Data Selection Matters: Towards Robust Instruction Tuning of Large Multimodal Models

Xu Yang

City University of Hong Kong

Chen Liu*

City University of Hong Kong

Ying Wei*

Zhejiang University

*Corresponding authors

Outline

Background

Related Work

Motivation

Proposed Methods

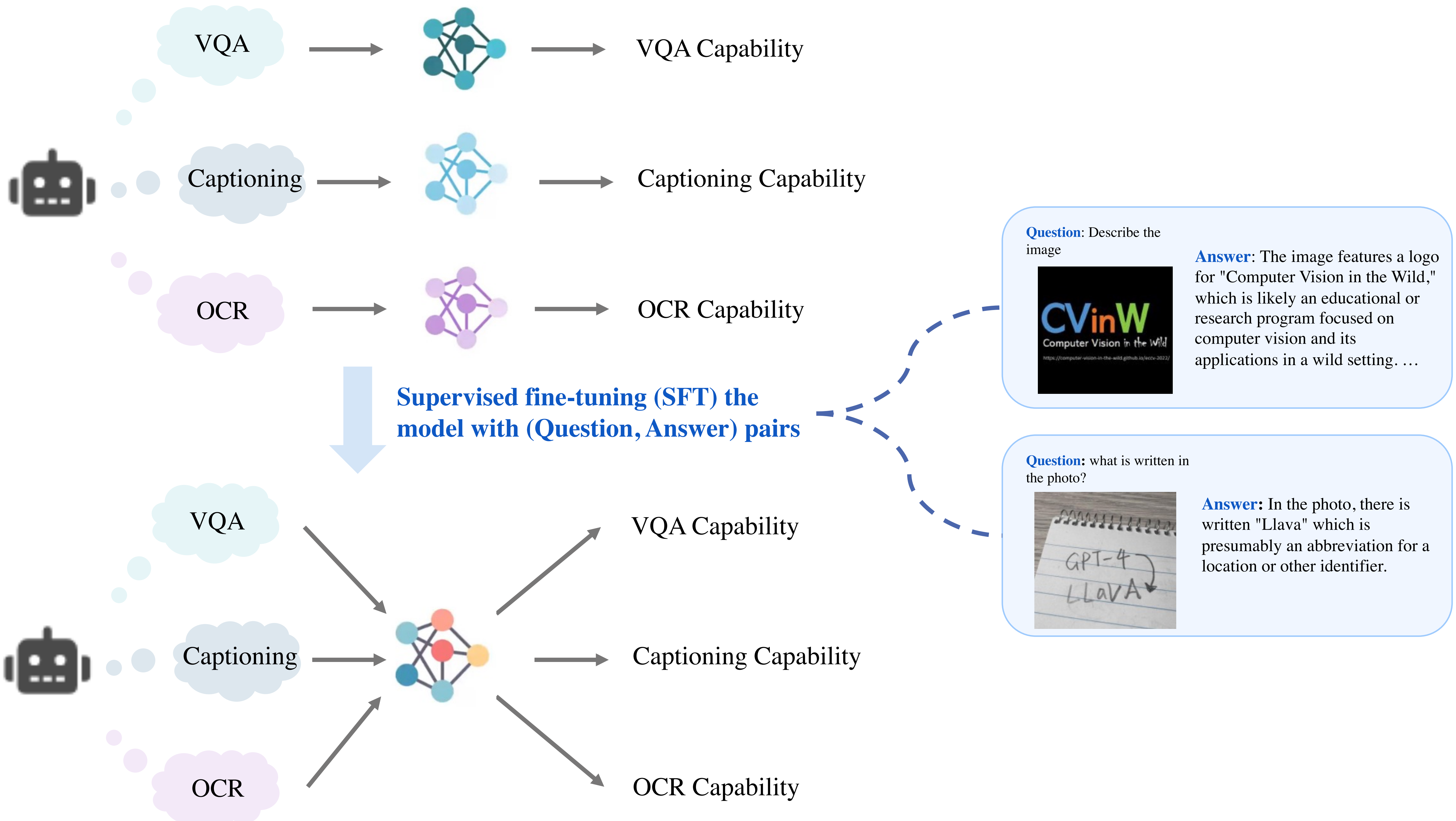
Framework

Experiments

Analysis

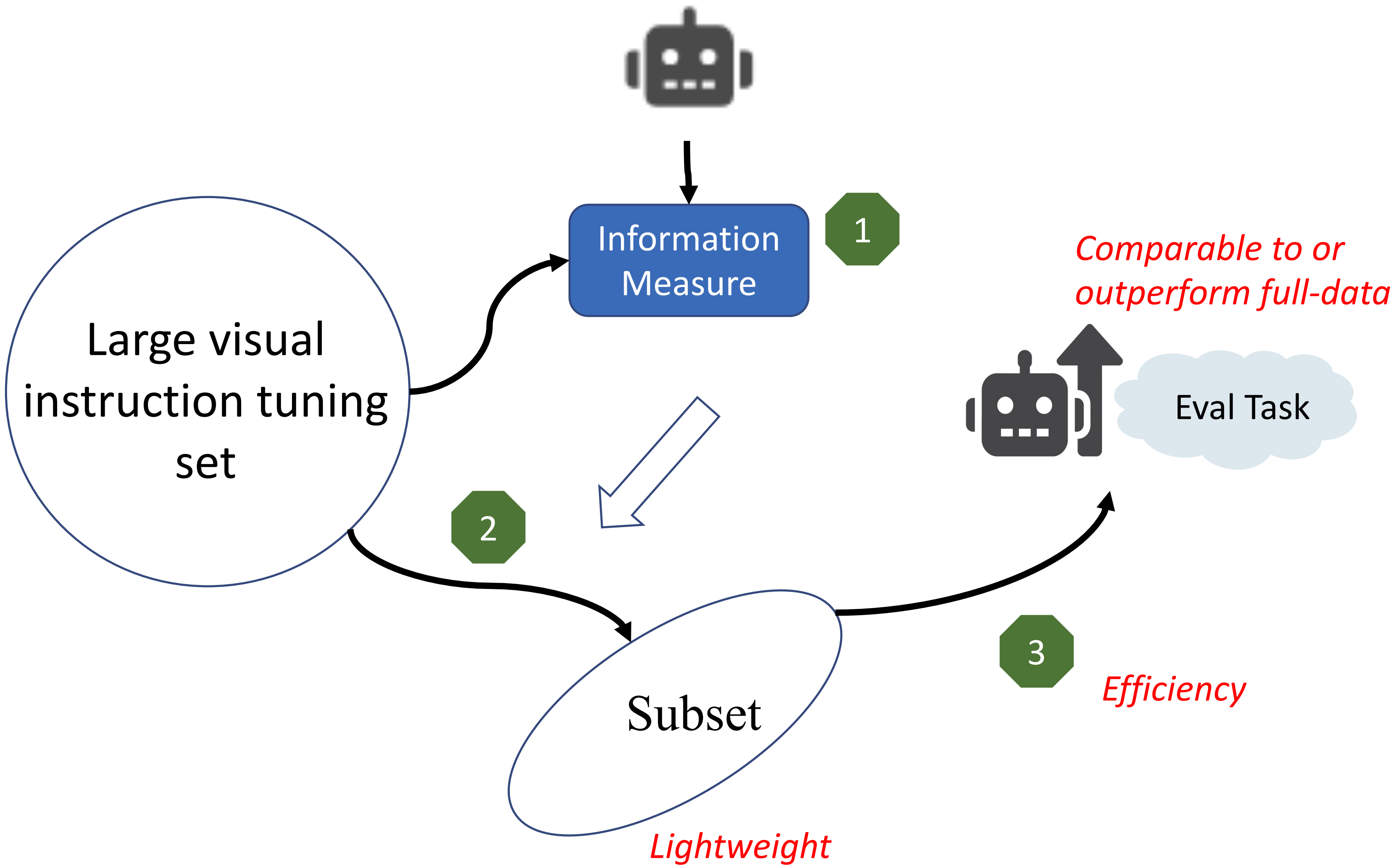
Visual Instruction Tuning for Aligning LMMs

