Sightation Counts: Leveraging Sighted User Feedback in Building a BLV-aligned Dataset of Diagram Descriptions

Wan Ju Kang　Eunki Kim　Na Min An　Sangryul Kim
Haemin Choi　Ki Hoon Kwak　James Thorne

KAIST AI Kim Jaechul Graduate School　SUNGKYUNKWAN UNIVERSITY　YONSEI UNIVERSITY

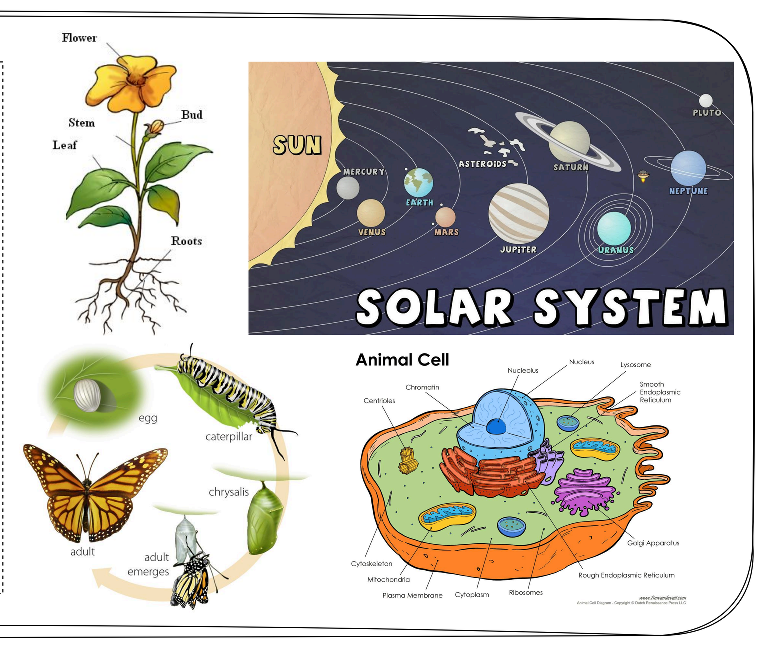
{soarhigh, eunkikim, naminan, sangryul, thorne}@kaist.ac.kr
chm1009@g.skku.edu
kihoon090@yonsei.ac.kr

The 63rd Annual Meeting of the Association for Computational Linguistics / July 27 - August 1, 2025 / Vienna, Austria

ACL 2025 VIENNA JULY 27 - AUGUST 1

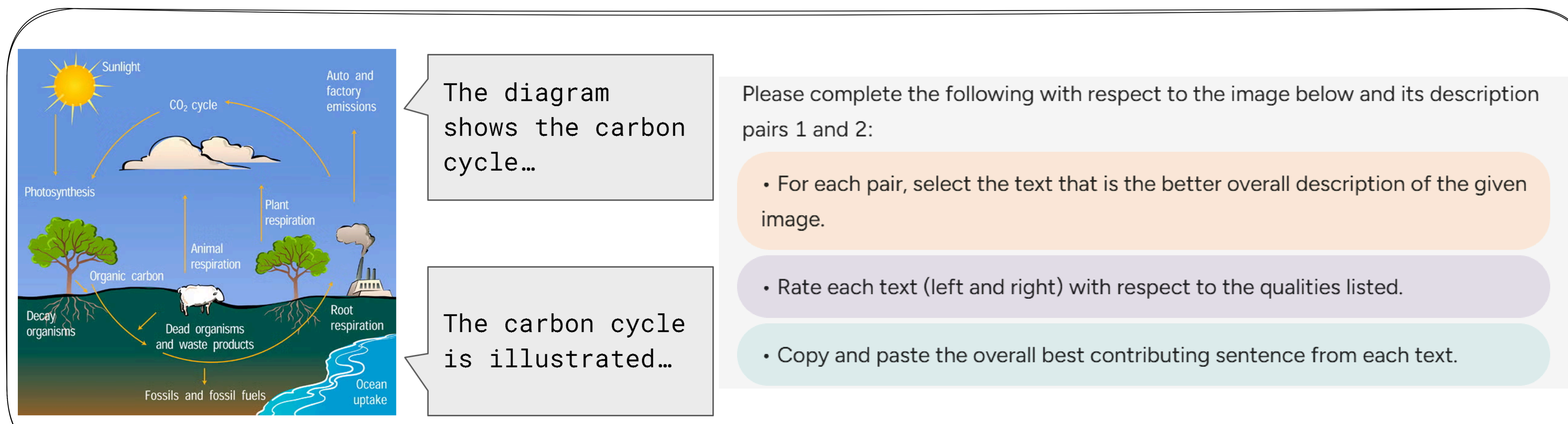
caption: cavendish bananas are the main commercial banana cultivars sold in the world market.

