From Head to Tail: Towards Balanced Representation in Large Vision-Language Models through Adaptive Data Calibration

Poster: Jun 13th, 16:00-18:00 ExHall D Poster #387

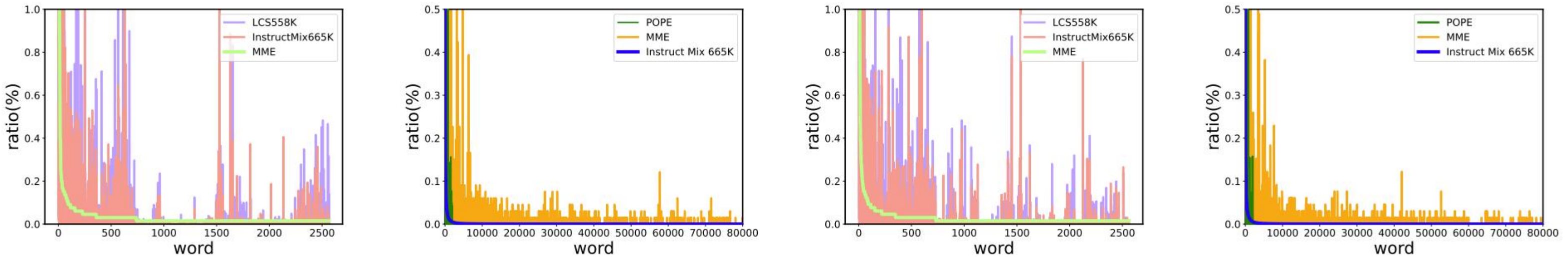
By: Mingyang Song

Outline

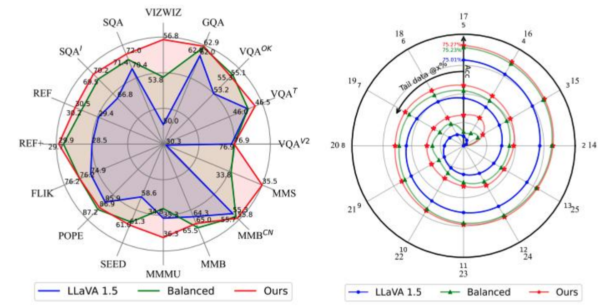
- **Motivation**
- **Pipeline**
- **Methods**
- **Experiments**

Motivation

1. **Current LVLM training data suffers from the “long-tail” problem (LT):**
The training datasets present highly imbalanced distributions, with many tail concepts, and often differ in distribution from the test set.



(a) MME-Tok (b) LLaVA-Tok (c) MME-Obj (d) LLaVA-Obj



(a) Performance Improvement Over LLaVA 1.5. (b) Tail 30% Acc. on VQAV2.

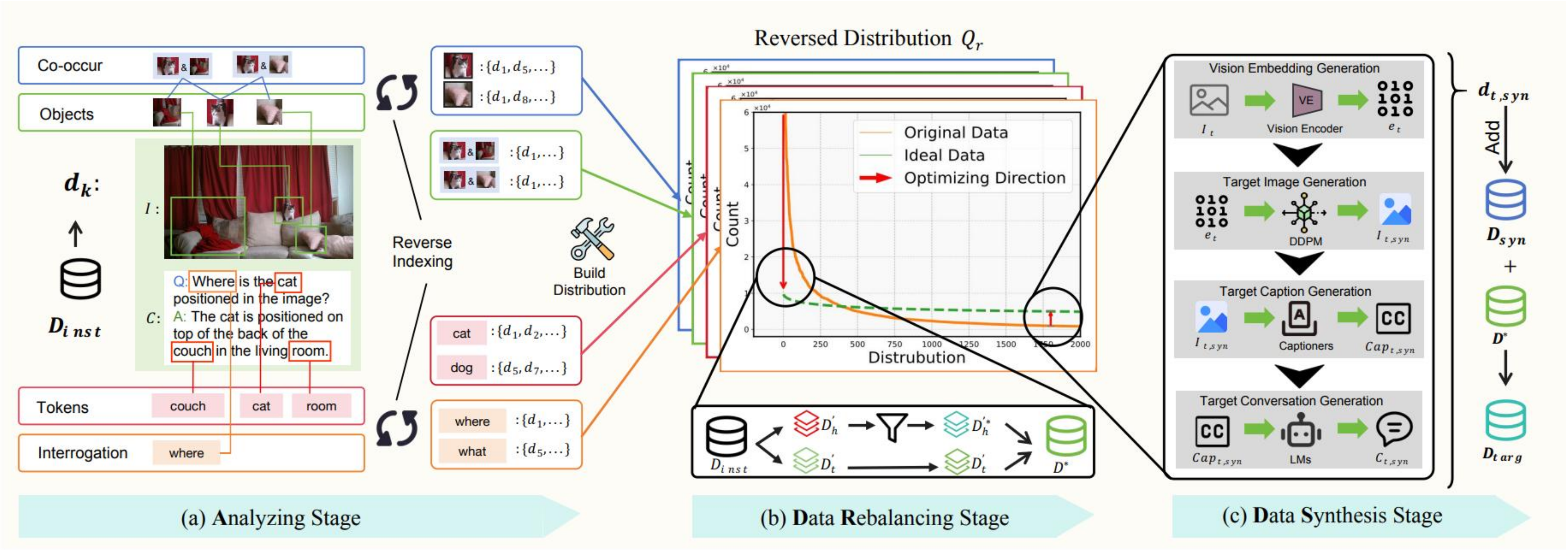
2. **The model is more prone to making mistakes on tail Concepts:** We analyze the distribution of entities in failed cases by computing their positions, and compare them with those in correct cases. Failed cases tend to appear later in the distribution.