➤ Visual instruction tuning refers to enable an LMM to *understand and act upon visual instructions*



Data Selection for Visual Instruction Tuning

Definition & Goal of Data Selection

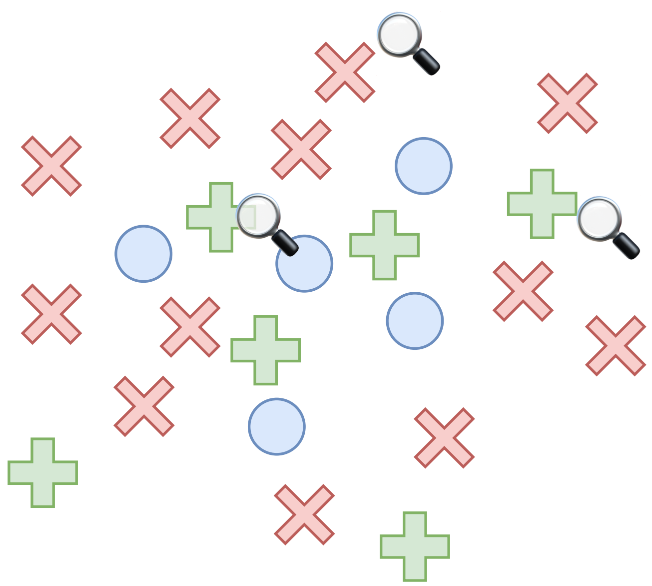


1. Estimate the training data information importance based on quality or diversity
2. Select a subset of data according to the importance score
3. Train the model on the built subset

Data Selection for Visual Instruction Tuning

Main Categories

Methods	Information Proxy			Objective	
	Score-based	Feature-based	Gradient-based	Quality	Diversity
EL2N (Paul et al., 2021)	✓	-	-	✓	-
Perplexity (Marion et al., 2023)	✓	-	-	✓	-
SemDeDup (Abbas et al., 2023)	-	✓	-	-	✓
COINCIDE (Lee et al., 2024)	-	✓	-	-	✓
LESS (Xia et al., 2024)	-	-	✓	✓	-



Training corpus $\mathcal{D}$

Singe Score Metric

- ☐ Error L2-Norm score: $\| p(x) - y \|_2$
- ☐ Prediction perplexity: $\exp(-\mathbb{E}[\log p(x)])$

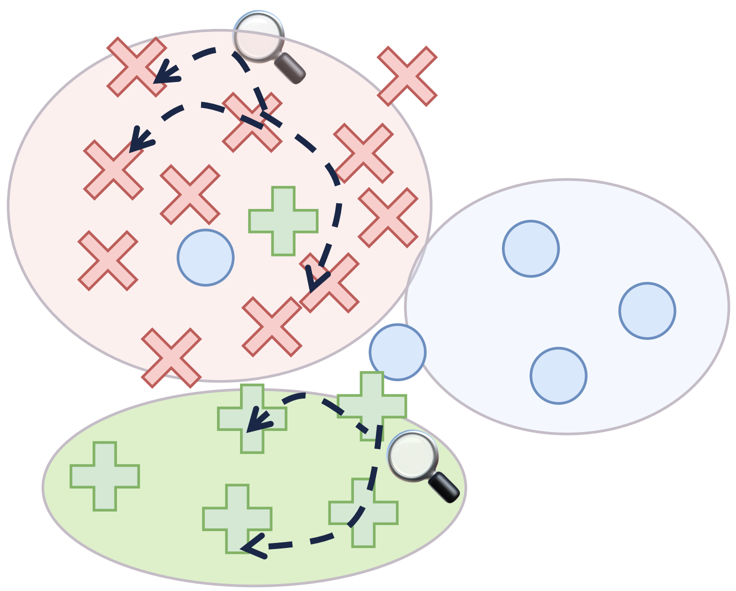
- $p(\cdot)$: reference model prediction
- y: ground truth

Easy to overlook the diversity of data!

Data Selection for Visual Instruction Tuning

Main Categories

Methods	Information Proxy			Objective	
	Score-based	Feature-based	Gradient-based	Quality	Diversity
EL2N (Paul et al., 2021)	✓	-	-	✓	-
Perplexity (Marion et al., 2023)	✓	-	-	✓	-
SemDeDup (Abbas et al., 2023)	-	✓	-	-	✓
COINCIDE (Lee et al., 2024)	-	✓	-	-	✓
LESS (Xia et al., 2024)	-	-	✓	✓	-