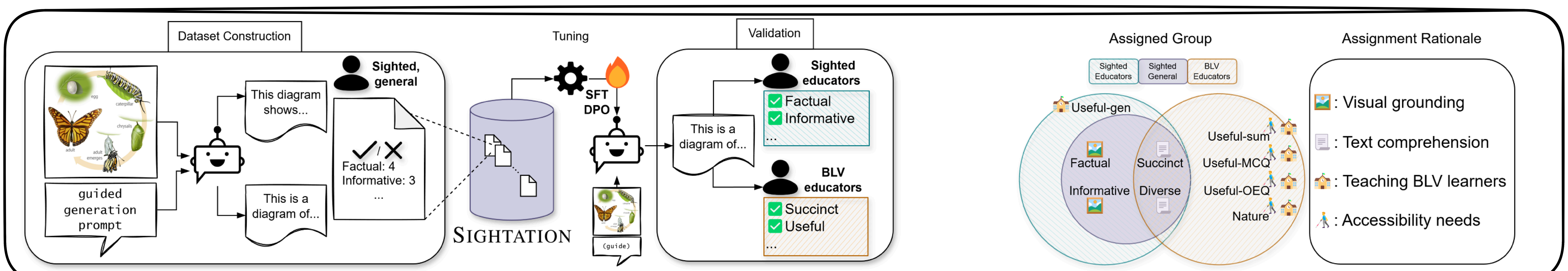
description: grocery store photo of several bunches of bananas

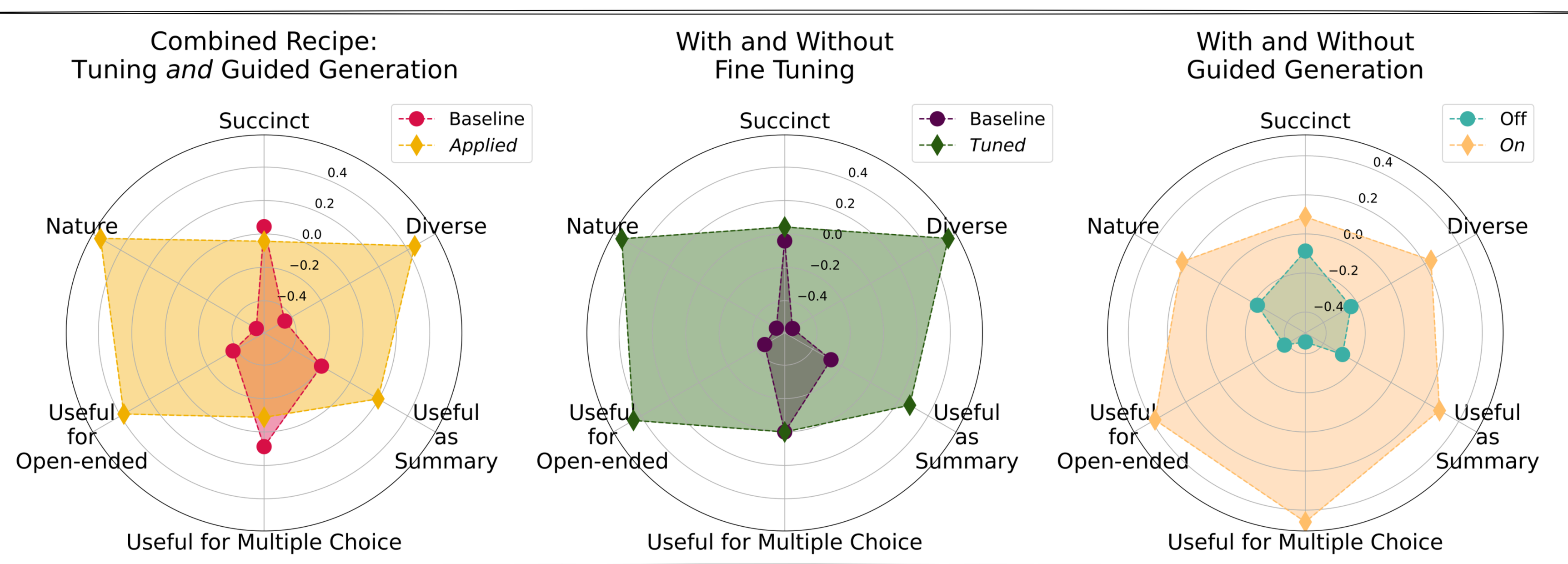


- Existing vision-language models are often trained for generating captions.
- This leaves out blind and low-vision individuals in need of descriptions.
- We created a dataset of diagram descriptions for training VLMs, driving them to generate more BLV-aligned text.

- We let VLMs generate descriptions then had them assessed by crowdworkers.
- Process leverages sighted user feedback for cost-effective, bias-reduced supervision.
- Dataset quality was validated by BLV educators at schools for the blind.





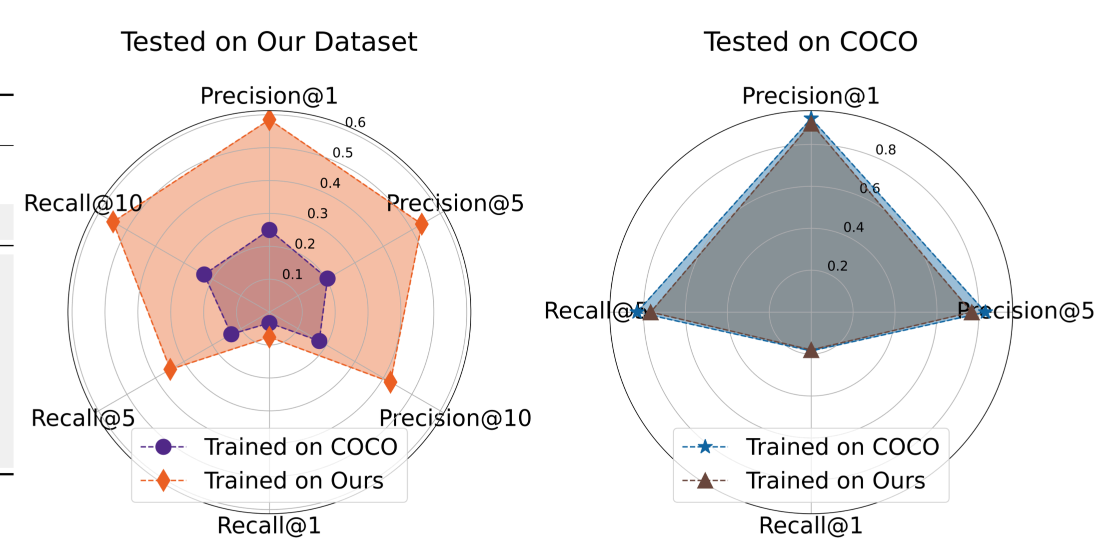


- We trained VLMs on our dataset and measured the effectiveness of the training with BLV and sighted educators across 9 quality aspects.
- Shown are the 6 aspects rated by BLV educators.
- Fine-tuned 2B model shows significant gain in usefulness and diversity.

- We also tested our dataset against existing datasets.
- BLIP2 trained on our data generalizes well to COCO.
- However, COCO-trained BLIP2 performs poorly on our dataset.

2-way Cross-validation of **BLIP-2**

Train set	N/A (Pre-trained)		COCO		SIGHTATIONRETRIEVAL (Ours)	
Test set	COCO	Ours	COCO	Ours	COCO	Ours
Recall@1	0.171	0.048	0.185	0.033	0.180	0.076
Recall@5	0.767	0.210	0.831	0.134	0.766	0.348
Recall@10	—	0.340	—	0.229	—	0.549
Precision@1	0.856	0.371	0.924	0.250	0.900	0.585
Precision@5	0.767	0.324	0.831	0.204	0.766	0.535
Precision@10	—	0.263	—	0.175	—	0.425



Aspect	**Combined Effect Size** 2B	7B
Succinct	-0.09	1.69
Diverse	0.90	0.46
Useful-Sum	0.39	0.53
Useful-MCQ	-0.18	0.20
Useful-OEQ	0.76	0.00
Average	0.36	0.58
Nature	1.08	-2.38

	Tuning Effect Size			
Aspect	2B	2B+GG	7B	7B+GG
Succinct	0.06	0.08	0.37	-0.11
Diverse	0.87	1.08	-0.06	0.00
Useful-Sum	0.20	0.55	0.14	0.36
Useful-MCQ	0.29	0.00	-0.54	0.00
Useful-OEQ	1.01	0.90	-0.74	-0.19
Average	0.49	0.52	-0.17	0.01
Nature	1.49	1.06	-3.14	-0.31

	Guided Generation Effect Size		
Aspect	GPT	2B Base	2B DPO
Succinct	0.18	-0.17	0.17
Diverse	-0.13	-0.13	0.47
Useful-Sum	0.48	-0.17	0.57
Useful-MCQ	0.13	-0.20	0.92
Useful-OEQ	0.76	-0.07	0.77
Average	0.28	-0.15	0.58
Nature	0.33	0.08	3.17

Experiment ID		**Assessments for**	
Description Generators	**Metrics**	**Desc[chartgemma]**	**Desc[q2bsft]**
Experiment 3c CHARTGEMMA (3B) vs. FINE-TUNED QWEN2-VL-2B-INSTRUCT	CLIP Score	0.450	**0.550**
	SigLIP Score	0.872	**0.940**
	BLIP-2 Retrieval Score	**0.511**	0.490
	Self-BLEU	**0.305**	0.280
	PAC-Score	0.705	**0.716**
	LongClip-B	0.316	**0.684**
	LongClip-L	**0.559**	0.441
	· VLM-as-a-Judge Evaluation Average	2.951	**3.860**
	Factuality	3.068	**4.119**
	Informativeness	2.848	**3.967**
	Succinctness	3.253	**3.925**
	Diversity	2.635	**3.428**

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2024-00457882, AI Research Hub Project and No.RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2024-00406715).

Sightation@ACL2025
https://hf.co/Sightation

ACL 2025 VIENNA JULY 27 - AUGUST 1