CSCE 689 - Special Topics in NLP for Science

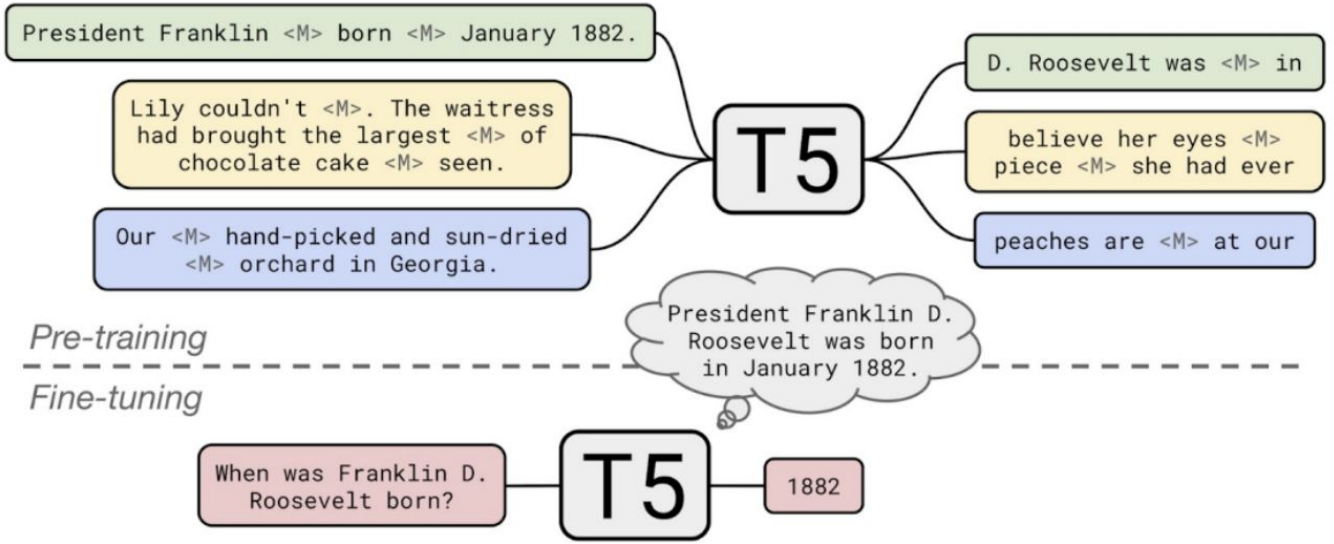
Scientific VLMs: Geometry

Shuo Xing

February 13, 2025

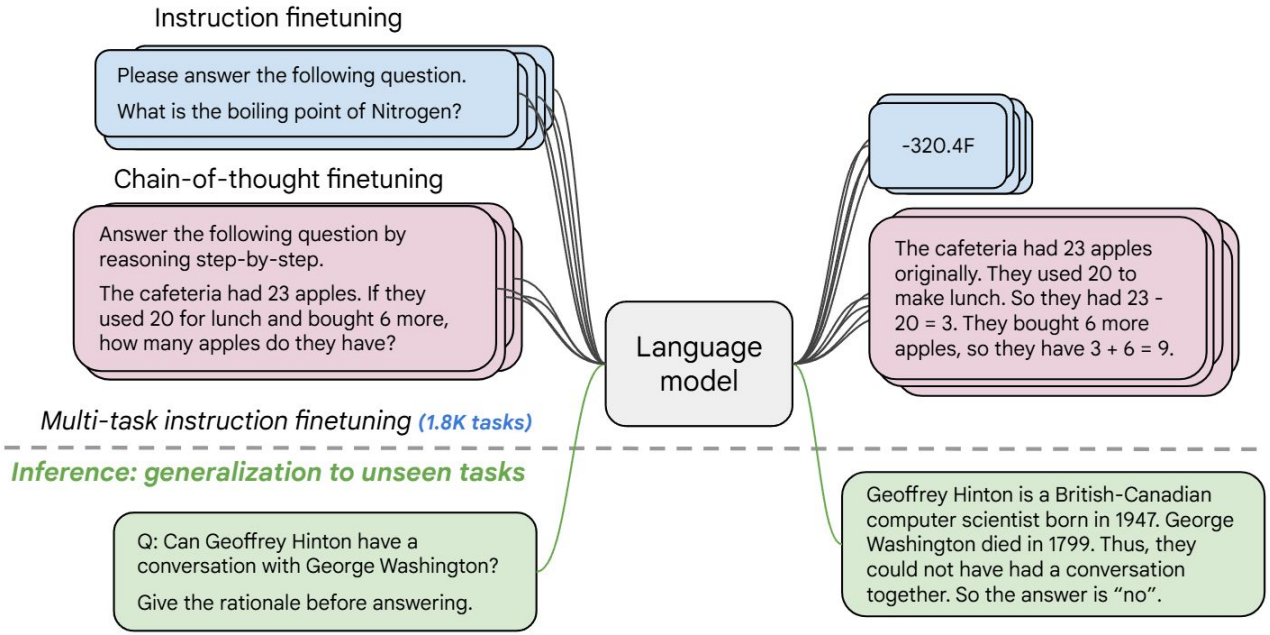
Preliminaries: T5

- T5: Text-to-Text Transfer Transformer
- Pre-training: Mask out spans of texts; generate the original spans
- Fine-tuning: Convert every task into a sequence-to-sequence generation problem



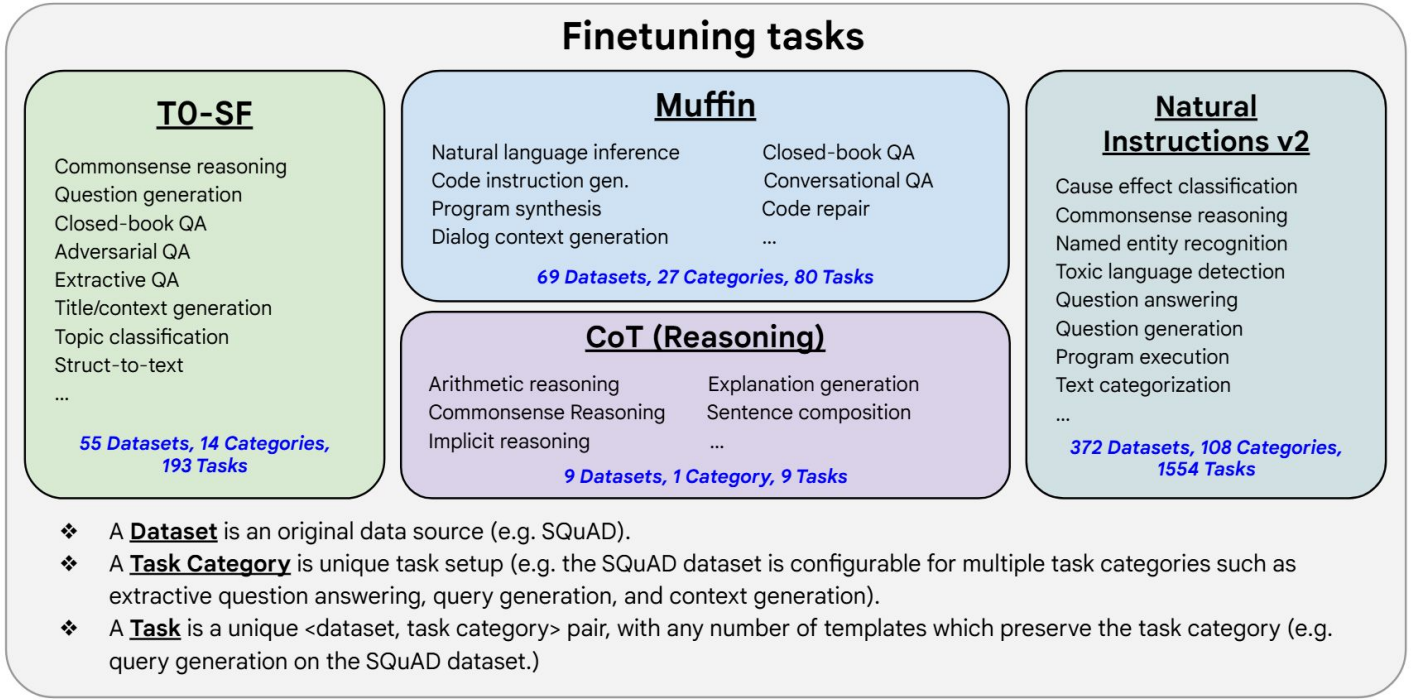
Preliminaries: FLAN-T5

Finetuning T5 on a collection of datasets phrased as ***instructions*** to improve model performance and generalization to unseen tasks



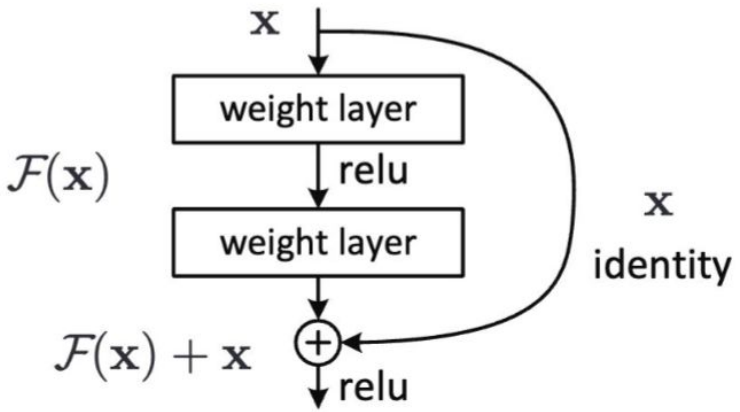
Preliminaries: FLAN-T5

Finetuning data comprises ***473 datasets, 146 task categories, and 1,836 total tasks***



Preliminaries: Residual Block

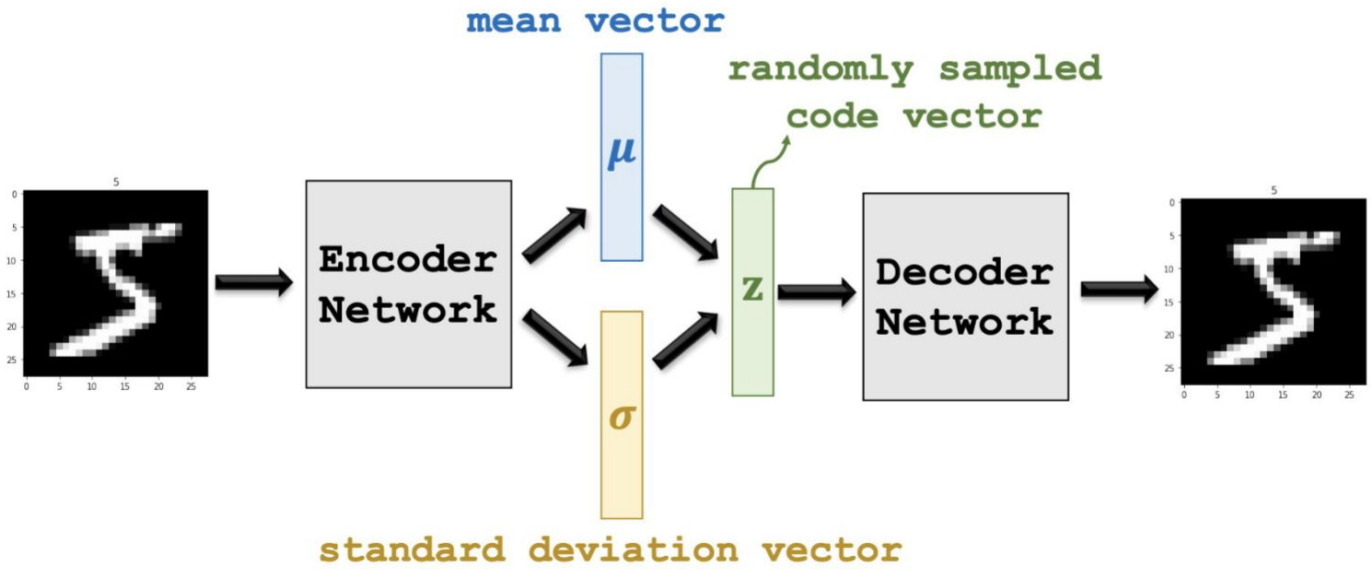
- Let weight layers fit residual function F(x)
- Let F(x) + x be the subnet output
- If identity is near-optimal
  - push weights to small
  - encourage small changes
- Initialize with small or zero weights



Deep Residual Learning for Image Recognition. CVPR 2016

Preliminaries: VQ-VAE

Variational Autoencoder with ***discrete latent space***

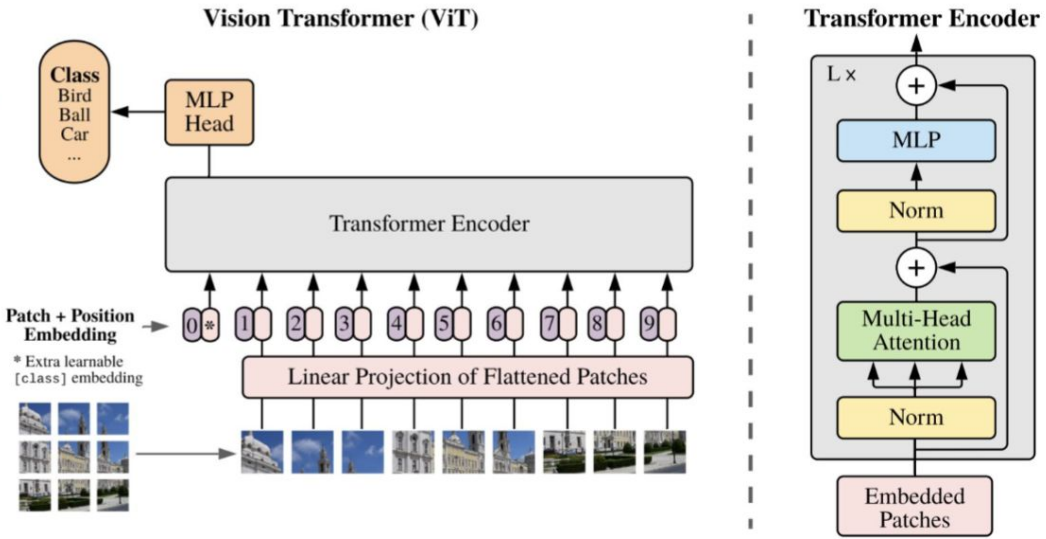


Neural Discrete Representation Learning. NeurIPS 2017

Preliminaries: ViT

Vision Transformer (ViT)

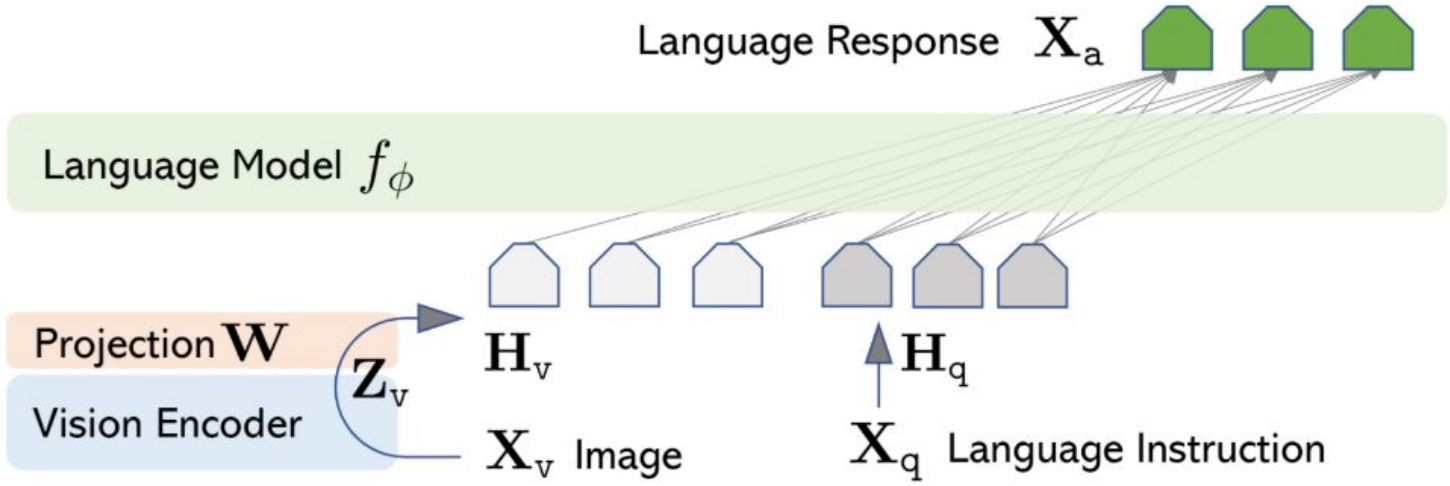
- Patchify images
- 1D processing of patches (same to NLP transformers)
- Apply self-attention layers to process the patch embedding



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021

Preliminaries: VLMs

Incorporating ***vision encoders*** to process image patches into tokens and aligning them with the ***text token space*** of LLMs



Visual Instruction Tuning. NeurIPS 2023 (Oral)

Agenda

- UniMath: A Foundational and Multimodal Mathematical Reasoner
- G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model
- Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models

Agenda

- UniMath: A Foundational and Multimodal Mathematical Reasoner
- G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model
- Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models

Mathematical Modalities

Math word problems (MWP):

textual information and execution of ***symbolic reasoning***

Text2Text

PROBLEM:
Text: Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Jack have now?
Equation: 8 - 3 = 5

QUESTION SENSITIVITY VARIATION:
Text: Jack had 8 pens and Mary had 5 pens. Jack gave 3 pens to Mary. How many pens does Mary have now?
Equation: 5 + 3 = 8

REASONING ABILITY VARIATION:
Text: Jack had 8 pens and Mary had 5 pens. Mary gave 3 pens to Jack. How many pens does Jack have now?
Equation: 8 + 3 = 11

STRUCTURAL INVARIANCE VARIATION:
Text: Jack gave 3 pens to Mary. If Jack had 8 pens and Mary had 5 pens initially, how many pens does Jack have now?
Equation: 8 - 3 = 5

Are NLP Models really able to Solve Simple Math Word Problems? NAACL 2021

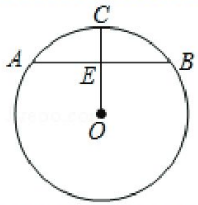
Mathematical Modalities

Geometry problem-solving:

visual context and ***reasoning on spatial relations***

Image2Text: require image encoder

As shown in the figure, in ⊙O, AB is the chord, OC⊥AB, if the radius of ⊙O is 5 (N0) and CE=2 (N1), then the length of AB is ()



A. 2 B. 4 C. 6 D. 8

Answer: D. 8

Problem Type: Length Calculation

Knowledge Points: Vertical Diameter, Pythagorean Theorem

Problem Solving Explanations:

OE=OC-CE=5-2=3. According to the Pythagorean Theorem,

AE = $\sqrt{OA^2 - OE^2}$= $\sqrt{5^2 - 3^2}$= 4. Thus, AB=2AE=8.

Annotated Programs:

Minus | N0 | N1 | PythagoreanMinus | N0 | V0 | Double | V1

Step1: Minus(N0, N1) = 5 – 2 = 3 (V0)

Step2: PythagoreanMinus(N0, V0) = $\sqrt{5^2 - 3^2}$ = 4 (V1)

Step3: Double(V1) = 2×4 = 8 (V2)

GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. Findings of ACL 2021

Mathematical Modalities

Table based math problem-solving: processing **structured table content** to **extract relevant information** for problem-solving

square beads	$2.97 per kilogram
oval beads	$3.41 per kilogram
flower-shaped beads	$2.18 per kilogram
star-shaped beads	$1.95 per kilogram
heart-shaped beads	$1.52 per kilogram
spherical beads	$3.42 per kilogram
rectangular beads	$1.97 per kilogram

Question: If Tracy buys 5 kilograms of spherical beads, 4 kilograms of star-shaped beads, and 3 kilograms of flower-shaped beads, how much will she spend? (unit: $)
Answer: 31.44
Solution:
Find the cost of the spherical beads. Multiply: $3.42 × 5 = $17.10.
Find the cost of the star-shaped beads. Multiply: $1.95 × 4 = $7.80.
Find the cost of the flower-shaped beads. Multiply: $2.18 × 3 = $6.54.
Now find the total cost by adding: $17.10 + $7.80 + $6.54 = $31.44.
She will spend $31.44.

transform tables into texts

Shop	Tuna	Egg salad
City Cafe	6	5
Sandwich City	3	12
Express Sandwiches	7	17
Sam's Sandwich Shop	1	6
Kelly's Subs	3	4

Question: As part of a project for health class, Cara surveyed local delis about the kinds of sandwiches sold. Which shop sold fewer sandwiches, Sandwich City or Express Sandwiches?
Options: (A) Sandwich City (B) Express Sandwiches
Answer: (A) Sandwich City
Solution:
Add the numbers in the Sandwich City row. Then, add the numbers in the Express Sandwiches row.
Sandwich City: 3 + 12 = 15. Express Sandwiches: 7 + 17 = 24.
15 is less than 24. Sandwich City sold fewer sandwiches.

Mathematical Modalities

The semi-structured format is created by converting the raw table text into a ***flattened token sequence***

Image format

Field day schedule		
Event	**Begin**	**End**
water balloon toss	11:30 A.M.	11:50 A.M.
obstacle course	12:05 P.M.	12:25 P.M.
parachute ball toss	12:30 P.M.	1:30 P.M.
jump rope race	1:40 P.M.	2:05 P.M.
balloon stomp	2:15 P.M.	2:35 P.M.
relay race	2:50 P.M.	3:40 P.M.
hula hoop contest	3:55 P.M.	4:30 P.M.
potato sack race	4:40 P.M.	5:15 P.M.

Semi-structured format

Table title: Field day schedule

Table text:

Event | Begin | End
water balloon toss | 11:30 A.M. | 11:50 A.M.
obstacle course | 12:05 P.M. | 12:25 P.M.
parachute ball toss | 12:30 P.M. | 1:30 P.M.
jump rope race | 1:40 P.M. | 2:05 P.M.
balloon stomp | 2:15 P.M. | 2:35 P.M.
relay race | 2:50 P.M. | 3:40 P.M.
hula hoop contest | 3:55 P.M. | 4:30 P.M.

Structured format

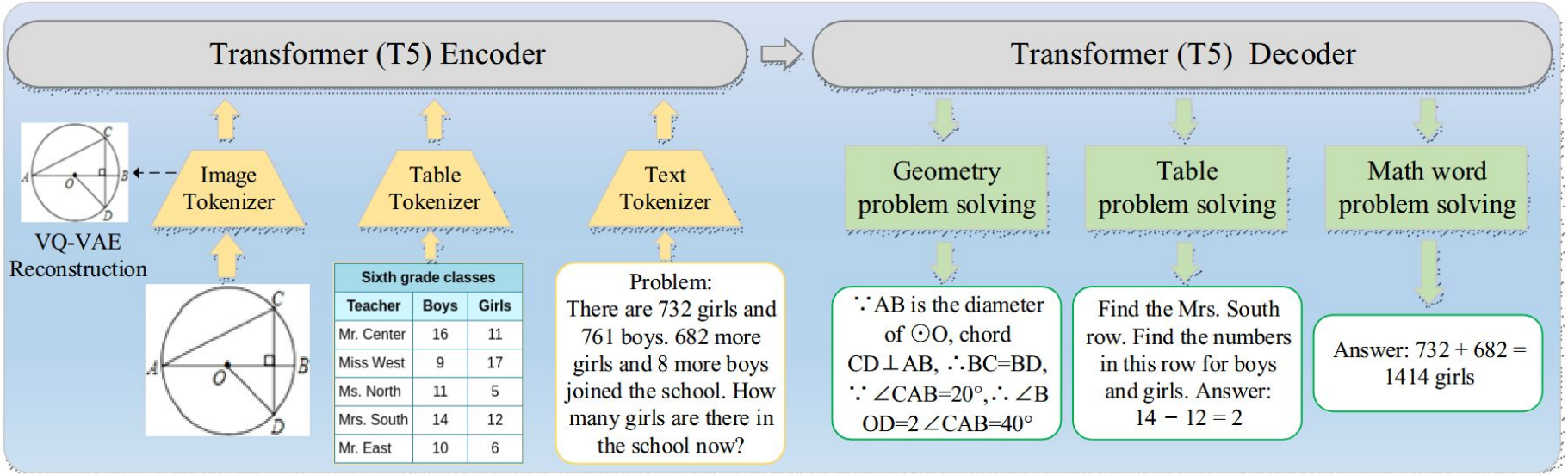
Table title: Field day schedule

	Event	Begin	End
0	water balloon toss	11:30 A.M.	11:50 A.M.
1	obstacle course	12:05 P.M.	12:25 P.M.
2	parachute ball toss	12:30 P.M.	1:30 P.M.
3	jump rope race	1:40 P.M.	2:05 P.M.
4	balloon stomp	2:15 P.M.	2:35 P.M.
5	relay race	2:50 P.M.	3:40 P.M.
6	hula hoop contest	3:55 P.M.	4:30 P.M.
7	potato sack race	4:40 P.M.	5:15 P.M.

UniMath

A unified system designed for multimodal mathematical reasoning tasks

Jointly training the model on three datasets - SVAMP, GeoQA, and TableMWP

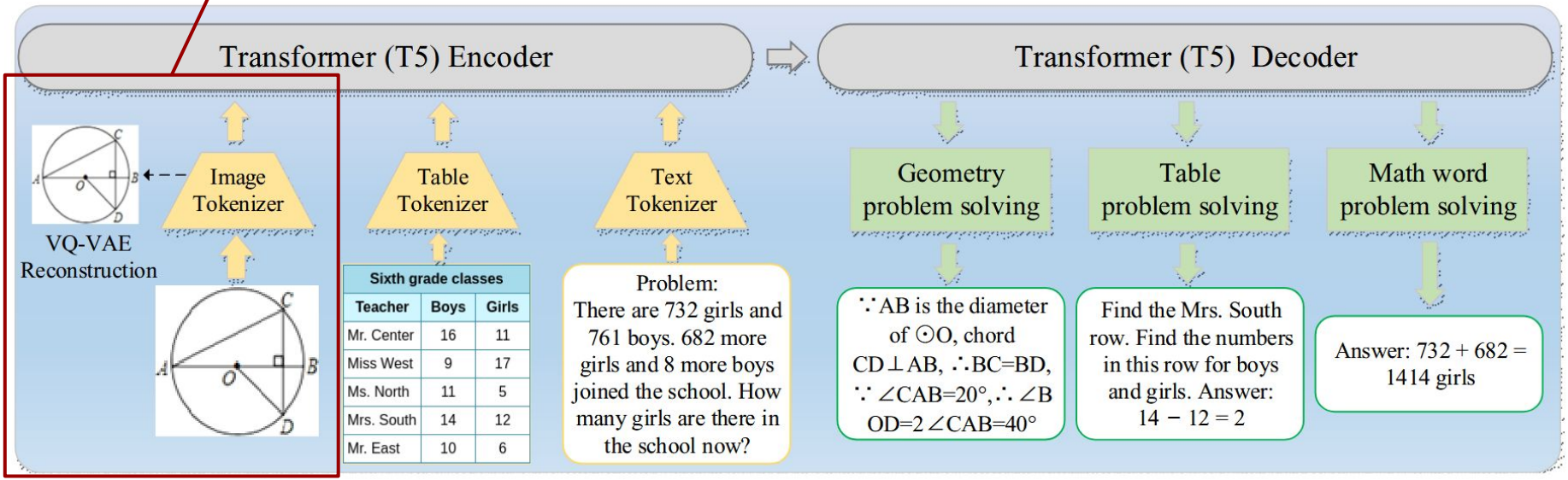


UniMath: A Foundational and Multimodal Mathematical Reasoner. EMNLP 2023

UniMath

Textual tokes are discrete

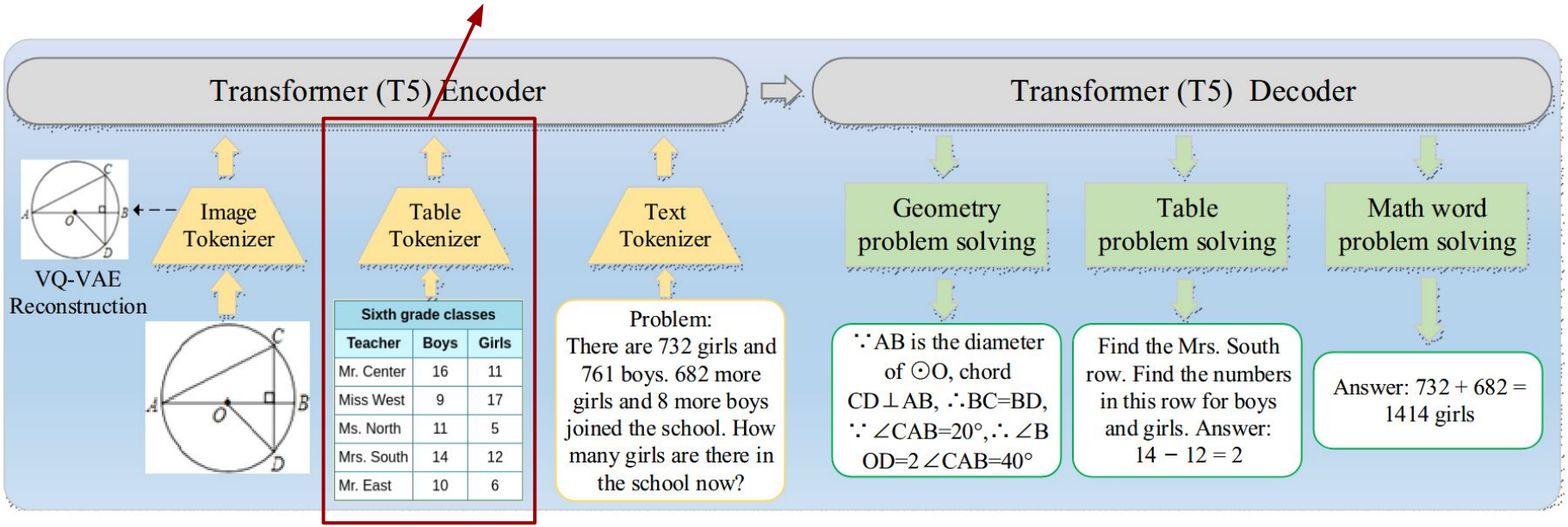
2-layer ResBlocks VQ-VAE as image encoder to transform image patches to ***new tokens*** and concatenate them with the ***textual tokens*** as the input.



UniMath: A Foundational and Multimodal Mathematical Reasoner. EMNLP 2023

UniMath Generated by Chain-of-Thought

The ***explanation*** and ***answer*** of the TableMWP dataset are separated into two targets during training controlled by different prefixes.



UniMath: A Foundational and Multimodal Mathematical Reasoner. EMNLP 2023

UniMath

- Effective unified mathematical reasoner with very competitive accuracy against state-of-the-art baselines

	Held-in Tasks			Held-out Tasks	
	SVAMP	GeoQA	TableMWP	MathQA	UniGeo-Proving
Best Fine-tuned Baseline	**47.3**^{a}	46.8^{b*}	58.5^{c}	78.6^{a}	80.6^{b*}
Train Individually on T5-base	29.8	43.7	62.7	82.3	82.7
Train Individually on Flan-T5-base	30.5	45.1	64.5	82.0	**83.0**
UniMath-T5-base	37.3	49.6	65.4	**83.3**	82.9
UniMath-Flan-T5-base	41.8	**50.0**	**66.5**	82.7	**83.0**

UniMath: A Foundational and Multimodal Mathematical Reasoner. EMNLP 2023

UniMath

- Able to generalize and help improve the fine-tuning on held-out tasks

	Held-in Tasks			Held-out Tasks	
	SVAMP	GeoQA	TableMWP	MathQA	UniGeo-Proving
Best Fine-tuned Baseline	**47.3**^{a}	46.8^{b*}	58.5^{c}	78.6^{a}	80.6^{b*}
Train Individually on T5-base	29.8	43.7	62.7	82.3	82.7
Train Individually on Flan-T5-base	30.5	45.1	64.5	82.0	**83.0**
UniMath-T5-base	37.3	49.6	65.4	**83.3**	82.9
UniMath-Flan-T5-base	41.8	**50.0**	**66.5**	82.7	**83.0**

UniMath: A Foundational and Multimodal Mathematical Reasoner. EMNLP 2023

Takeaways

- Process diverse math modalities separately
- Joint training yields better performance

Agenda

- UniMath: A Foundational and Multimodal Mathematical Reasoner
- G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model
- Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models

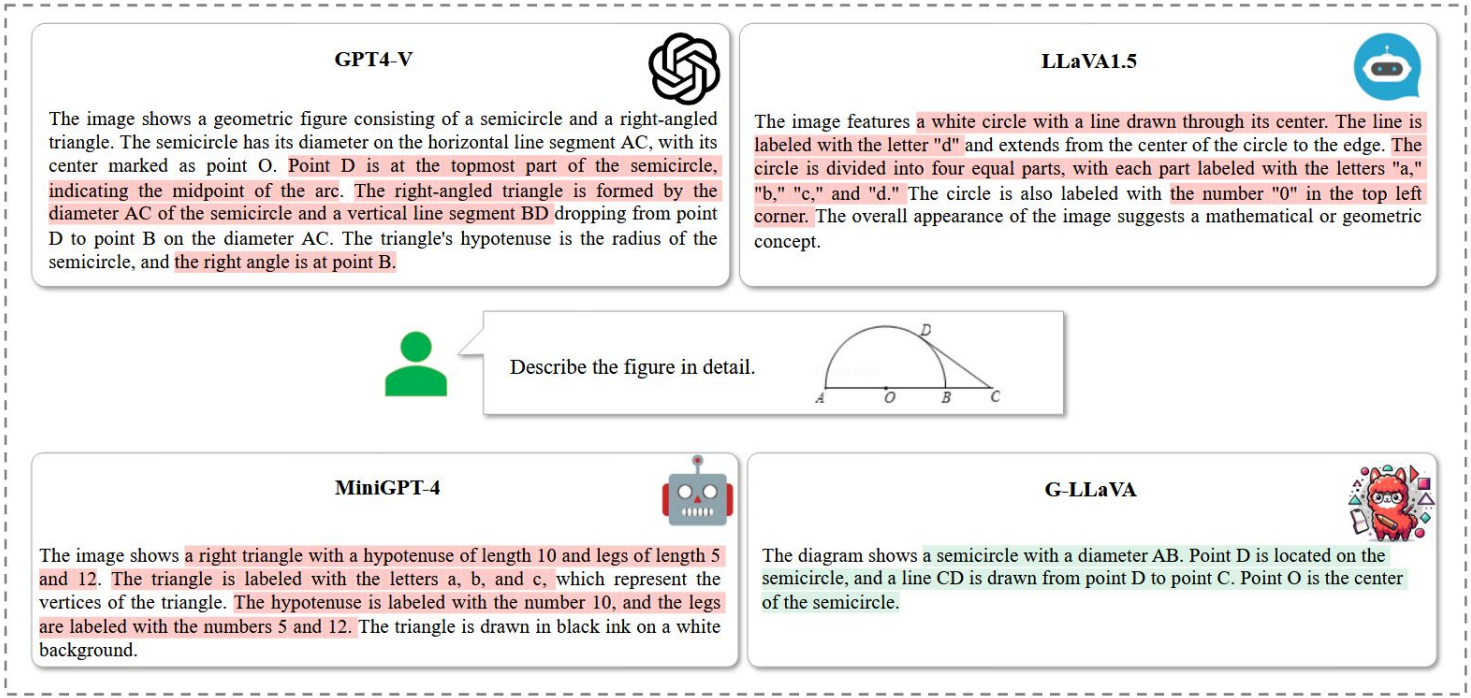
G-LLaVA

Vision encoder and LLMs are pretrained separately

SOTA VLMs suffer severe ***hallucination*** on ***geometric figures***, which greatly hinders their abilities on solving geometric problems.

Poor spatial reasoning

"View the image as a bag of items"



G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

G-LLaVA

Two-stage finetuning:

- Cross-modal alignment

  To mitigate the hallucinations of VLMs on understanding geometric figures

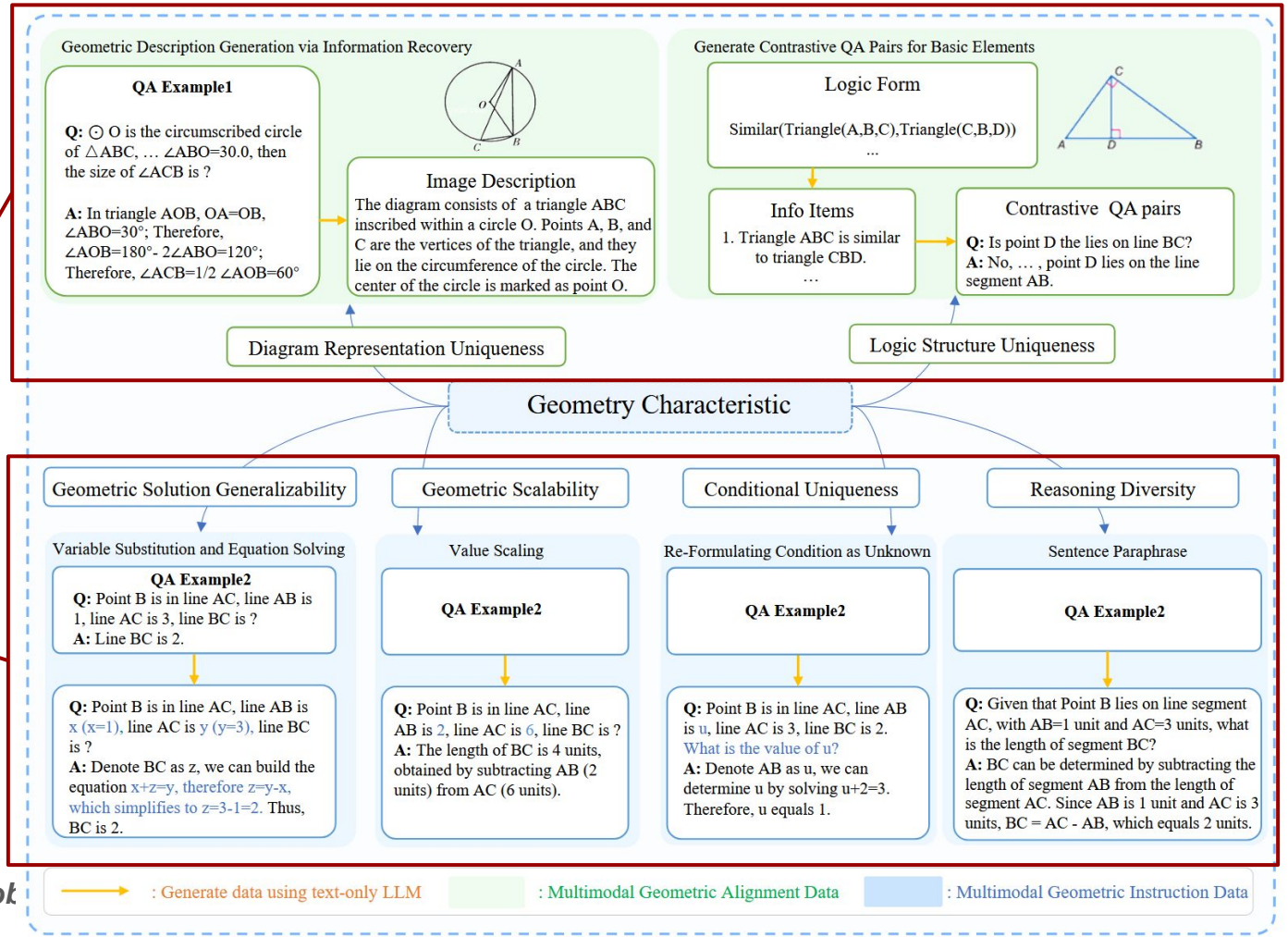
- Instruction-following tuning

  Enhance the capabilities of VLMs on addressing geometric math problems

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

G-LLaVA

A multi-modal geometry dataset based upon existing dataset (an ***alignment dataset*** + an ***geometric instruction data*** dataset).



G-LLaVA: Solving Geometric Prob
Model. arXiv 2023

G-LLaVA

Alignment dataset:

- Geometric Image Caption Generation: use ***text-only ChatGPT 3.5*** to create image captions based on these ***human-labeled QA pairs***, which can be considered as a type of inverse ***information recovery***.

Geometric Description Generation via Information Recovery

QA Pair:

Question: As shown in the figure, circle O is the circumscribed circle of triangle ABC, and it is known that angle ABO = 30.0, then the size of angle ACB is ()

Answer: In triangle AOB, OA=OB, angle ABO=30°; Therefore, angle AOB=180°- 2 angle ABO =120°; Therefore, angle ACB=1/2angle AOB=60°

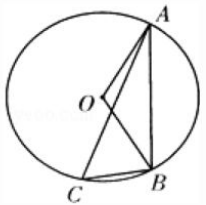


Diagram Description:

The diagram consists of a triangle ABC inscribed within a circle, where the circle is denoted as circle O. Points A, B, and C are the vertices of the triangle, and they all lie on the circumference of the circle. The center of the circle is marked as point O.

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

G-LLaVA

Alignment dataset:

- Contrastive QA Pairs:

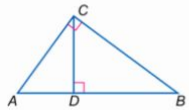
1. Use text-only ChatGPT to convert ***logical forms into clear descriptions***.

2. Produce ***contrastive QA pairs***

Contrastive QA Pairs for Basic Elements

Logic Form:

```
Similar(Triangle(A,B,C),Triangle(C,B,D))
Triangle(A,B,C)
Triangle(A,C,B)
Equals(LengthOf(Line(C, B)), 5)
Equals(LengthOf(Line(C, D)), 3)
Equals(LengthOf(Line(B, D)), 4)
PointLiesOnLine(D, Line(A, B))
Perpendicular(Line(A, C), Line(C, B))
Perpendicular(Line(C, D), Line(A, D))
```



Phase-1 Generation

Info Items:

The geometric figure have the following information:
Info 1: The figure represents a triangle with vertices A, B, and C.
Info 2: Triangle ABC is similar to triangle CBD.
Info 3: Point D lies on the line segment AB.
Info 4: Line segment AC is perpendicular to line segment CB.
...

Summary:

The figure depicts a triangle ABC and triangle CBD, which is similar to triangle ABC. Additionally, triangle ACB is also present in the figure. The lengths of the segments CB, CD, and BD are 5, 3, and 4 units respectively. Point D lies on the line segment AB, and perpendicularity is indicated between the line segments AC and CB, as well as between the line segments CD and AD.

Phase-2 Generation

QA pairs:

1. Question: Does the diagram include a triangle formed by vertices A, B, and D?
Answer: **No**, the diagram does not include a triangle formed by vertices A, B, and D. The triangles present are ABC and CBD, and the point D lies on the line segment AB.
2. Question: Does the diagram contain a triangle with vertices A, C, and B?
Answer: **Yes**, the diagram contains triangle ACB, which is the same as triangle ABC.
3. Question: Is point D the lies on line BC?
Answer: **No**, point D is not on lines BC. The point D lies on the line segment AB.
4. Question: Is line segment AC perpendicular to line segment CB?
Answer: **Yes**, line segment AC is perpendicular to line segment CB.

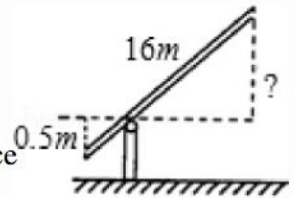
G-LLaVA: Solving Geometric Problem wit
Model. arXiv 2023

G-LLaVA

Geometric Instruction Data: construct an instruction tuning dataset based on existing datasets with the help of powerful LLMs.

Original Example

Question:
As shown in the figure, the length of the short arm of the railing at the railway crossing is 1.0 and the length of the long arm is 16.0. When the end of the short arm drops by 0.5, the end of the long arm rises ()



Answer:
By the property of similar triangles, the lengths of corresponding sides are proportional. Since the long arm is 16 times longer than the short arm, the height it rises is 16 times the height the short arm drops, which is 0.5 multiplied by 16, resulting in a rise of 8 meters.

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

G-LLaVA

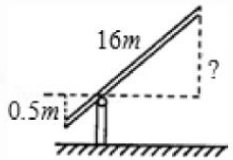
Geometric Instruction Data:

- Equation Solving (ES):

Replace the ***specific values*** in the original QA pairs with ***unknown variables*** and prompt the LLM to construct the solution by ***solving equation***

Variable Substitution and Equation Solving

Question:
Given a railway crossing barrier with two arms, where the shorter arm has a length of 'a' units (a = 1.0) and the longer arm has a length of 'b' units (b = 16.0), determine the elevation 'y' units of the longer arm's end when the shorter arm descends by 'd' units (d = 0.5). Express your answer as a function of 'a', 'b', and 'd'.



Answer:
To find the elevation 'y' of the long arm's end, we apply the property of similar triangles and establish the equation $\frac{d}{y} = \frac{a}{b}$. Substituting the given values, we get $\frac{0.5}{y} = \frac{1.0}{16.0}$. Solving for 'y' leads to $y = \frac{b \times d}{a}$, which simplifies to $y = \frac{16.0 \times 0.5}{1.0}$. Therefore, 'y' equals 8.0 units, indicating that the long arm rises by 8.0 units when the short arm descends by 0.5 units.

Question:
In a railway crossing lever system, the lengths of the shorter and longer arms are 's' units (s = 1.0) and 'l' units (l = 16.0) respectively. If the shorter arm lowers by 'deltaS' units (deltaS = 0.5), what is the increase 'deltaL' in the length of the longer arm?

Answer:
The increase 'deltaL' in the longer arm can be found using the proportionality relationship of similar triangles. The equation can be expressed as: $\frac{deltaS}{deltaL} = \frac{s}{l}$. Inserting the given values, the equation becomes $\frac{0.5}{deltaL} = \frac{1.0}{16.0}$. Solving for 'deltaL' yields $deltaL = \frac{l \times deltaS}{s}$, which simplifies to $deltaL = \frac{16.0 \times 0.5}{1.0}$. Thus, deltaL is 8.0 units.

Question:
he barrier is 1.0 unit long and the long arm is 16.0 units long. If the short will the long arm rise?

Generalize its understanding of the problem

Answer:
The triangles formed by the shorter and longer arms as they descend are similar. The rise of the long arm 'h' can be calculated using the property of similar triangles, the equation is 0.5/h = 1/16. Solving for 'h' gives h = 16 * 0.5 / 1, which equals 8.0 units.

G-LLaVA: Solving Geometric Problem wi
Model. arXiv 2023

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

G-LLaVA

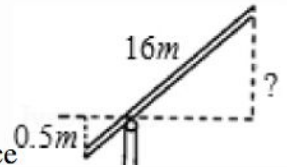
Geometric Instruction Data:

Flexible in handling different numerical inputs

- Value Scaling (VS): augment the data by scaling the length values in the QA pairs

Original Example

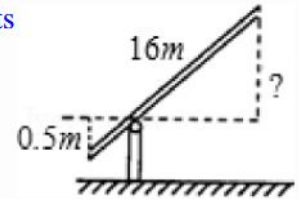
Question:
As shown in the figure, the length of the short arm of the railing at the railway crossing is 1.0 and the length of the long arm is 16.0. When the end of the short arm drops by 0.5, the end of the long arm rises ()



Answer:
By the property of similar triangles, the lengths of corresponding sides are proportional. Since

Value Scaling

Question:
At a railroad crossing, the short arm of the barrier is 2.0 unit long and the long arm is 32 units long. If the short arm drops 1 units, by how many units will the long arm rise?



Answer:
Denote the rise of the long arm as 'h'. In similar triangles, the ratios of their corresponding sides are in proportion. h can be calculated using the equation 1/h = 2/32. Solving for 'h' gives h = 32 * 1 / 2, which equals 16 units. Therefore, the long arm rises by 16 units.

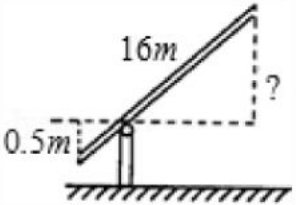
G-LLaVA

Geometric Instruction Data: Flexible in handling different numerical inputs

- Re-Formulating Condition as Unknown (RCU): reformulate questions to ask for the values originally present in the condition, and retain the generated data with correct answer only

Re-Formulating Condition as Unknown (Weng et al., 2023; Yu et al., 2023)

Question:
At a railroad crossing, the short arm of the barrier is u unit long and the long arm is 16.0 units long. When the end of the short arm drops by 0.5, the end of the long arm rises 8 units. What is the value of unknown variable u?



Answer:
Denote the short arm of the barrier as variable u. By the property of similar triangles, we can determine u by solving the equation 0.5/8 = u/16. Therefore, u equals 1.

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

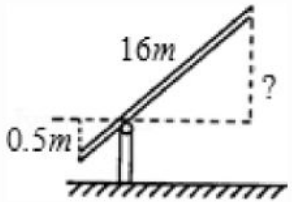
G-LLaVA

Geometric Instruction Data: Handle similar questions with different phrasings and provide accurate responses

- Sentence Paraphrase (SP): paraphrasing for both the question and answer pairs

Sentence Paraphrase

Question:
In the illustration, the railing at the railway crossing has a short arm measuring 1.0 unit in length and a long arm measuring 16.0 units. When the short arm drops by 0.5 units, what is the corresponding rise in the long arm?



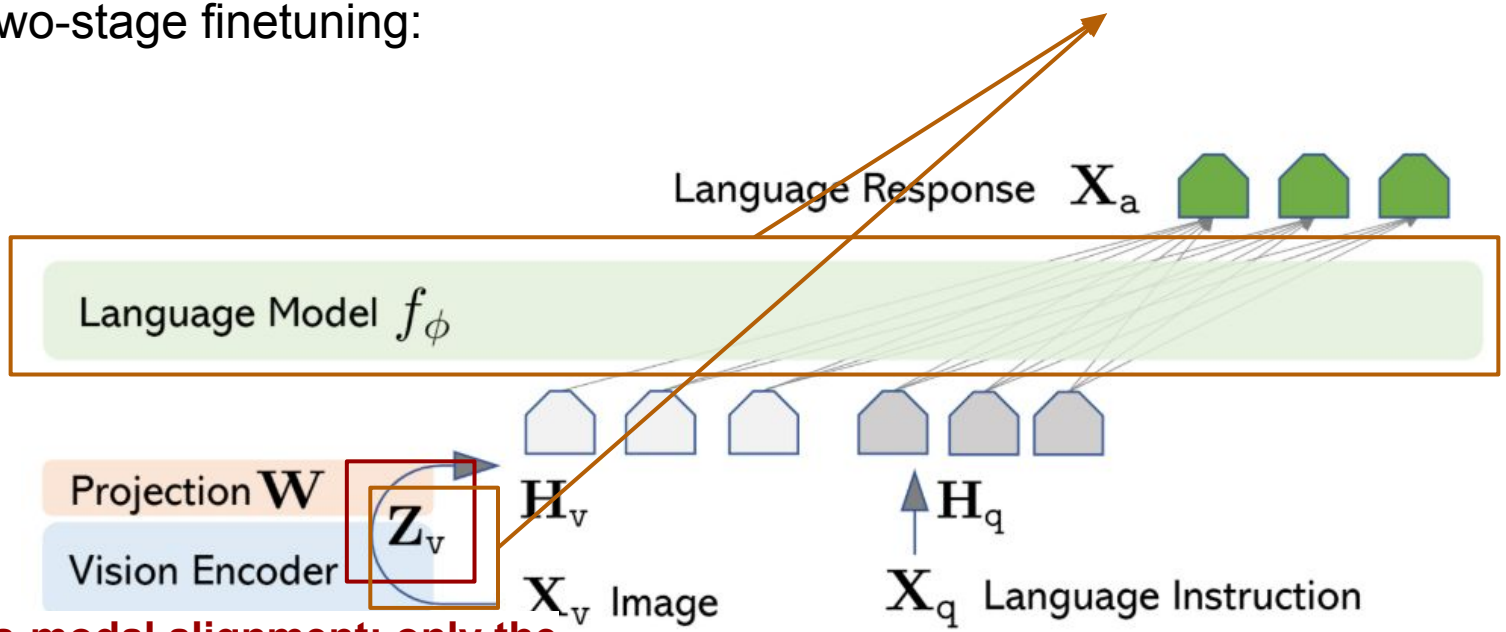
Answer:
The triangles are similar, and their corresponding sides are proportional. The long arm is 16 times longer than the short arm, resulting in an 8-meter rise when the short arm drops by 0.5 meters.

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

G-LLaVA

Two-stage finetuning:

Instruction-following tuning: both the projection linear layer and the language model are trainable



Cross-modal alignment: only the projection linear layer is trainable

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

G-LLaVA

Data Gen:

GeoQA+ and Geometry3K

Testing:

MathVista

Model	Input	Accuracy (%)
Heuristics Baseline		
Random Chance	-	21.6
Frequent Guess	-	34.1
Human	Q, I	48.4
Close Source Model		
Text-Only LLMs		
2-shot CoT Claude-2	Q	29.8
2-shot CoT ChatGPT	Q	36.5
2-shot CoT GPT-4	Q	44.7
2-shot PoT ChatGPT	Q	30.8
2-shot PoT GPT-4	Q	33.2
Visual-Augmented LLMs		
2-shot CoT Claude-2	Q, I_c, I_t	31.7
2-shot CoT ChatGPT	Q, I_c, I_t	29.3
2-shot CoT GPT-4	Q, I_c, I_t	31.7
2-shot PoT ChatGPT	Q, I_c, I_t	26.4
2-shot PoT GPT-4	Q, I_c, I_t	39.4
Multimodal LLMs		
Multimodal Bard	Q, I	47.1
Gemini Nano 1	Q, I	21.6
Gemini Nano 2	Q, I	23.6
Gemini Pro	Q, I	40.4
Gemini Ultra	Q, I	56.3
GPT4-V	Q, I	50.5
Open Source Model		
IDEFICS (9B-Instruct)	Q, I	21.1
mPLUG-Owl (LLaMA-7B)	Q, I	23.6
miniGPT4 (LLaMA-2-7B)	Q, I	26.0
LLaMA-Adapter-V2 (7B)	Q, I	25.5
LLaVAR	Q, I	25.0
InstructBLIP (Vicuna-7B)	Q, I	20.7
LLaVA (LLaMA-2-13B)	Q, I	29.3
G-LLaVA-7B	Q, I	53.4
G-LLaVA-13B	Q, I	**56.7**

G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model. arXiv 2023

Takeaways

- Two-stage finetuning pipeline: alignment + instruction-following
- Two-phase data augmentation pipeline

- Potential future work:

  Using preference optimization or contrastive learning during the first finetuning stage for aligning the VLMs

Agenda

- UniMath: A Foundational and Multimodal Mathematical Reasoner
- G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model
- Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models

Math-LLaVA

Existing open-source image instruction fine-tuning datasets, containing limited question-answer pairs per image, do not fully exploit ***visual information*** to enhance the multimodal mathematical reasoning capabilities of VLMs

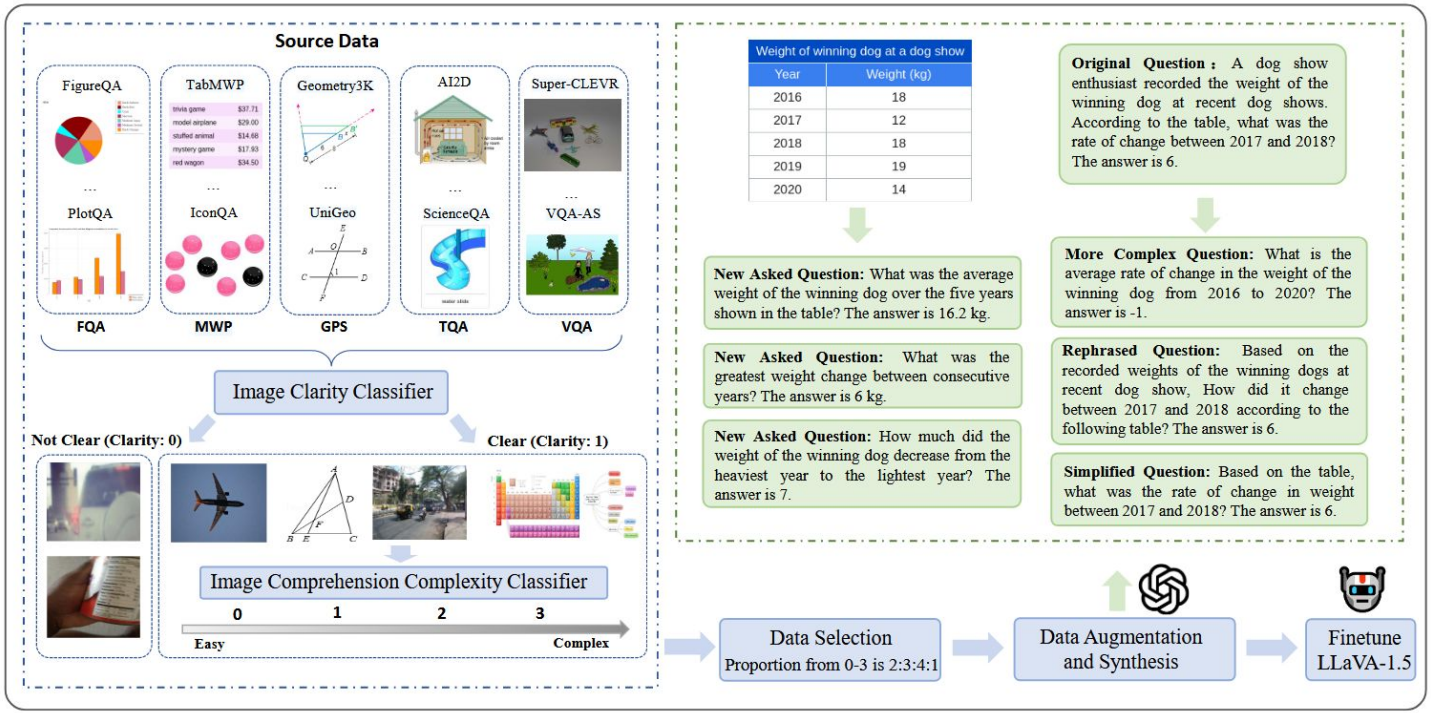
Contribution:

- Collecting 40K high-quality images from 24 existing datasets and synthesizing 320K new pairs, creating the MathV360K dataset
- Fine-tuning LLaVA-1.5 with MathV360K, we developed Math-LLaVA

Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. EMNLP 2024

Math-LLaVA

- Collecting Data
- Data Augmentation
- Finetuning



Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. EMNLP 2024

Math-LLaVA

Collecting Data

24 visual question answering and multimodal mathematical reasoning datasets

Dataset	Task	Visual Context	Training Images	Clear Images	Image Complexity			
					0	1	2	3
DocVQA (2022)	FQA	Document Image	8535	8227	2086	6007	125	9
FigureQA (2017)	FQA	Charts and Plots	18173	18173	687	16792	694	0
DVQA (2018)	FQA	Bar Chart	19092	19092	21	18021	1045	5
PlotQA (2020)	FQA	Bar, Line, Scatter	18782	18782	13	18759	10	0
ChartQA (2022)	FQA	Charts and Plots	3699	3699	0	3649	50	0
MapQA (2022)	FQA	Map Chart	10020	10016	1	10015	0	0
IconQA (2021b)	MWP	Abstract Scene	20000	19068	10991	8055	22	0
CLEVR-Math (2022)	MWP	Synthetic Scene	17552	17551	1	17550	0	0
TabMWP (2022b)	MWP	Table	20000	20000	14919	5081	0	0
GEOS (2015)	GPS	Geometry Diagram	66	64	2	57	5	0
Geometry3K (2021a)	GPS	Geometry Diagram	2101	2101	21	1508	568	4
GeoQA+ (2022)	GPS	Geometry Diagram	6027	5956	103	4399	1454	0
UniGeo (2022)	GPS	Geometry Diagram	3499	3432	72	2514	846	0
TQA (2017)	TQA	Scientific Figure	1499	1497	20	949	498	30
AI2D (2016)	TQA	Scientific Figure	3247	3235	32	2321	823	59
ScienceQA (2022a)	TQA	Scientific Figure	6218	6061	1533	4251	273	4
A-OKVQA (2022)	VQA	Natural Image	16540	14526	10	11724	2743	49
VQA2.0 (2017)	VQA	Natural Image	16912	14521	45	12783	1672	21
PMC-VQA (2023a)	VQA	Medical Image	19682	9846	62	2989	3501	3294
VizWiz (2018)	VQA	Natural Image	20,000	16400	790	14800	770	40
Super-CLEVR (2023)	VQA	Synthetic Scene	2000	1950	1	1568	381	0
VQA-AS (2015)	VQA	Abstract Scene	14065	14065	7	13996	62	0
VQA-RAD (2018)	VQA	Medical Image	259	248	0	91	95	62
TextVQA (2019)	VQA	Natural Image	15815	11350	179	9497	1598	76

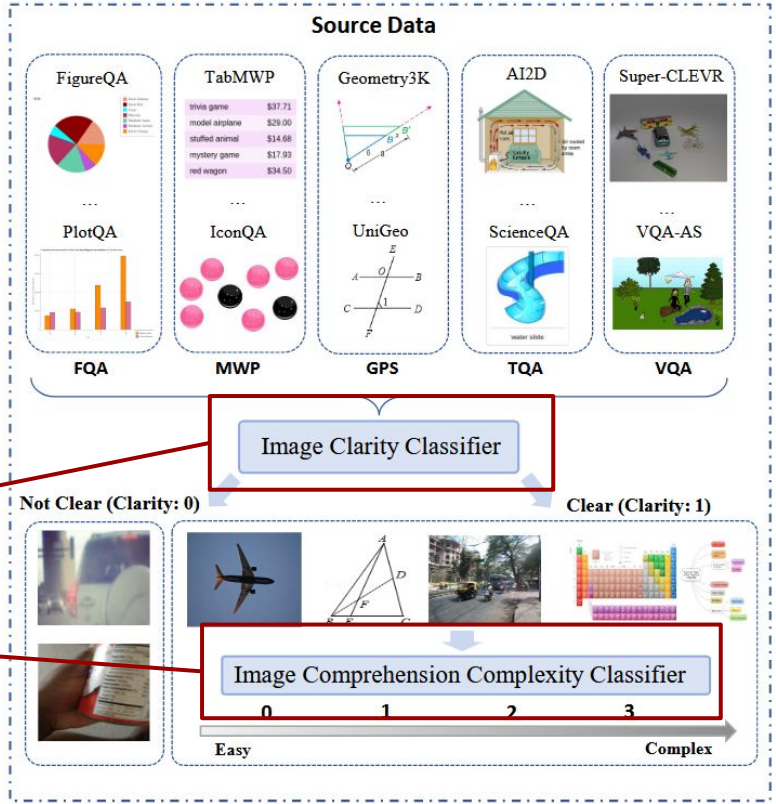
Math-LLaVA: Bootstrapping Mathematic Language Models. EMNLP 2024

Math-LLaVA

Collecting Data

- Image Filtering and Proportioning based on:
  - ***clarity of the images***
  - ***comprehension complexity***

Training two ViTs using data annotated by GPT4V



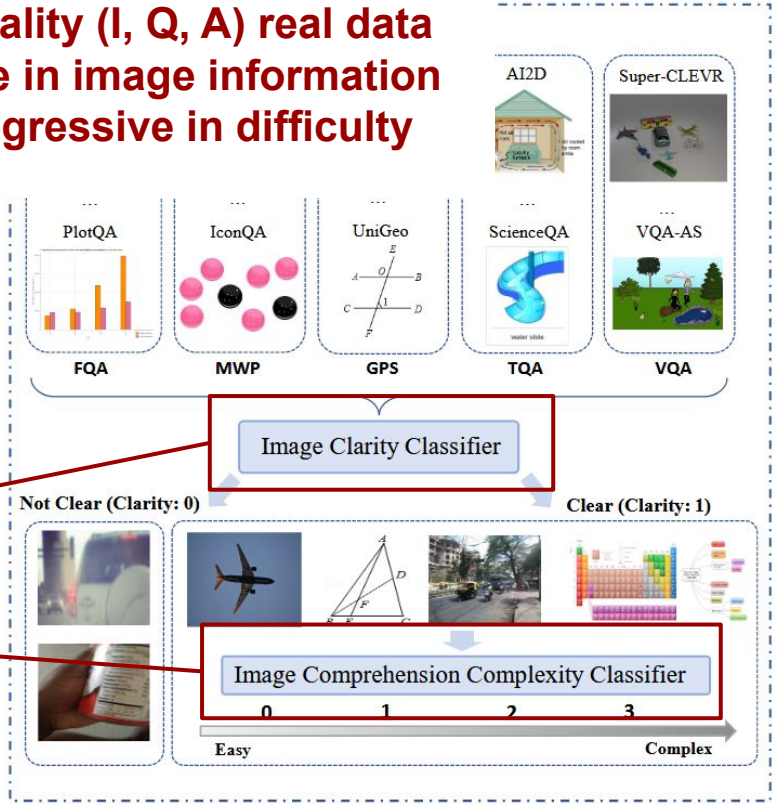
Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. EMNLP 2024

Math-LLaVA

Obtained 40K high-quality (I, Q, A) real data points that are diverse in image information and questions are progressive in difficulty

Collecting Data

- Image Filtering and Proportioning based on:
  - ***clarity of the images***
  - ***comprehension complexity***



Training two ViTs using data annotated by GPT4V

Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. EMNLP 2024

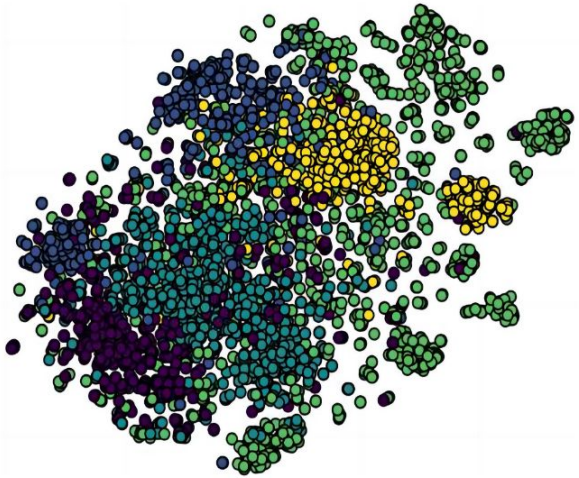
Math-LLaVA

Term Frequency–Inverse Document Frequency (TF-IDF): a numerical statistic often used in information retrieval and text mining to assess how important a word is to a document within a collection of documents

Data Augmentation

- Clustering:
  - using ***TF-IDF*** to extract features of ***text questions***
  - clustered using ***K-Means*** into FQA, GPS, MWP, TQA, VQA

To construct few-shot examples for generating new questions



Questions Clustering of IconQA

Cluster 0: How many scooters are there? ...

Cluster 1: Move the ruler to measure the length of the nail to the nearest inch. The nail is about (_) inches long. ...

Cluster 2: The first picture is a paw. Which picture is eighth? ...

Cluster 3: If you select a marble without looking, which color are you more likely to pick? ...

Cluster 4: Rick is waking up in the morning. The clock by his bed shows the time. What time is it? ...

Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. EMNLP 2024

Math-LLaVA

To fully exploit visual information of an image

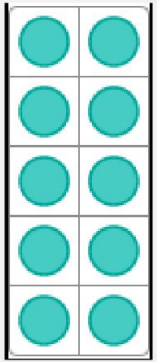
Data Augmentation

- Generate additional questions:

  Using ***few-shot prompting*** to generate ***5 new questions*** based on the original images and questions

Prompt-Ask New Questions:

Input Image



Q: How many dots are there?

[ROLE] You are an expert at understanding images and good at asking and answering questions based on the given images.

[TASK] You will be given some question examples. Please refer to the format of the examples to ask up to five high-quality questions on the given image. The original question of the image will also be given, please avoid asking the same question. Please provide the correct answer within ten words or answer with only an integer or float number.

[EXAMPLES]
From IconQA:
Question: There is 1 ball in the top row. How many balls are in the bottom row?
Question: What has been done to this letter? ...
From CLEVR-Math: ...
From TabMWP: ...
[ORIGINAL QUESTION] {Q}
[REQUIREMENT] Please follow and make full use of the image information. Please avoid asking questions for which you are not confident to give the definite correct answer. Please do not completely copy the content of the example questions. Ensure that provide final correct answer for each question.
[OUTPUT FORMAT] Your output MUST be "The question is [YOUR QUESTION]. The answer is [YOUR CORRECT ANSWER]."

Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. EMNLP 2024

Math-LLaVA

Obtained 8 x 40K = 320K synthetic data

Data Augmentation

- Augmentation of original question:
  - ***more complex questions*** based on the original image and corresponding inquiries
  - ask the ***same question in different ways*** without changing the answer
  - ***simplified*** the original questions without affecting their semantic understanding

Prompt-Complexity:
You will be given the question for the given image. Please ask a more complex question that requires more steps to answer than the given question.
Question: {Q}

Prompt-Logical Consistency:
You are an AI assistant to help me rephrase questions. Please ask the same question in a different way but have to make sure the answer won't be changed.
Question: {Q}
Rephrase the above question:

Prompt-Underspecification:
You are an AI assistant to help me rephrase question of the given image. Please simplify the question into a concise question, but does not affect the understanding and answering question with the image.
Question: {Q}
Simplify the above question:

Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models. EMNLP 2024

Math-LLaVA

Finetune LLaVA-v1.5-13B on MathV360K (SFT)

Test on MathVista

ALG: algebraic reasoning
ARI: arithmetic reasoning
GEO: geometry reasoning
LOG:logical reasoning
NUM: numeric commonsense
SCI: scientific reasoning
STA: statistical reasoning

Model	MathVista												
	ALL	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA
Heuristics Baselines													
Random Chance	17.9	18.2	21.6	3.8	19.6	26.3	21.7	14.7	20.1	13.5	8.3	17.2	16.3
Frequent Guess (Lu et al., 2023)	26.3	22.7	34.1	20.4	31.0	24.6	33.1	18.7	31.4	24.3	19.4	32.0	20.9
Human	60.3	59.7	48.4	73.0	63.2	55.9	50.9	59.2	51.4	40.7	53.8	64.9	63.9
Close-Source Multimodal Large Langugae Models (MLLMs)													
Gemini 1.0 Nano 2 (Team et al., 2023)	30.6	28.6	23.6	30.6	41.8	31.8	27.1	29.8	26.8	10.8	20.8	40.2	33.5
Qwen-VL-Plus (Bai et al., 2023)	43.3	**54.6**	38.5	31.2	55.1	34.1	39.1	32.0	39.3	18.9	26.4	59.0	56.1
Gemini 1.0 Pro (Team et al., 2023)	45.2	47.6	40.4	39.2	61.4	**39.1**	45.2	38.8	41.0	10.8	**32.6**	54.9	**56.8**
Claude 3 Haiku (Anthropic, 2024)	46.4	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V (OpenAI)	**49.9**	43.1	**50.5**	**57.5**	**65.2**	38.0	**53.0**	**49.0**	**51.0**	**21.6**	20.1	**63.1**	55.8
Open-Source Multimodal Large Langugae Models (MLLMs)													
mPLUG-Owl-7B (Ye et al., 2023)	22.2	22.7	23.6	10.2	27.2	27.9	23.6	19.2	23.9	13.5	12.7	26.3	21.4
miniGPT4-7B (Zhu et al., 2023)	23.1	18.6	26.0	13.4	30.4	30.2	28.1	21.0	24.7	16.2	16.7	25.4	17.9
LLaVAR-13B (Zhang et al., 2023b)	25.2	21.9	25.0	16.7	34.8	30.7	24.2	22.1	23.0	13.5	15.3	42.6	21.9
InstructBLIP-7B (Dai et al., 2024)	25.3	23.1	20.7	18.3	32.3	35.2	21.8	27.1	20.7	18.9	20.4	33.0	23.1
LLaVA-13B (Liu et al., 2023)	26.1	26.8	29.3	16.1	32.3	26.3	27.3	20.1	28.8	24.3	18.3	37.3	25.1
SPHINX-V1-13B (Lin et al., 2023b)	27.5	23.4	23.1	21.5	39.9	34.1	25.6	28.1	23.4	16.2	17.4	40.2	23.6
LLaVA-1.5-13B (Liu et al., 2024)	27.6	-	-	-	-	-	-	-	-	-	-	-	-
LLaVA-1.5-13B† (Liu et al., 2024)	27.7	23.8	22.7	18.3	40.5	30.2	25.3	26.4	22.8	21.6	26.4	35.3	23.6
OmniLMM-12B (OpenBMB, 2024)	34.9	45.0	17.8	26.9	44.9	39.1	23.1	32.3	20.9	18.9	27.8	45.9	44.2
SPHINX-V2-13B (Lin et al., 2023b)	36.7	**54.6**	16.4	23.1	41.8	**43.0**	20.6	33.4	17.6	**24.3**	21.5	43.4	**51.5**
G-LLaVA-13B (Gao et al., 2023a)	-	-	56.7	-	-	-	-	-	-	-	-	-	-
Math-LLaVA-DS	38.2	33.5	47.2	41.4	36.7	34.6	38.4	34.3	45.6	18.9	33.3	45.9	35.2
Math-LLaVA	**46.6**	37.2	**57.7**	**56.5**	**51.3**	33.5	**53**	**40.2**	**56.5**	16.2	**33.3**	**49.2**	43.9

Math-LLaVA

Finetune LLaVA-v1.5-13B on MathV360K (SFT)

Test on MathVista

1. achieves 46.6% overall accuracy with a significant improvement compare with vanilla LLaVA
2. achieves 57.7% accuracy on GPS subset, outperforming G-LLaVA-13B
3. achieving comparable performance to GPT-4V

ALG: algebraic reasoning
ARI: arithmetic reasoning
GEO: geometry reasoning
LOG:logical reasoning
NUM: numeric commonsense
SCI: scientific reasoning
STA: statistical reasoning

Model	MathVista												
	ALL	FQA	GPS	MWP	TQA	VQA	ALG	ARI	GEO	LOG	NUM	SCI	STA
Frequ												2	16.3
												0	20.9
												9	63.9
Gemini 1												2	33.5
Qwen-VL-Plus (Bai et al., 2023)	43.3	**54.6**	38.3	31.2	55.1	34.1	39.1	32.0	39.3	18.9	26.4	59.0	56.1
Gemini 1.0 Pro (Team et al., 2023)	45.2	47.6	40.4	39.2	61.4	**39.1**	45.2	38.8	41.0	10.8	**32.6**	54.9	**56.8**
Claude 3 Haiku (Anthropic, 2024)	46.4	-	-	-	-	-	-	-	-	-	-	-	-
GPT-4V (OpenAI)	**49.9**	43.1	**50.5**	**57.5**	**65.2**	38.0	**53.0**	**49.0**	**51.0**	**21.6**	20.1	**63.1**	55.8
Open-Source Multimodal Large Langugae Models (MLLMs)													
mPLUG-Owl-7B (Ye et al., 2023)	22.2	22.7	23.6	10.2	27.2	27.9	23.6	19.2	23.9	13.5	12.7	26.3	21.4
miniGPT4-7B (Zhu et al., 2023)	23.1	18.6	26.0	13.4	30.4	30.2	28.1	21.0	24.7	16.2	16.7	25.4	17.9
LLaVAR-13B (Zhang et al., 2023b)	25.2	21.9	25.0	16.7	34.8	30.7	24.2	22.1	23.0	13.5	15.3	42.6	21.9
InstructBLIP-7B (Dai et al., 2024)	25.3	23.1	20.7	18.3	32.3	35.2	21.8	27.1	20.7	18.9	20.4	33.0	23.1
LLaVA-13B (Liu et al., 2023)	26.1	26.8	29.3	16.1	32.3	26.3	27.3	20.1	28.8	24.3	18.3	37.3	25.1
SPHINX-V1-13B (Lin et al., 2023b)	27.5	23.4	23.1	21.5	39.9	34.1	25.6	28.1	23.4	16.2	17.4	40.2	23.6
LLaVA-1.5-13B (Liu et al., 2024)	27.6	-	-	-	-	-	-	-	-	-	-	-	-
LLaVA-1.5-13B† (Liu et al., 2024)	27.7	23.8	22.7	18.3	40.5	30.2	25.3	26.4	22.8	21.6	26.4	35.3	23.6
OmniLMM-12B (OpenBMB, 2024)	34.9	45.0	17.8	26.9	44.9	39.1	23.1	32.3	20.9	18.9	27.8	45.9	44.2
SPHINX-V2-13B (Lin et al., 2023b)	36.7	**54.6**	16.4	23.1	41.8	**43.0**	20.6	33.4	17.6	**24.3**	21.5	43.4	**51.5**
G-LLaVA-13B (Gao et al., 2023a)	-	-	56.7	-	-	-	-	-	-	-	-	-	-
Math-LLaVA-DS	38.2	33.5	47.2	41.4	36.7	34.6	38.4	34.3	45.6	18.9	33.3	45.9	35.2
Math-LLaVA	**46.6**	37.2	**57.7**	**56.5**	**51.3**	33.5	**53**	**40.2**	**56.5**	16.2	**33.3**	**49.2**	43.9

Math-LLaVA

massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning

Evaluation experiments using the ***MMMU benchmark*** (generalization capability) sub-domains

Model	MMMU	Art & Design	Business	Sci.	Health & Med.	Human. & Social Sci.	Tech. & Eng.
Random Chance	22.1	29.2	24.7	18.0	20.7	20.0	21.4
Frequent Guess	26.8	23.3	**29.3**	27.3	30.0	25.8	24.8
miniGPT4-7B	26.8	29.2	21.3	28.7	30.7	29.2	23.8
mPLUG-Owl-7B	32.7	45.8	24.7	22.7	32.0	45.8	31.0
SPHINX-13B	32.9	48.3	24.7	26.7	30.7	50.0	26.2
InstructBLIP-7B	32.9	40.0	28.0	**32.7**	28.7	47.5	27.1
LLaVA-1.5-13B	36.4	51.7	22.7	29.3	38.7	53.3	31.4
Math-LLaVA-DS	36.9	**55.0**	24.7	23.3	38.7	56.7	32.4
Math-LLaVA	**38.3**	53.3	24.7	30.7	**38.7**	**58.3**	**33.3**

significantly outperforms the base model, LLaVA-1.513B, as well as several other open-source MLLMs on all six

UniMath: A Foundational and Multimodal Mathematical Reasoner. EMNLP 2023

Takeaways

Data is all you need !

Questions

(1) Paper: UniMath: A Foundational and Multimodal Mathematical Reasoner

Question: This paper did tests on three different tasks, mainly (symbol pre-processing, image tokenization and COT). could these three tasks be done at once to gain better results ?

(2) Paper: UniMath: A Foundational and Multimodal Mathematical Reasoner

Question: In table 1, section 3.4: What do you think why prompt-based LLMs were excluded during model comparison? It seems the paper did not justify the reason.

(3) Paper: UniMath: A Foundational and Multimodal Mathematical Reasoner

Question: Considering the breakthrough of recent DeepSeek-R1, why multi-modality is so important in these math reasoning papers except the geometry problems?

Thank you