Training corpus $\mathcal{D}$

Clustering

1. Clustering the feature embedding
2. Reduce redundancy

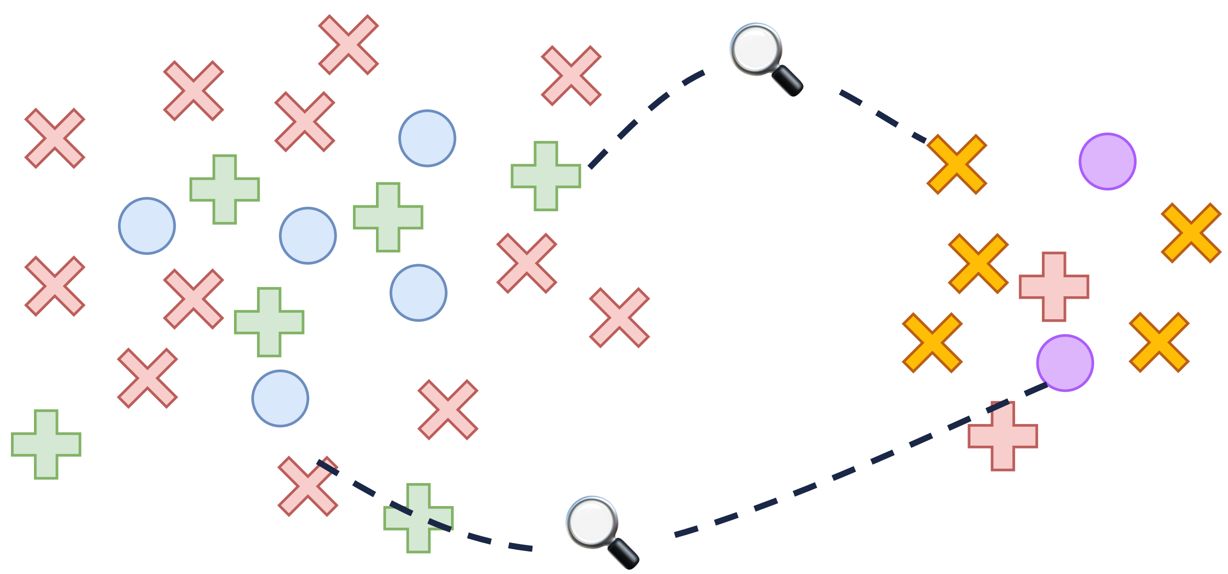
- Remove *semantically duplicated* data
- Prioritize selection from *lower* cluster density

Require a good feature representation space!

Data Selection for Visual Instruction Tuning

Main Categories

Methods	Information Proxy			Objective	
	Score-based	Feature-based	Gradient-based	Quality	Diversity
EL2N (Paul et al., 2021)	✓	-	-	✓	-
Perplexity (Marion et al., 2023)	✓	-	-	✓	-
SemDeDup (Abbas et al., 2023)	-	✓	-	-	✓
COINCIDE (Lee et al., 2024)	-	✓	-	-	✓
LESS (Xia et al., 2024)	-	-	✓	✓	-



Training corpus $\mathcal{D}_{\mathrm{tr}}$ Test corpus $\mathcal{D}_{te}$

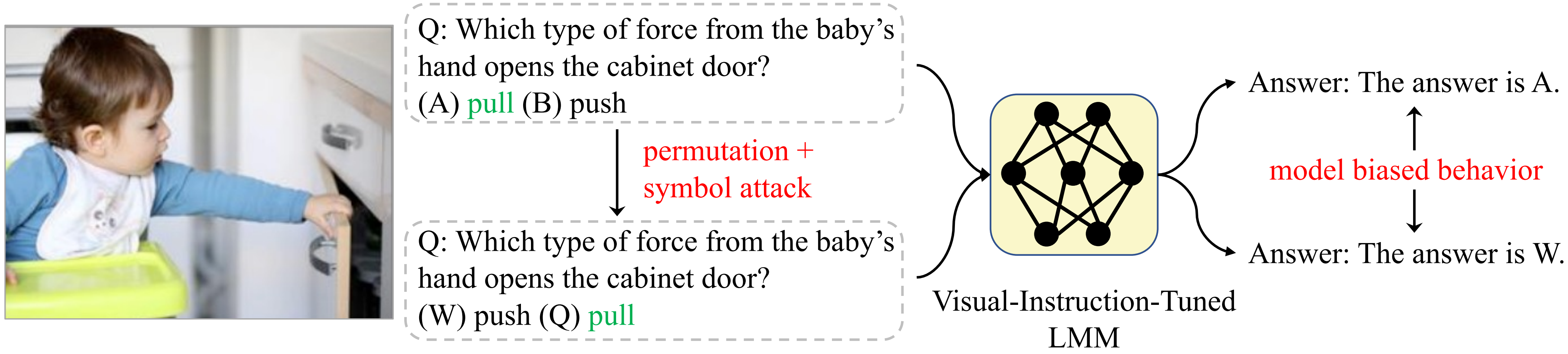
Computationally expensive!
Requirement of Downstream Data!

Influence Function

$$\mathrm{Inf}_{\mathrm{Adam}}(\boldsymbol{z}, \boldsymbol{z}') \triangleq \sum_{i=1}^{N} \bar{\eta}_i \mathrm{cosine}\big(\nabla \ell(\boldsymbol{z}'; \boldsymbol{\theta}_i), \Gamma(\boldsymbol{z}, \boldsymbol{\theta}_i)\big)$$

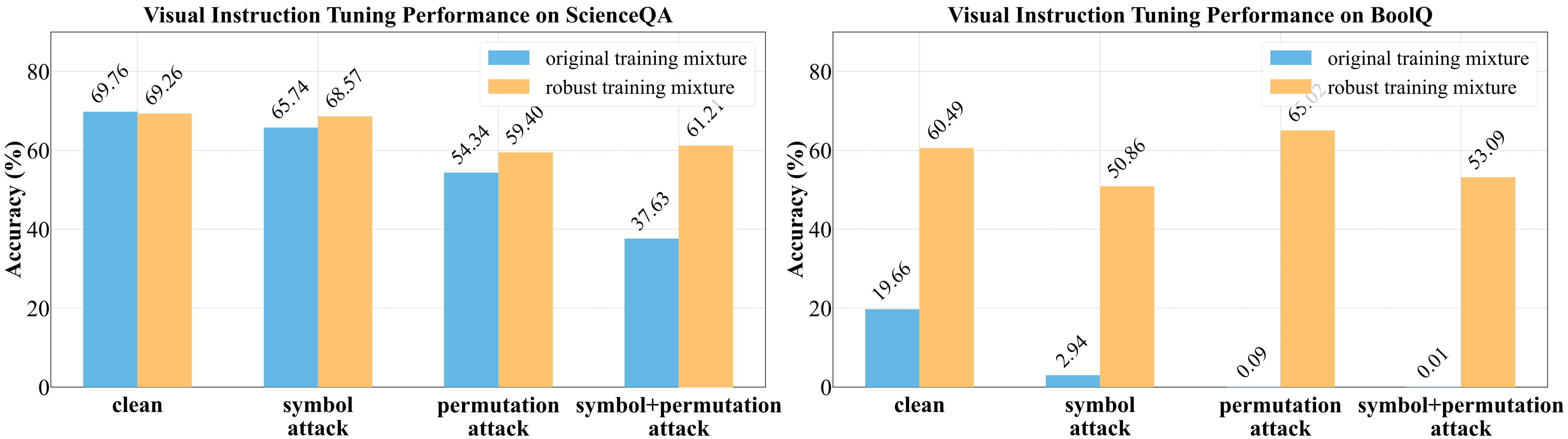
- $\boldsymbol{z}$: training sample from $\mathcal{D}_{\mathrm{tr}}$
- $\boldsymbol{z}'$: sample from target task from $\mathcal{D}_{te}$
- $\bar{\eta}_i$: learning rate at the i-th epoch
- N: number of epoch
- $\tilde{\Gamma}$: gradient calculated by Adam