Methods	MME						POPE					
	Tok-C	Tok-W	Obj-C	Obj-W	Co-C	Co-W	Tok-C	Tok-W	Obj-C	Obj-W	Co-C	Co-W
Max	9738	10377	2708	3222	247315	257107	2242	2772	1085	1100	130043	141722
		+639		+514		+9792		+30		+15		+11679
Min	1	1	60	131	12732	20741	1	1	17	21	926	1033
		+0		+71		+8009		+0		+4		+107
Mean	1035	1068	842	1035	71123	79104	313	340	319	336	27457	30989
		+33		+193		+7981		+27		+17		+3532

3. LT in LVLMs poses unique, under-explored challenges due to cross-modal complexity, interactions, and distinct co-occurrence patterns.

[1]Song, Mingyang, et al. "From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration." CVPR 2025.

Pipeline



An overview of our Adaptive Data Refinement Framework (ADR)

a. **Analyzing Stage:** we first extract tokens, objects, co-occurrences, and interrogations from the training instances, then construct corresponding distribution using a reverse-indexed mapping.

b. **Data Rebalancing stage:** we analyze the optimizing direction and adaptively rebalance the redundant data based on the entitydistribution identified in the Analyzing stage.

c. **Data Synthesis stage**, we utilize DDPM and the latent representations of scarce image instances to synthesize the underrepresented data.

[1]Song, Mingyang, et al. "From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration." CVPR 2025.

Method

1. Analyzing Stage

1) Entity Distribution Construction

Specifically, we conduct the whole analysis procedure by constructing the frequency distribution of entities Q_e from these four perspectives among the whole training set.

Token: $e_t = \{n | n \in \text{Noun} \wedge n \subseteq C \text{ for } (I, C) \text{ in } D\}$

Object: $e_o = \{o | o \in I \text{ for } (I, C) \text{ in } D\}$

Co-occurrence: $e_c = \{(o_1, o_2) || o_1 \in I \wedge o_2 \in I \text{ for } (I, C) \text{ in } D\}$

Interrogation: $e_w = \{q | q \in Q \wedge q \in C \text{ for } (I, C) \text{ in } D\}$

Data	Level	thres	% E	% DI
LLaVA [27]	Tok	120	98.7	10.0
	Obj	304	98.0	10.0
	Co	24	92.7	25.0
	Int	4895	99.6	10.0
Avg.	-	-	**97.25**	**13.75**

2) Reverse Indexing

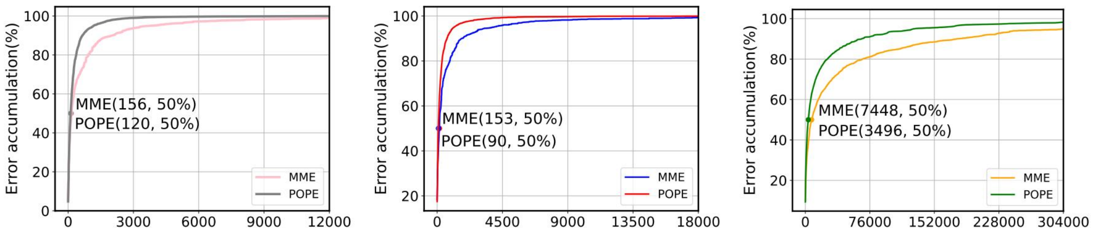
Subsequently, we use the number of data instances corresponding to each entity as frequency to build the reversed distribution Q_r: $Q_r = \{e_1 : N_{e_1}, e_2 : N_{e_2}, \ldots, e_n : N_{e_n}\}$ where e_i means entity item and N_{e_i} means the number of corresponding data instances of e_i.

Surprisingly, among four perspectives, an average of **97.25%** entries account for only **13.75%** data instances on average,

3) Discovery

a. Tail data accounts for more failed cases.

b. Distribution varies between train and test data.



(a) