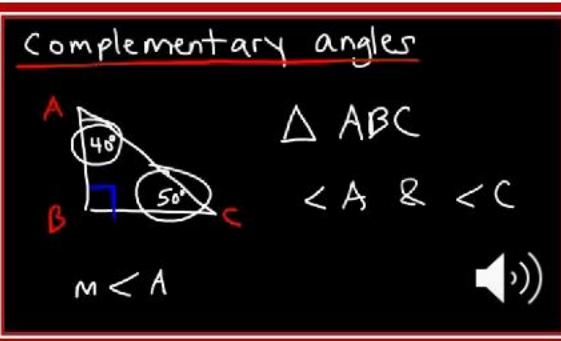
High-quality Corpus

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

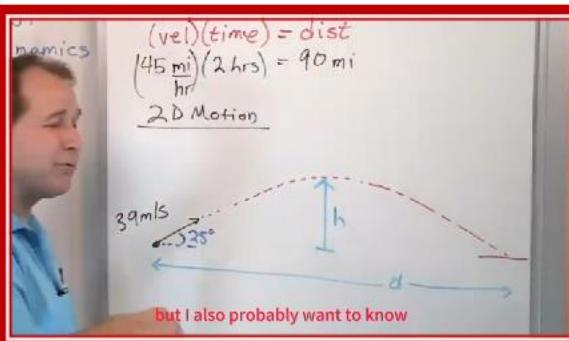


Vision-Language Models

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

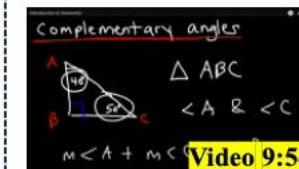
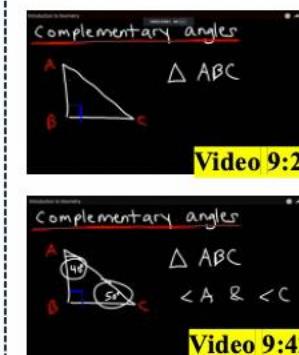


Video-to-Textbook



«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.....



these richly contextualized and professionally curated
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.



The figure shows the Hugging Face Model Hub interface. It features three main sections: 'Models', 'Spaces', and 'Datasets'. The 'Models' section on the left lists repositories like 'microsoft/phi-4', 'hexgrad/Kokoro-82M', 'deepseek-ai/DeepSeek-V3', 'NovaSky-AI/Sky-T1-32B-Preview', and 'black-forest-labs/FLUX.1-dev'. The 'Spaces' section in the center lists applications like 'Kokoro TTS', 'TRELLIS', 'IC Light V2', '2024 AI Timeline', and 'Stable Point-Aware 3D'. The 'Datasets' section on the right lists datasets like 'fka/awesome-chatgpt-prompts', 'DAMO-NLP-SG/multimodal_textbook' (which is highlighted with a red box), 'NovaSky-AI/Sky-T1_data_17k', 'cfahlgren1/react-code-instructions', and 'HumanLLMs/Human-Like-DPO-Dataset'. Each item in the lists includes the repository name, last updated time, commit count, and star count.

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

你已转帖
merve @mervenoyann · 1月10日
Alibaba released Multimodal Textbook: a new multimodal pre-training set from online instructional videos (22k hours)  

6,5M images interleaved with 800K text on math, physics, chemistry 

Our Multimodal Textbook



Massive Instructional Videos
22000 Class Hours
2.5 Years Duration

Multi-Level Extraction & Filtering

Keyframe ASR & OCR

High-quality Corpus
More clean image-text relation
Rich visual and textual knowledge
More coherent image sequences

Textbook-Level interleaved Dataset

Textbook: Mathematics

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

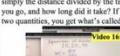
Let's consider a right triangle, and we will label it as triangle ABC. The symbol for this triangle is as follows: triangle ABC.

Complementary angles 

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.

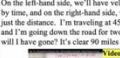
Complementary angles 

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.

Tutorial Text Extract From Video: 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 

On the left-hand side, we'll have velocity multiplied by just time, and on the right-hand side, we'll be left with just distance. I'm going to run at 5 miles per hour, and I've gone 20 miles. Now, if I run for two hours, how far will I have gone? It's clear for two hours, how far will I have gone?

Video 16:19

Tutorial Text Extract From Video: The Appalachians Mountains in eastern North America contain the remains and the remains of the shells of marine animals. 

These animals lived in a shallow ocean more than 400 million years ago. Around 300 million years ago,

Video 2:13

Tutorial Text Extract From 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,

Textbook: Physics

Video 9:42

Textbook: Earth Science

Video 1:58

Textbook: Engineering

Video 6:11

Video 6:17

Video 18:16

Video 18:18

Video 18:19

Video 18:20

Video 18:21

Video 18:22

Video 18:23

Video 18:24

Video 18:25

Video 18:26

Video 18:27

Video 18:28

Video 18:29

Video 18:30

Video 18:31

Video 18:32

Video 18:33

Video 18:34

Video 18:35

Video 18:36

Video 18:37

Video 18:38

Video 18:39

Video 18:40

Video 18:41

Video 18:42

Video 18:43

Video 18:44

Video 18:45

Video 18:46

Video 18:47

Video 18:48

Video 18:49

Video 18:50

Video 18:51

Video 18:52

Video 18:53

Video 18:54

Video 18:55

Video 18:56

Video 18:57

Video 18:58

Video 18:59

Video 18:60

Video 18:61

Video 18:62

Video 18:63

Video 18:64

Video 18:65

Video 18:66

Video 18:67

Video 18:68

Video 18:69

Video 18:70

Video 18:71

Video 18:72

Video 18:73

Video 18:74

Video 18:75

Video 18:76

Video 18:77

Video 18:78

Video 18:79

Video 18:80

Video 18:81

Video 18:82

Video 18:83

Video 18:84

Video 18:85

Video 18:86

Video 18:87

Video 18:88

Video 18:89

Video 18:90

Video 18:91

Video 18:92

Video 18:93

Video 18:94

Video 18:95

Video 18:96

Video 18:97

Video 18:98

Video 18:99

Video 18:100

Video 18:101

Video 18:102

Video 18:103

Video 18:104

Video 18:105

Video 18:106

Video 18:107

Video 18:108

Video 18:109

Video 18:110

Video 18:111

Video 18:112

Video 18:113

Video 18:114

Video 18:115

Video 18:116

Video 18:117

Video 18:118

Video 18:119

Video 18:120

Video 18:121

Video 18:122

Video 18:123

Video 18:124

Video 18:125

Video 18:126

Video 18:127

Video 18:128

Video 18:129

Video 18:130

Video 18:131

Video 18:132

Video 18:133

Video 18:134

Video 18:135

Video 18:136

Video 18:137

Video 18:138

Video 18:139

Video 18:140

Video 18:141

Video 18:142

Video 18:143

Video 18:144

Video 18:145

Video 18:146

Video 18:147

Video 18:148

Video 18:149

Video 18:150

Video 18:151

Video 18:152

Video 18:153

Video 18:154

Video 18:155

Video 18:156

Video 18:157

Video 18:158

Video 18:159

Video 18:160

Video 18:161

Video 18:162

Video 18:163

Video 18:164

Video 18:165

Video 18:166

Video 18:167

Video 18:168

Video 18:169

Video 18:170

Video 18:171

Video 18:172

Video 18:173

Video 18:174

Video 18:175

Video 18:176

Video 18:177

Video 18:178

Video 18:179

Video 18:180

Video 18:181

Video 18:182

Video 18:183

Video 18:184

Video 18:185

Video 18:186

Video 18:187

Video 18:188

Video 18:189

Video 18:190

Video 18:191

Video 18:192

Video 18:193

Video 18:194

Video 18:195

Video 18:196

Video 18:197

Video 18:198

Video 18:199

Video 18:200

Video 18:201

Video 18:202

Video 18:203

Video 18:204

Video 18:205

Video 18:206

Video 18:207

Video 18:208

Video 18:209

Video 18:210

Video 18:211

Video 18:212

Video 18:213

Video 18:214

Video 18:215

Video 18:216

Video 18:217

Video 18:218

Video 18:219

Video 18:220

Video 18:221

Video 18:222

Video 18:223

Video 18:224

Video 18:225

Video 18:226

Video 18:227

Video 18:228

Video 18:229

Video 18:230

Video 18:231

Video 18:232

Video 18:233

Video 18:234

Video 18:235

Video 18:236

Video 18:237

Video 18:238

Video 18:239

Video 18:240

Video 18:241

Video 18:242

Video 18:243

Video 18:244

Video 18:245

Video 18:246

Video 18:247

Video 18:248

Video 18:249

Video 18:250

Video 18:251

Video 18:252

Video 18:253

Video 18:254

Video 18:255

Video 18:256

Video 18:257

Video 18:258

Video 18:259

Video 18:260

Video 18:261

Video 18:262

Video 18:263

Video 18:264

Video 18:265

Video 18:266

Video 18:267

Video 18:268

Video 18:269

Video 18:270

Video 18:271

Video 18:272

Video 18:273

Video 18:274

Video 18:275

Video 18:276

Video 18:277

Video 18:278

Video 18:279

Video 18:280

Video 18:281

Video 18:282

Video 18:283

Video 18:284

Video 18:285

Video 18:286

Video 18:287

Video 18:288

Video 18:289

Video 18:290

Video 18:291

Video 18:292

Video 18:293

Video 18:294

Video 18:295

Video 18:296

Video 18:297

Video 18:298

Video 18:299

Video 18:300

Video 18:301

Video 18:302

Video 18:303

Video 18:304

Video 18:305

Video 18:306

Video 18:307

Video 18:308

Video 18:309

Video 18:310

Video 18:311

Video 18:312

Video 18:313

Video 18:314

Video 18:315

Video 18:316

Video 18:317

Video 18:318

Video 18:319

Video 18:320

Video 18:321

Video 18:322

Video 18:323

Video 18:324

Video 18:325

Video 18:326

Video 18:327

Video 18:328

Video 18:329

Video 18:330

Video 18:331

Video 18:332

Video 18:333

Video 18:334

Video 18:335

Video 18:336

Video 18:337

Video 18:338

Video 18:339

Video 18:340

Video 18:341

Video 18:342

Video 18:343

Video 18:344

Video 18:345

Video 18:346

Video 18:347

Video 18:348

Video 18:349

Video 18:350

Video 18:351

Video 18:352

Video 18:353

Video 18:354

Video 18:355

Video 18:356

Video 18:357

Video 18:358

Video 18:359

Video 18:360

Video 18:361

Video 18:362

Video 18:363

Video 18:364

Video 18:365

Video 18:366

Video 18:367

Video 18:368

Video 18:369

Video 18:370

Video 18:371

Video 18:372

Video 18:373

Video 18:374

Video 18:375

Video 18:376

Video 18:377

Video 18:378

Video 18:379

Video 18:380

Video 18:381

Video 18:382

Video 18:383

Video 18:384

Video 18:385

Video 18:386

Video 18:387

Video 18:388

Video 18:389

Video 18:390

Video 18:391

Video 18:392

Video 18:393

Video 18:394

Video 18:395

Video 18:396

Video 18:397

Video 18:398

Video 18:399

Video 18:400

Video 18:401

Video 18:402

Video 18:403

Video 18:404

Video 18:405

Video 18:406

Video 18:407

Video 18:408

Video 18:409

Video 18:410

Video 18:411

Video 18:412

Video 18:413

Video 18:414

Video 18:415

Video 18:416

Video 18:417

Video 18:418

Video 18:419

Video 18:420

Video 18:421

Video 18:422

Video 18:423

Video 18:424

Video 18:425

Video 18:426

Video 18:427

Video 18:428

Video 18:429

Video 18:430

Video 18:431

Video 18:432

Video 18:433

Video 18:434

Video 18:435

Video 18:436

Video 18:437

Video 18:438

Video 18:439

Video 18:440

Video 18:441

Video 18:442

Video 18:443

Video 18:444

Video 18:445

Video 18:446

Video 18:447

Video 18:448

Video 18:449

Video 18:450

Video 18:451

Video 18:452

Video 18:453

Video 18:454

Video 18:455

Video 18:456

Video 18:457

Video 18:458

Video 18:459

Video 18:460

Video 18:461

Video 18:462

Video 18:463

Video 18:464

Video 18:465

Video 18:466

Video 18:467

Video 18:468

Video 18:469

Video 18:470

Video 18:471

Video 18:472

Video 18:473

Video 18:474

Video 18:475

Video 18:476

Video 18:477

Video 18:478

Video 18:479

Video 18:480

Video 18:481

Video 18:482

Video 18:483

Video 18:484

Video 18:485

Video 18:486

Video 18:487

Video 18:488

Video 18:489

Video 18:490

Video 18:491

Video 18:492

Video 18:493

Video 18:494

Video 18:495

Video 18:496

Video 18:497

Video 18:498

Video 18:499

Video 18:500

Video 18:501

Video 18:502

Video 18:503

Video 18:504

Video 18:505

Video 18:506

Video 18:507

Video 18:508

Video 18:509

Video 18:510

Video 18:511

Video 18:512

Video 18:513

Video 18:514

Video 18:515

Video 18:516

Video 18:517

Video 18:518

Video 18:519

Video 18:520

Video 18:521

Video 18:522

Video 18:523

Video 18:524

Video 18:525

Video 18:526

Video 18:527

Video 18:528

Video 18:529

Video 18:530

Video 18:531

Video 18:532

Video 18:533

Video 18:534

Video 18:535

Video 18:536

Video 18:537

Video 18:538

Video 18:539

Video 18:540

Video 18:541

Video 18:542

Video 18:543

Video 18:544

Video 18:545

Video 18:546

Video 18:547

Video 18:548

Video 18:549

Video 18:550

Video 18:551

Video 18:552

Video 18:553

Video 18:554

Video 18:555

Video 18:556

Video 18:557

Video 18:558

Video 18:559

Video 18:560

Video 18:561

Video 18:562

Video 18:563

Video 18:564

Video 18:565

Video 18:566

Video 18:567

Video 18:568

Video 18:569

Video 18:570

Video 18:571

Video 18:572

Video 18:573

Video 18:574

Video 18:575

Video 18:576

Video 18:577

Video 18:578

Video 18:579

Video 18:580

Video 18:581

Video 18:582

Video 18:583

Video 18:584

Video 18:585

Video 18:586

Video 18:587

Video 18:588

Video 18:589

Video 18:590

Video 18:591

Video 18:592

Video 18:593

Video 18:594

Video 18:595

Video 18:596

Video 18:597

Video 18:598

Video 18:599

Video 18:600

Video 18:601

Video 18:602

Video 18:603

Video 18:604

Video 18:605

Video 18:606

Video 18:607

Video 18:608

Video 18:609

Video 18:610

Video 18:611

Video 18:612

Video 18:613

Video 18:614

Video 18:615

Video 18:616

Video 18:617

Video 18:618

Video 18:619

Video 18:620

Video 18:621

Video 18:622

Video 18:623

Video 18:624

Video 18:625

Video 18:626

Video 18:627

Video 18:628

Video 18:629

Video 18:630

Video 18:631

Video 18:632

Video 18:633

Video 18:634

Video 18:635

Video 18:636

Video 18:637

Video 18:638

Video 18:639

Video 18:640

Video 18:641

Video 18:642

Video 18:643

Video 18:644

Video 18:645

Video 18:646

Video 18:647

Video 18:648

Video 18:649

Video 18:650

Video 18:651

Video 18:652

Video 18:653

Video 18:654

Video 18:655

Video 18:656

Video 18:657

Video 18:658

Video 18:659

Video 18:660

Video 18:661

Video 18:662

Video 18:663

Video 18:664

Video 18:665

Video 18:666

Video 18:667

Video 18:668

Video 18:669

Video 18:670

Video 18:671

Video 18:672

Video 18:673

Video 18:674

Video 18:675

Video 18:676

Video 18:677

Video 1

Recent discussions on scaling up have sparked widespread debates, with some claiming that "scale up is dead." We argue, however, that high-quality data is the true key to effective scaling, particularly textbook-grade, high-quality knowledge corpora. In our recent work, we

显示更多

Online Instructional Videos

Complementary angles

STEAL THIS!

Learn To Program

Physics & Geometry

CREATE AN ONLINE CLASS IN MINUTES!

You Should Be Able To Answer These

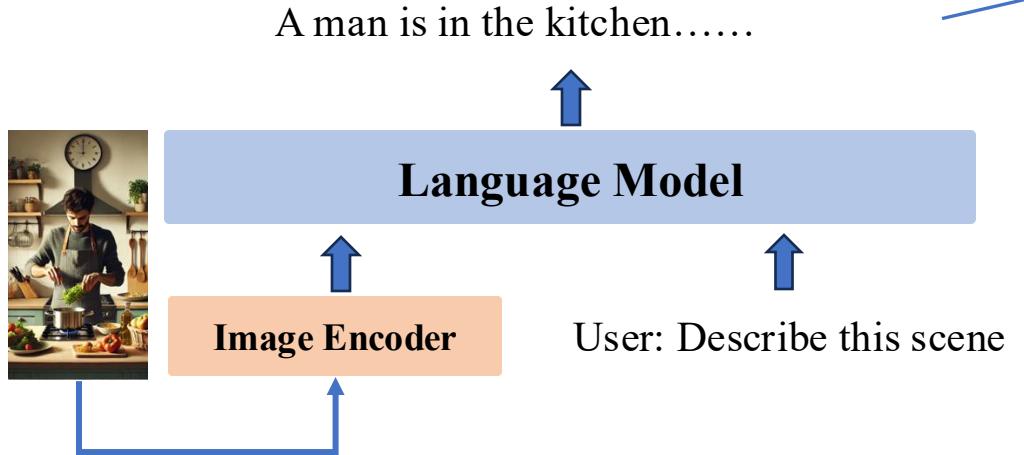
Breadth First Search

collect instructional videos across various disciplines from the Internet.

0:59

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

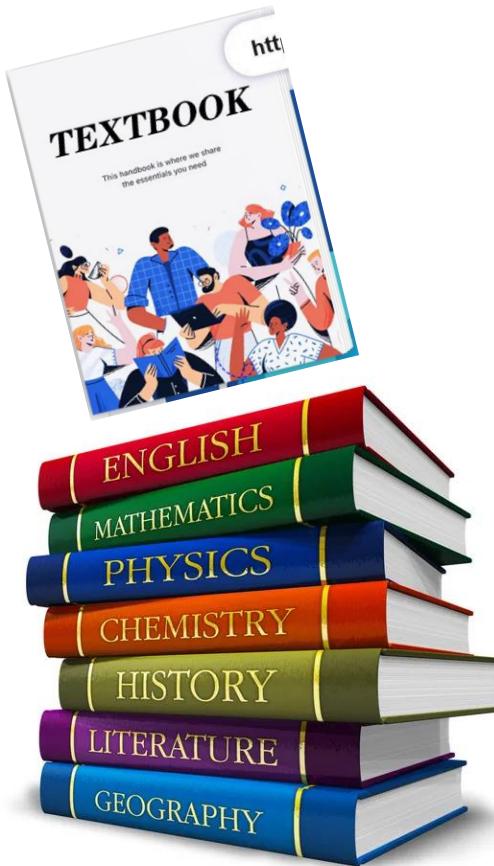
➤ Simple descriptive text

➤ Only statically describe images

➤ Difficult to collect

➤ Dynamically describe continuous actions and complex processes

What should the human learning process be like?



Characteristics of Life

2

2.0 CHAPTER PREVIEW

In this chapter we will discuss:

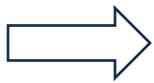
- The meaning of "organism."
- The properties that all living things share.
- The common things that almost all organisms on earth require to live.

2.1 OVERVIEW

It may seem silly, but when studying life science, one must know which things are alive and which things are not. You can often tell what is alive and what is not by looking at that thing. We use the term "organism" to describe things that are alive. You know that a rock is not alive, but it's not an organism, but a tree, a tree, or a robin is alive, so they are organisms. But what about yeast that makes bread rise? (It is an organism) and very often, people think that yeast is not alive. But if you look at the properties of "alive," one must know what is and is not alive. All organisms share common properties whether they are single cell organisms, such as bacteria and yeasts, or multicellular organisms, such as plants and human beings. Let's dive in and learn what they are.

Figure 2.1.1

It's easy to tell if something is alive—if it's an organism—just by looking at it, and we can also often tell by looking at a dead organism that it was once alive. But with some organisms, it's not so easy. For example, coral looks like a rock, but it's not a rock, it's a living organism. It looks (and feels) like a rock, but it's not, it's a colony of tiny, tiny organisms that live together. So, how do we tell if something is alive? Well, scientists outline the common properties that all living things—*all organisms*—have.



- 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 possess 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 ...

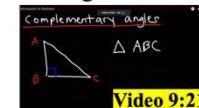


The high zirconium oxide content of the ceramic blade of these TB knives makes it a premium tool ...

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**....



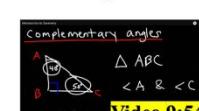
Video 9:21

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



Video 9:42

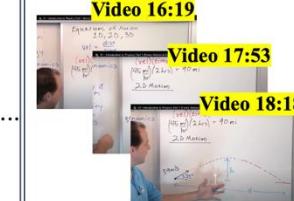
angle A measures 40 degrees and angle C measures 50 degrees. In this case, we can say that angle A



Video 9:54

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

«Textbook: Physics»



Video 16:19
Video 17:53
Video 18:18

<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»



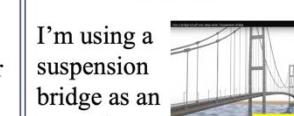
Video 1:58
Video 2:13

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



Formation of the Appalachian Mountains
- 400 Million Years Ago
Laurentia
- 200 Million Years Ago
Gondwana
Assembly of Pangaea

«Textbook: Engineering»



Video 0:11
Video 0:27

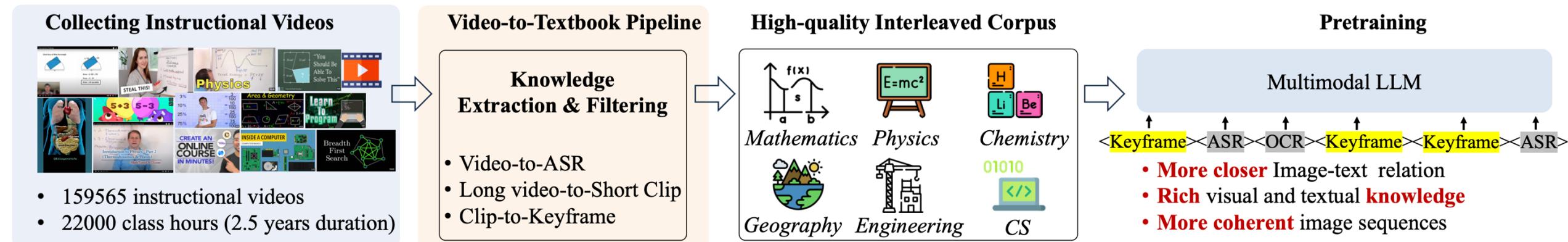
I'm using a suspension bridge as an example...



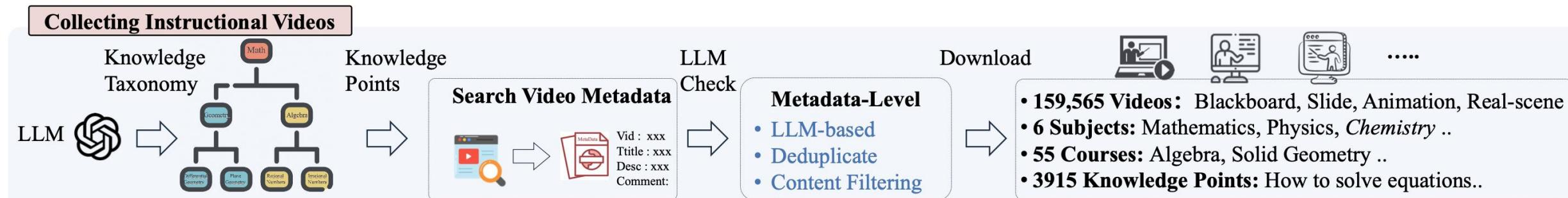
The parts of a suspension bridge include towers

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

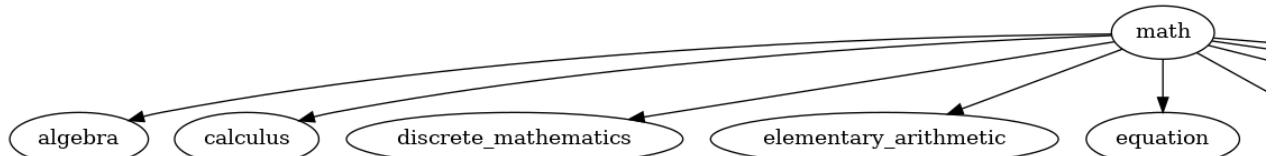
Course: Algebra

Sub-course: Multivariate equations

Knowledge Point:

- Definition of equations: video₁
- Application of multivariate equations: video₂₋₄
- How to solve linear equations: video₅₋₇

.....

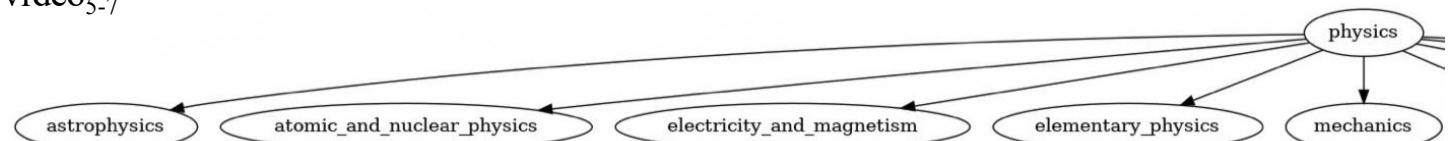


Sub-course: Functions and Equations

Knowledge Point:

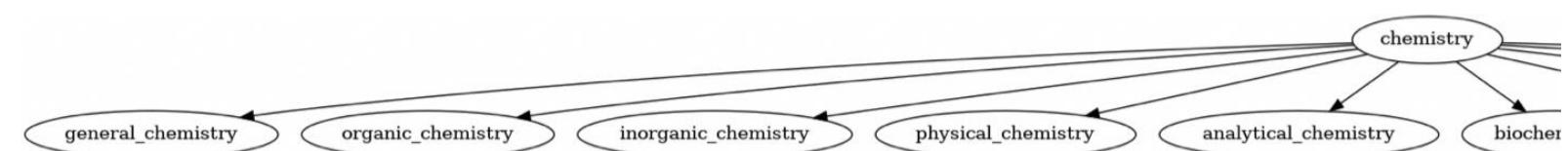
- Functions and equations: video₈

.....



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.



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



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.



Multimodal Textbook

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

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

- 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

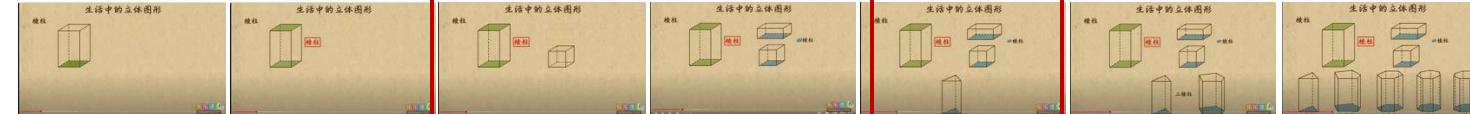


Advertising video Non-teaching videos

Video clip: 15s ~ 50s

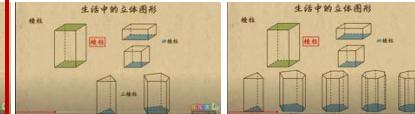


human face is too large Too little visual knowledge



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 A and B. 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

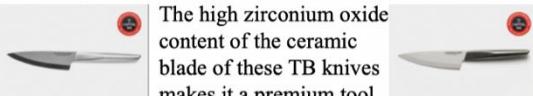


The Firearm Licensing and Registration Act would establish licensing requirements to possess a firearm and ammunition, including a psychological evaluation and insurance policy. Individuals hospitalized with a mental illness would be denied a license. File photo\n\nOCEANSIDE



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			#Text Token			In-sample Image $SIM^L \uparrow$					Source	
	Min.	Max.	Avg.	Min.	Max.	Avg.	$L=4$	$L=5$	$L=6$	$L=7$	$L=8$		
<i>Image-text Paired Dataset</i>													
COYO-700M	1	1	1	1	811	16	-	-	-	-	-	-	Common Crawl
LAION-5B	1	1	1	6	683	27	-	-	-	-	-	-	Common Crawl
<i>Image-text Interleaved Dataset</i>													
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):

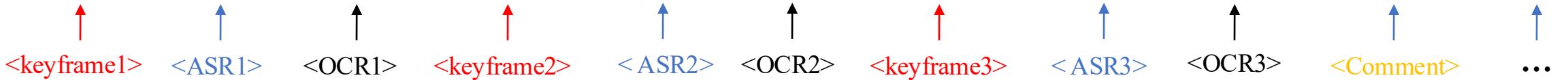
- Mostly crawled from websites, such as Wikipedia
- A small number of images
- Low connections between images and texts
- Lack of logical relations between images
- Low knowledge density

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



Action Teaching



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



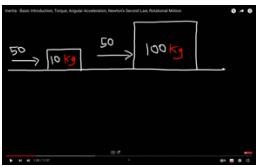
Then he took a shovel and stirred it in the hot pot, watching as the pot gradually began to emit hot steam.



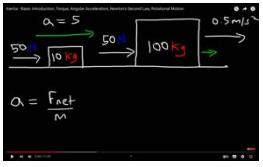
After that, he took some seasonings and added them to the hot pot.

.....

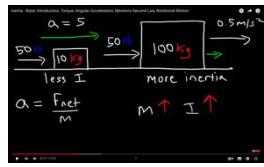
Physics. Newtonian mechanics



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...



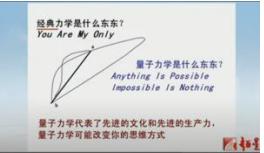
According to Newton's second law, the resultant force acting on an object is equal to the product of its mass and acceleration...



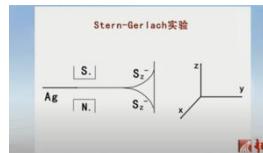
You can see in the figure Like!! The junior high school physics class, taught by this teacher is very clear, and it helped me master the concept of acceleration....

From easy to difficult

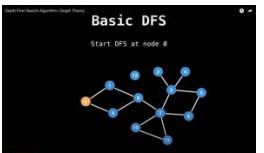
Physics. Rigid Body Mechanics



First, let's introduce the differences between classical mechanics



Physics. Quantum Mechanics



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



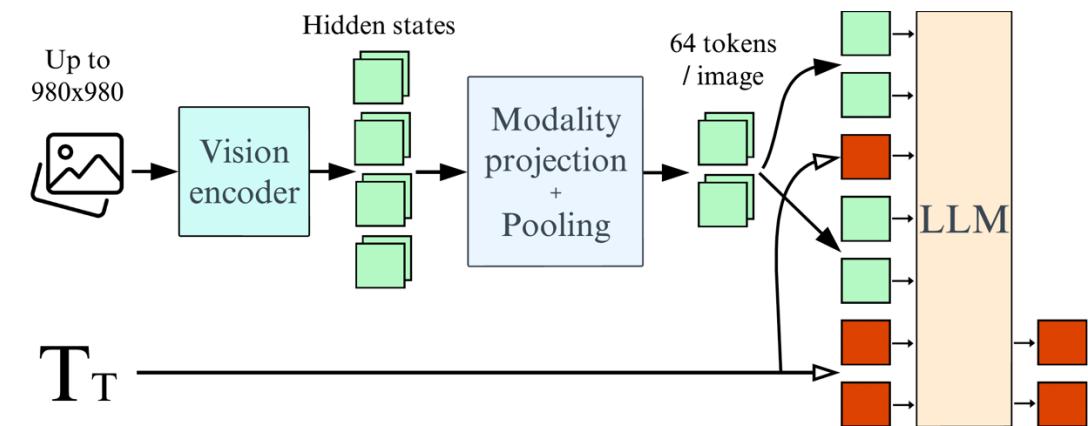
At node 8, we arbitrarily choose an edge and proceed to node 7. At node 7, there are multiple edges to....