Dataset Biases



(a)

(a) Model biased behaviors



(b)

(b) Robustness on a multimodal task (left) and on a pure-text task (right) under symbol and permutation attacks

*This motivates us to explore **alternative data selection objectives**, aiming to design carefully curated training mixtures that **go beyond efficiency, quality, and diversity.***

- The results highlight a ***significant decline in accuracy*** under simple input perturbations, and ***text-only catastrophic forgetting*** further amplifies the vulnerability.
- We hypothesis such vulnerabilities are often attributed to ***dataset biases*** that inadvertently encourage shortcut learning or spurious correlations.

Data Selection for Robustness

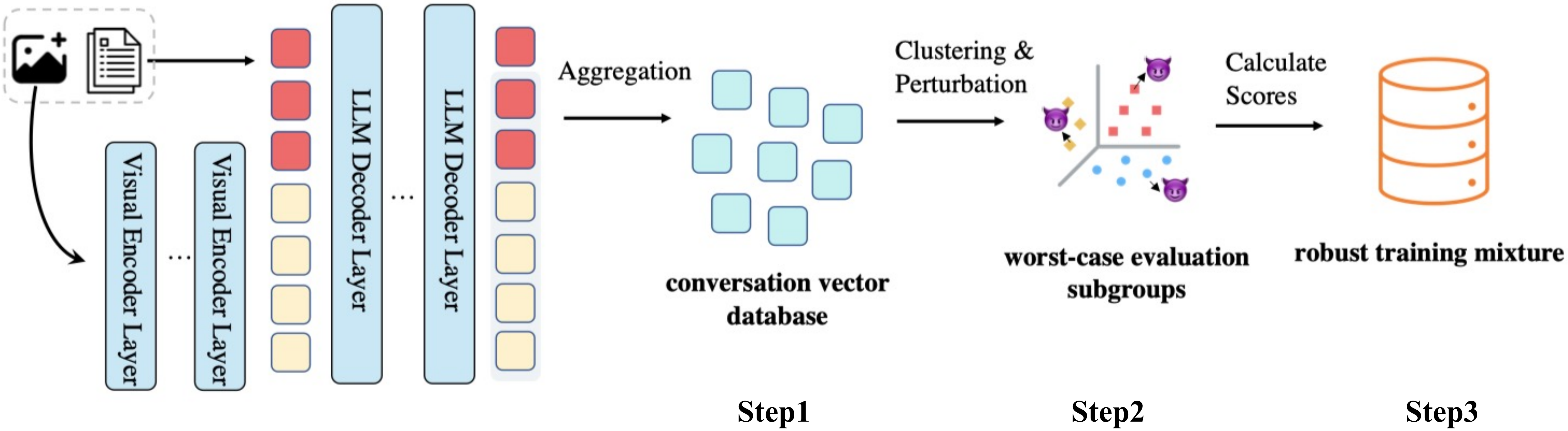
Table 1: Comparisons of existing visual instruction-following data selection methods with large multimodal models. *Information Proxy* indicates the representation used to compute the information measure. *Objective* means the selection goal emphasized when ranking samples. *Task-Aware Selection* denotes methods explicitly target a specific task. *Downstream-Data-free* marks no downstream-task samples are required during selection.

Method	Information Proxy	Objective	Task-Aware Selection	Downstream-Data-free
LESS [107]	Gradient	Quality	✓	✗
ICONS [106]	Gradient	Quality	✓	✗
TIVE [68]	Gradient	Diversity	✓	✓
COINCIDE [51]	Feature	Diversity	✗	✓
ARDS (Ours)	Feature	Robustness	✓	✓

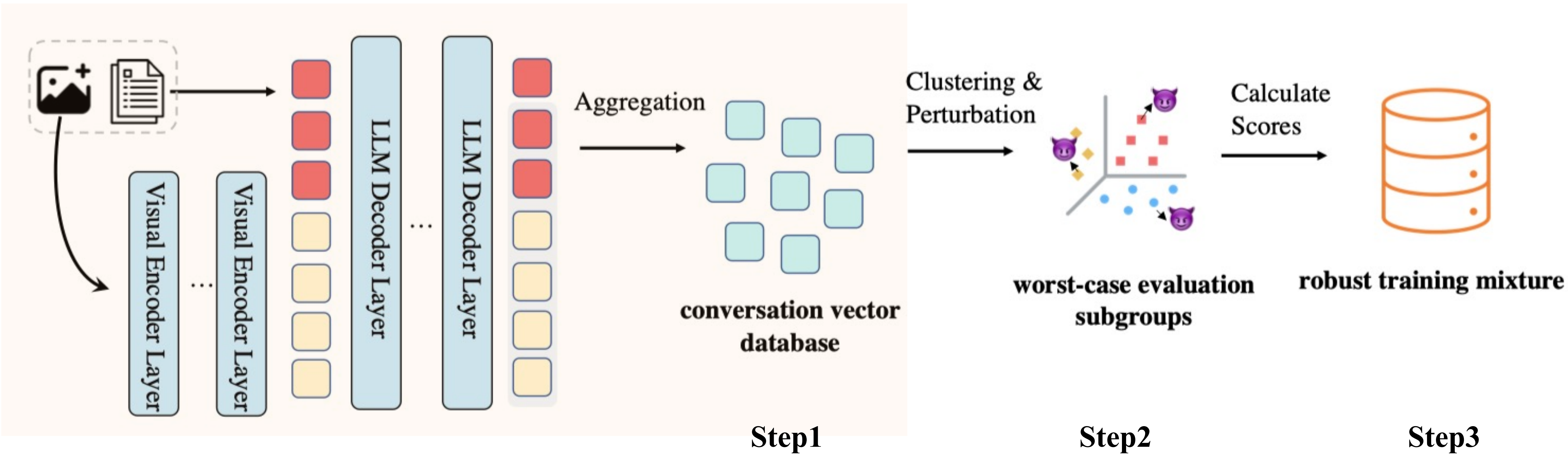
We want to propose a data selection method to:

- ✓ **Curate a robust training mixture**
- ✓ **Gradient-free**
- ✓ **Do not require a well-trained reference model**
- ✓ **Do not require few-shot examples in downstream tasks**

The Proposed ARDS



The Proposed ARDS



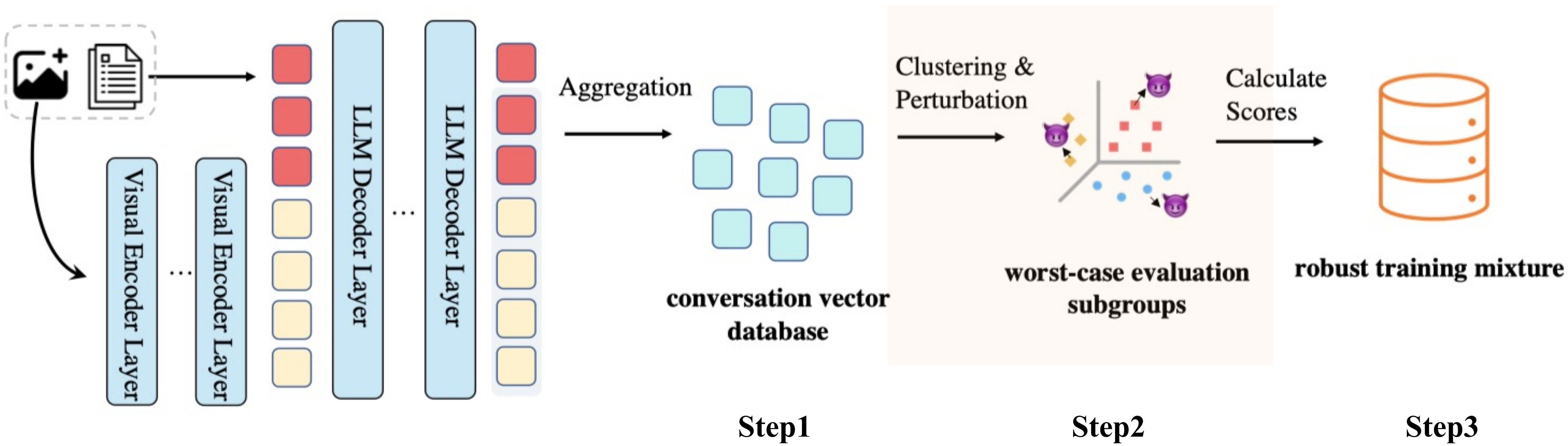
Conversation Vector Database

$$\hat{\mathbf{H}} = \sum_{t=1}^{L-1} \mathbf{A}_{L,t} \cdot \mathbf{H}_t$$
$$r_i = [\mathbf{H}_L; \hat{\mathbf{H}}]$$

- for an input with L tokens
- $\mathbf{H}_t$: token embeddings
- **A:** attention-score matrix

➢ Introduce the ***attention-score weighted mechanism*** to aggregate the conversation vector from the token-level embeddings based on their relevance

The Proposed ARDS



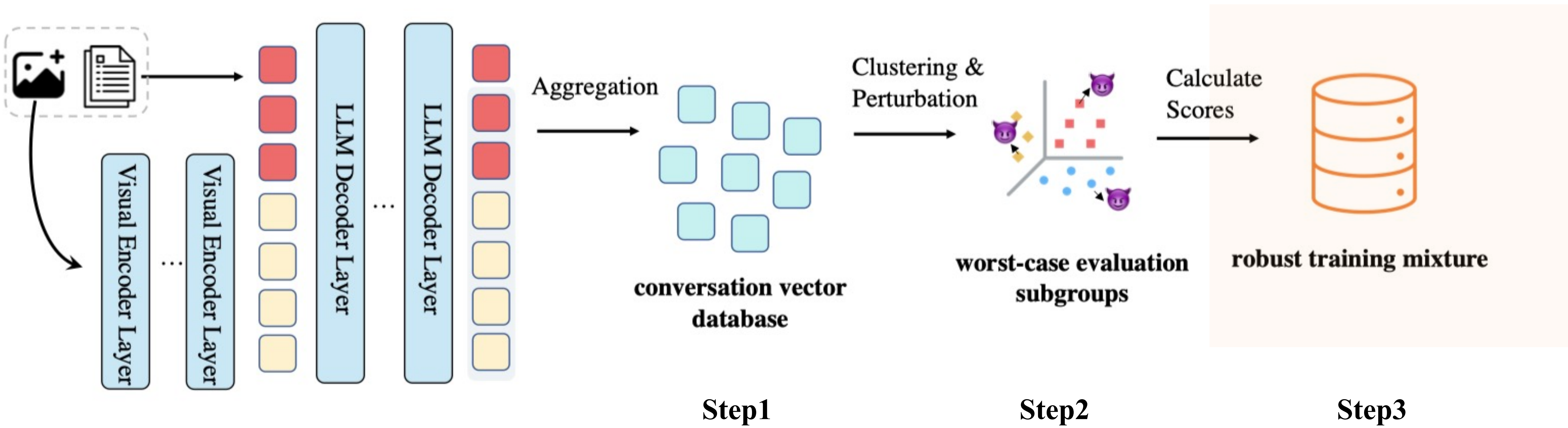
Worst-case Evaluation Subgroups

$$\mathcal{S}_m = \text{top}_B\{\mathbf{x} \in \mathcal{C}_m : |\ell(\mathbf{x}) - \ell(\mathbf{x}')|\}$$

- $\mathcal{C}_m$: m-th subgroup
- ℓ: cross-entropy loss
- $\mathbf{x}'$: corrupted conversation

- *Cluster* M subgroups over the built conversation vector database
- Inject ***task-aware perturbations*** designed to improve robustness against specific attacks (i.e., symbol attack, permutation attack)
- Retrieve top_B conversations with the ***largest loss difference***

The Proposed ARDS



Robust Training Mixture

$$d_{i\mathcal{S}_m} = \frac{1}{B}\sum_{j\in\mathcal{S}_m} \cos\left(r_{\text{tr}}^{i}; r_{\mathcal{S}_m}^{j}\right)$$

$$\mathcal{I}(x_i) = \frac{\sum_{m=1}^{M} \exp\left(\ell_{\mathcal{S}_m}\right)\cdot d_{i\mathcal{S}_m}}{\sum_{m'=1}^{M} \exp\left(\ell_{\mathcal{S}_{m'}}\right)}$$

- cos: cosine similarity
- $\ell_{\mathcal{S}_m}$: average loss
- $\mathcal{I}$: information score

- Quantify the importance of each training sample
- Weight each similarity by the ***subgroup's difficulty*** using a SoftMax normalization
- Select training conversations with the highest scores to build the final ***robust training mixture***

Experiment Results

Zero-shot robust accuracies of LLaVA-1.5-7B against SA: symbol attacks; PA: permutation attacks

Selection Method	Data Percentage	ScienceQA Clean	ScienceQA PA	ScienceQA SA	ScienceQA SA + PA	ScienceQA Avg.	SEED-Bench Clean	SEED-Bench PA	SEED-Bench SA	SEED-Bench SA + PA	SEED-Bench Avg.	MMBench-EN Clean	MMBench-EN PA	MMBench-EN SA	MMBench-EN SA + PA	MMBench-EN Avg.	MMBench-CN Clean	MMBench-CN PA	MMBench-CN SA	MMBench-CN SA + PA	MMBench-CN Avg.
Full	100%	69.76	54.34	65.74	37.63	56.87	59.65	41.92	54.83	22.40	44.69	74.84	61.15	69.39	41.09	61.62	69.95	52.34	65.33	34.90	55.63
Random	30%	69.76	52.60	59.44	23.75	51.39	56.84	35.74	46.58	12.73	37.97	74.20	57.75	65.49	31.83	57.32	69.76	49.50	63.78	34.33	54.34
LESS-SciQA [107]	30%	68.42	55.63	64.70	34.95	55.93	55.82	36.30	52.32	18.19	40.66	72.14	57.89	67.54	34.51	58.02	67.38	48.49	62.05	30.68	52.15
RHO-LOSS [76]	30%	64.01	36.89	59.44	21.42	45.44	53.97	25.07	48.36	11.26	34.67	70.82	49.90	66.94	32.83	55.12	68.05	43.68	65.03	31.90	52.16
COINCIDE [51]	30%	67.72	52.21	61.08	28.06	52.27	57.49	36.02	48.93	15.88	39.58	73.78	58.65	68.10	37.65	59.54	69.48	49.64	64.84	35.97	54.98
ARDS (ours)	30%	69.26	**59.40**	**68.57**	**47.60**	**61.21**	58.11	40.73	**56.83**	**31.52**	**46.80**	74.43	61.03	**72.37**	**53.22**	**65.26**	**70.48**	**53.73**	**68.98**	**46.02**	**59.80**

Selection Method	Data Percentage	A-OKVQA Clean	A-OKVQA PA	A-OKVQA SA	A-OKVQA SA + PA	A-OKVQA Avg.	MMMU Clean	MMMU PA	MMMU SA	MMMU SA + PA	MMMU Avg.	ARC-e Clean	ARC-e PA	ARC-e SA	ARC-e SA + PA	ARC-e Avg.	BoolQ Clean	BoolQ PA	BoolQ SA	BoolQ SA + PA	BoolQ Avg.
Full	100%	80.52	72.31	78.34	55.02	71.54	35.06	10.15	33.65	4.84	20.92	36.76	11.11	25.25	0.83	18.48	37.77	23.64	4.53	0.09	16.50
Random	30%	78.25	66.29	70.13	35.72	62.59	34.00	9.21	**35.77**	5.43	21.10	38.95	12.38	33.99	1.36	21.67	55.93	29.79	37.22	3.39	31.58
LESS-SciQA [107]	30%	78.60	66.72	74.41	45.94	66.42	37.43	11.81	33.53	4.49	21.82	37.86	13.57	35.18	3.03	22.41	57.58	40.86	39.36	3.27	35.27
RHO-LOSS [76]	30%	76.86	55.02	71.00	37.64	60.13	34.00	5.31	32.23	3.19	18.68	38.21	5.49	34.39	1.27	19.84	43.79	8.41	37.80	0.61	22.65
COINCIDE [51]	30%	77.55	65.59	72.66	44.10	64.97	**37.90**	9.80	33.29	3.54	21.13	38.25	11.86	36.06	2.64	22.20	55.14	29.20	41.01	5.20	32.63
ARDS (ours)	30%	78.34	71.09	77.64	**64.72**	**72.95**	37.54	**12.75**	34.24	**6.97**	**22.88**	**39.92**	**16.95**	**37.15**	**8.26**	**25.57**	**58.62**	**46.45**	**46.85**	**17.25**	**42.29**

- With only 30% of the original data, our method ARDS **consistently holds the advantage to boost robustness** comparing with baseline methods

Cross-Architecture-Scale Transferability

Proxy Model	Target Model	Selection Method	Data Percentage	ScienceQA Clean	ScienceQA PA	ScienceQA SA	ScienceQA SA + PA	ScienceQA Avg.	SEED-Bench Clean	SEED-Bench PA	SEED-Bench SA	SEED-Bench SA + PA	SEED-Bench Avg.	MMBench-EN Clean	MMBench-EN PA	MMBench-EN SA	MMBench-EN SA + PA	MMBench-EN Avg.	MMBench-CN Clean	MMBench-CN PA	MMBench-CN SA	MMBench-CN SA + PA	MMBench-CN Avg.
-	LLaVA-1.5 (13B)	**Full**	100%	71.05	57.21	64.20	37.58	57.51	61.12	43.85	56.19	23.08	46.06	76.02	64.06	71.73	47.79	64.90	72.88	57.36	68.68	37.77	59.17
-	LLaVA-1.5 (13B)	**Random**	30%	70.25	54.69	63.76	31.33	55.01	59.08	39.06	52.09	16.17	41.60	75.70	59.92	69.74	39.50	61.22	72.23	53.98	65.28	31.74	55.81
LLaVA-1.5 (7B)	LLaVA-1.5 (13B)	**ARDS (ours)**	30%	**72.58**	**60.19**	**66.14**	**41.99**	**60.22**	59.94	**43.98**	**57.58**	**30.76**	**48.07**	**76.41**	**64.24**	**72.95**	**52.60**	**66.55**	71.49	56.18	67.45	**40.06**	58.80

Proxy Model	Target Model	Selection Method	Data Percentage	A-OKVQA Clean	A-OKVQA PA	A-OKVQA SA	A-OKVQA SA + PA	A-OKVQA Avg.	MMMU Clean	MMMU PA	MMMU SA	MMMU SA + PA	MMMU Avg.	ARC-e Clean	ARC-e PA	ARC-e SA	ARC-e SA + PA	ARC-e Avg.	BoolQ Clean	BoolQ PA	BoolQ SA	BoolQ SA + PA	BoolQ Avg.
-	LLaVA-1.5 (13B)	**Full**	100%	82.36	73.28	80.70	62.88	74.80	38.25	14.29	35.77	6.14	23.61	18.36	0.53	14.58	0.09	8.39	19.66	2.94	0.09	0.01	5.68
-	LLaVA-1.5 (13B)	**Random**	30%	79.74	69.61	77.21	50.22	69.19	38.84	12.63	35.42	4.72	22.90	45.63	17.35	41.37	7.51	27.96	56.57	39.89	63.09	40.67	50.06
LLaVA-1.5 (7B)	LLaVA-1.5 (13B)	**ARDS (ours)**	30%	80.96	72.66	79.83	**63.41**	74.22	**40.50**	**15.94**	**38.37**	**8.74**	**25.89**	**45.98**	**22.57**	**42.07**	**12.12**	**30.69**	**60.49**	**50.86**	**65.02**	**53.09**	**57.37**

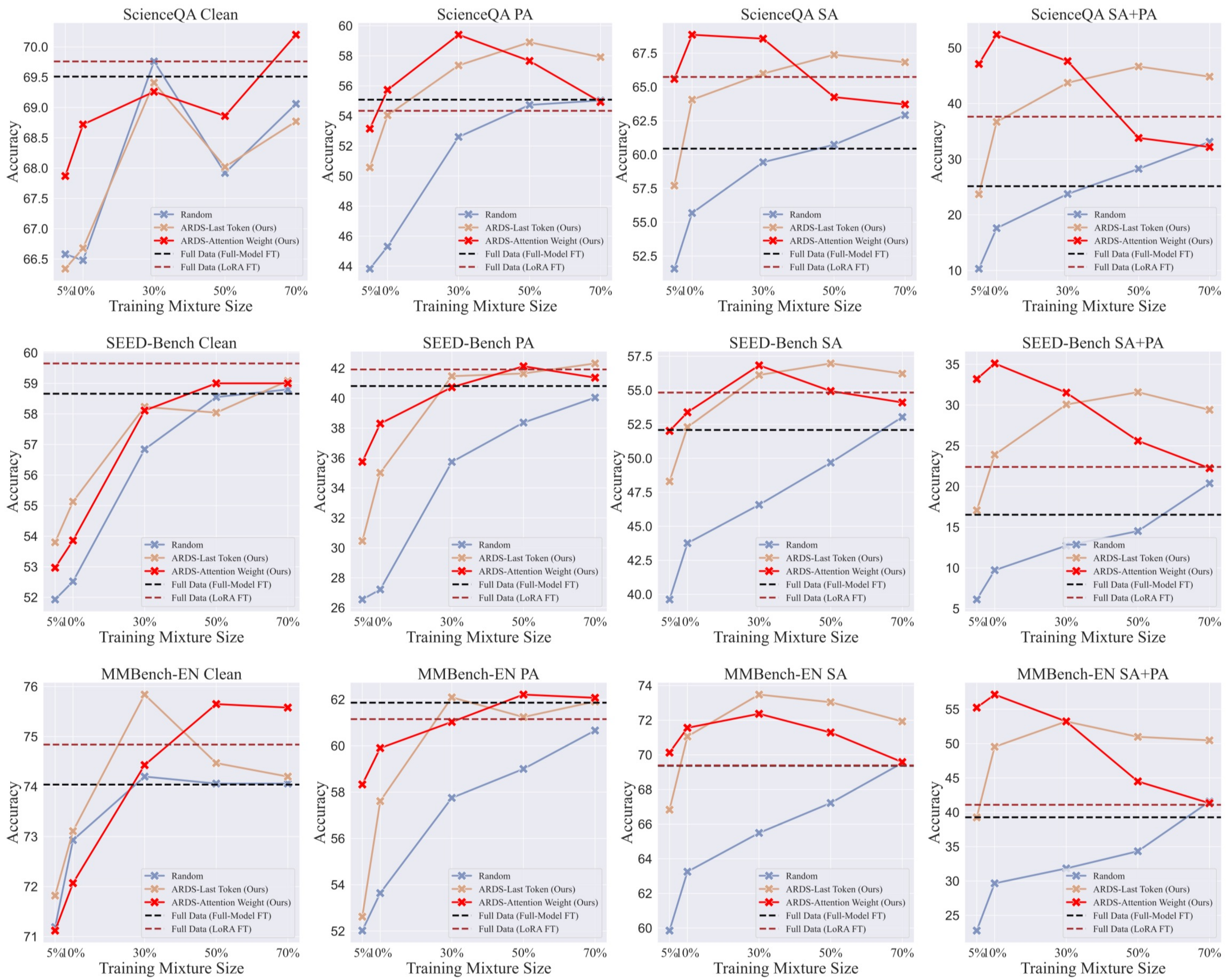
- The robust training mixture created with a smaller model (LLaVA-1.5-7B) can be **transferred effectively** to a larger model (LLaVA-1.5-13B)

Zero-shot robust accuracies of LLaVA-1.5-7B

Selection Method	Data Percentage	GQA Original	GQA OOD-All	GQA OOD-Head	GQA OOD-Tail	GQA Avg.
Full	100%	61.94	57.51	61.17	51.55	58.04
Random	50%	60.69	55.97	60.30	48.92	56.47
COINCIDE	50%	61.88	56.58	60.30	50.52	57.32
ARDS (ours)	50%	62.43	58.44	62.26	52.21	58.84

- ARDS improves robustness against **visual spurious correlation**

Experiment Results



Robust Accuracies (↑) across different sizes of training data

- Randomly removing training samples **does not necessarily** improve robustness
- Our method outperform baselines **across data scales**
- Why 30%? **best trade-off** between data efficiency and both clean and robust performance.

Analysis

Conversation Vector Variants

Conversation Vector	Data Percentage	ScienceQA Clean	ScienceQA PA	ScienceQA SA	ScienceQA SA + PA	ScienceQA Avg.	SEED-Bench Clean	SEED-Bench PA	SEED-Bench SA	SEED-Bench SA + PA	SEED-Bench Avg.	MMBench-EN Clean	MMBench-EN PA	MMBench-EN SA	MMBench-EN SA + PA	MMBench-EN Avg.	A-OKVQA Clean	A-OKVQA PA	A-OKVQA SA	A-OKVQA SA + PA	A-OKVQA Avg.
Last Token	10%	66.68	54.04	64.06	36.69	55.36	55.13	35.02	52.29	23.90	41.58	73.11	57.61	71.06	49.53	62.82	76.33	64.45	74.59	54.32	67.42
Attention Weight	10%	69.66	55.88	69.21	52.35	61.78	53.86	38.30	53.38	35.11	45.16	72.07	59.90	71.56	57.13	65.16	77.90	70.04	76.94	66.72	72.90
Last Token	30%	69.41	57.36	65.99	43.73	59.12	58.23	41.47	56.12	30.07	46.47	75.84	62.09	73.48	53.22	66.15	78.95	71.18	78.08	63.32	72.88
Attention Weight	30%	69.26	59.40	68.57	47.60	61.21	58.11	40.73	56.83	31.52	46.80	74.43	61.03	72.37	53.22	65.26	78.34	71.09	77.64	64.72	72.95

- Attention-weighted conversation vector consistently preserves more significant and useful semantics

Different Components for Worst-case Evaluation Subgroups

Worst-case Evaluation Subgroup Perturbation	Worst-case Evaluation Subgroup Clustering	Data Percentage	ScienceQA Clean	ScienceQA PA	ScienceQA SA	ScienceQA SA + PA	ScienceQA Avg.	SEED-Bench Clean	SEED-Bench PA	SEED-Bench SA	SEED-Bench SA + PA	SEED-Bench Avg.	MMBench-EN Clean	MMBench-EN PA	MMBench-EN SA	MMBench-EN SA + PA	MMBench-EN Avg.	A-OKVQA Clean	A-OKVQA PA	A-OKVQA SA	A-OKVQA SA + PA	A-OKVQA Avg.
✗	✗	30%	65.29	51.66	62.17	30.44	52.39	56.75	35.40	51.01	18.56	40.43	73.76	57.80	69.14	41.44	60.53	77.12	65.33	74.06	47.60	66.03
✓	✗	30%	67.43	54.34	64.35	36.49	55.65	**58.38**	40.42	56.24	26.57	45.40	74.15	60.89	71.96	49.36	64.09	**79.04**	70.92	76.77	59.56	71.57
✓	✓	30%	**69.26**	**59.40**	**68.57**	**47.60**	**61.21**	58.11	**40.73**	**56.83**	**31.52**	**46.80**	**74.43**	**61.03**	**72.37**	**53.22**	**65.26**	78.34	**71.09**	**77.64**	**64.72**	**72.95**

- First row: randomly sample the same number of samples
- Second row: retrieve top-MB samples from the training dataset with the largest loss difference

- Effectiveness of component to build the worst-case evaluation subgroups

Different Score Aggregation Strategies

Score Aggregation Strategy	Data Percentage	ScienceQA Clean	ScienceQA PA	ScienceQA SA	ScienceQA SA + PA	ScienceQA Avg.	SEED-Bench Clean	SEED-Bench PA	SEED-Bench SA	SEED-Bench SA + PA	SEED-Bench Avg.
Subgroup Maximum	30%	70.05	57.61	68.12	43.88	59.91	57.23	41.02	56.13	30.65	46.25
Subgroup Weighted Sum	30%	69.26	59.40	68.57	47.60	61.21	58.11	40.73	56.83	31.52	46.80

- Incorporating subgroup difficulty helps select training samples that more effectively target model-biased behaviors.

Transferability across large multimodal architectures

Proxy Model	Target Model	Selection Method	Data Percentage	ScienceQA Clean	ScienceQA PA	ScienceQA SA	ScienceQA SA + PA	ScienceQA Avg.	SEED-Bench Clean	SEED-Bench PA	SEED-Bench SA	SEED-Bench SA + PA	SEED-Bench Avg.	MMBench-EN Clean	MMBench-EN PA	MMBench-EN SA	MMBench-EN SA + PA	MMBench-EN Avg.
-	LLaVA-Mistral (7B)	**Full**	100%	73.03	60.78	68.32	42.79	61.23	59.22	39.65	56.62	28.98	46.11	77.04	62.05	73.30	47.05	64.86
-	LLaVA-Mistral (7B)	**Random**	30%	73.08	56.22	58.70	21.17	52.29	56.84	34.85	50.47	14.05	39.05	75.31	58.51	67.48	32.87	58.54
LLaVA-1.5 (7B)	LLaVA-Mistral (7B)	**ARDS**	30%	72.04	61.77	69.16	55.53	**64.63**	59.22	44.02	57.53	34.93	**48.93**	76.97	65.37	75.17	55.19	**68.18**
-	Qwen2.5-VL (7B)	-	-	77.05	63.71	67.08	33.71	60.38	48.61	24.72	53.09	10.60	34.25	71.31	52.48	72.14	35.16	57.77
-	Qwen2.5-VL (7B)	**Random**	30%	80.32	69.31	67.43	31.78	62.21	52.06	28.50	53.67	8.98	35.80	74.27	57.36	73.83	34.63	60.02
LLaVA-1.5 (7B)	Qwen2.5-VL (7B)	**ARDS**	30%	83.84	76.55	70.15	36.19	**66.68**	61.71	41.81	55.40	10.46	**42.35**	80.85	69.81	75.44	40.29	**66.60**

Proxy Model	Target Model	Selection Method	Data Percentage	MMBench-CN Clean	MMBench-CN PA	MMBench-CN SA	MMBench-CN SA + PA	MMBench-CN Avg.	A-OKVQA Clean	A-OKVQA PA	A-OKVQA SA	A-OKVQA SA + PA	A-OKVQA Avg.	MMMU Clean	MMMU PA	MMMU SA	MMMU SA + PA	MMMU Avg.
-	LLaVA-Mistral (7B)	**Full**	100%	71.63	52.34	66.99	38.51	57.36	80.00	68.38	77.99	59.21	71.39	38.84	12.51	35.54	6.49	23.34
-	LLaVA-Mistral (7B)	**Random**	30%	68.33	49.04	57.57	13.17	47.02	77.47	61.31	72.93	39.21	62.73	37.43	12.51	35.30	3.07	22.07
LLaVA-1.5 (7B)	LLaVA-Mistral (7B)	**ARDS**	30%	72.26	57.84	70.32	51.24	**62.92**	81.66	72.58	80.52	69.00	**75.94**	39.55	16.06	36.60	11.33	**25.89**
-	Qwen2.5-VL (7B)	-	-	63.62	36.59	73.60	36.71	52.63	82.18	67.34	75.90	41.48	66.72	52.66	26.21	45.45	11.92	34.06
-	Qwen2.5-VL (7B)	**Random**	30%	68.05	41.79	73.46	32.92	54.05	84.54	73.01	75.90	38.25	67.92	53.72	26.56	46.40	10.74	34.35
LLaVA-1.5 (7B)	Qwen2.5-VL (7B)	**ARDS**	30%	79.05	63.99	75.88	38.58	**64.38**	85.85	77.03	77.55	42.01	**70.61**	53.13	26.92	45.93	11.57	**34.39**

- The robust data mixture curated with Vicuna-based LLaVA-1.5 (7B) transfers effectively to other architectures across visual instruction tuning and post-training settings.

Take Away

1. This paper introduces *robustness* as a new and important data selection objective for visual instruction tuning.

Method	Information Proxy	Objective	Task-Aware Selection	Downstream-Data-free
LESS [107]	Gradient	Quality	✓	✗
ICONS [106]	Gradient	Quality	✓	✗
TIVE [68]	Gradient	Diversity	✓	✓
COINCIDE [51]	Feature	Diversity	✗	✓
ARDS (Ours)	Feature	Robustness	✓	✓

2. Our proposed ARDS is a simple yet effective gradient-free and robustness-aware data selection approach, curating a robust training mixture to enhance model robustness against underlying dataset biases.

Paper

Code

Thanks!