n = number of nodes in the graph
 ng = adjacency list representing graph
 $nvisted$ = [false, ..., false]
size n
function dfs(at):
 if visited[at] == true:
 return
 visited[at] = true
 for next in neighbours:
 if nvisted[next] == false:
 dfs(next)
Start DFS at node zero
start_node = 0
dfs(start_node)"

Computer. Algorithm

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)

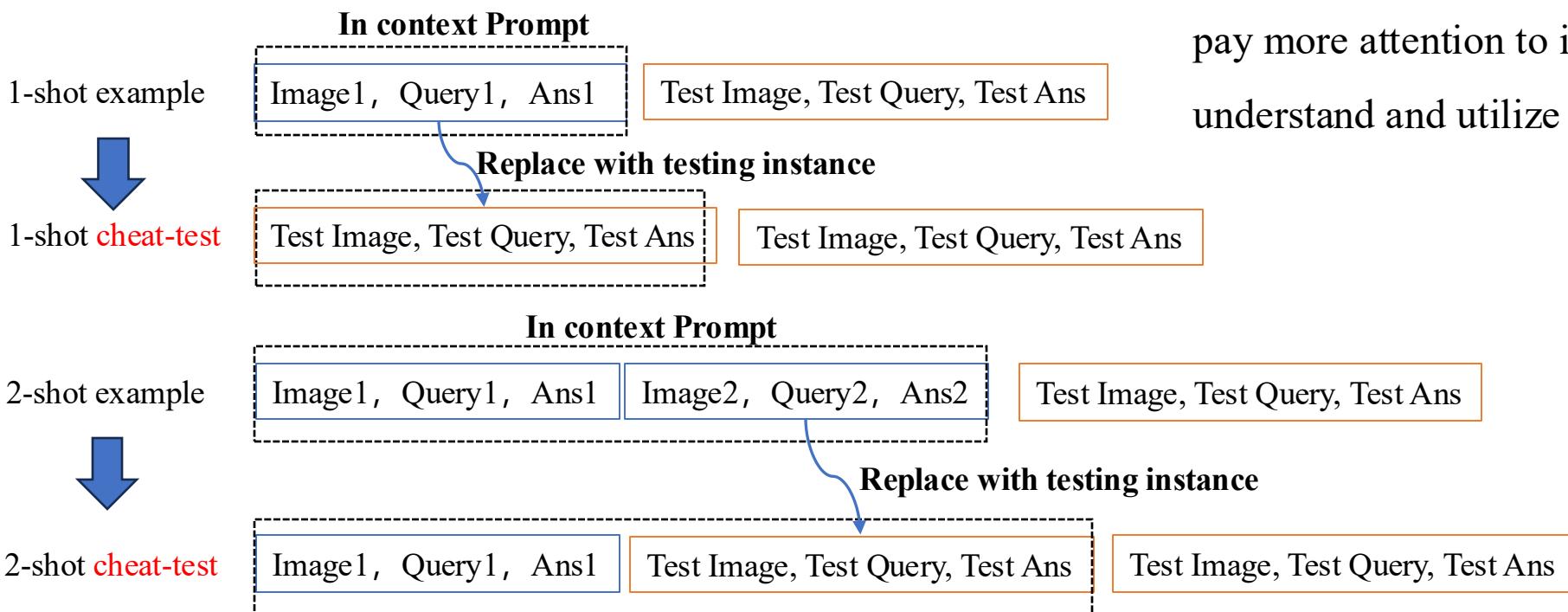
Experiment: Pre-training Performance

#Shot	0	1	2	4	0	1	2	4	0	1	2	4	0	1	2	4
Dataset																
ScienceQA ^{IMG}																
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																
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
MathVerse																
Avg.																

- 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
<i>1-shot Cheat:</i> 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
<i>2-shot Cheat:</i> 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.

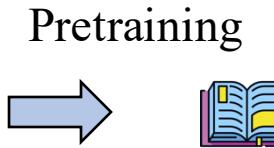
Future Works

- **Textbook-level multimodal corpora**



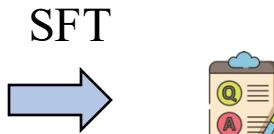
Create pretraining corpora that include basic knowledge of various disciplines

- **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

- **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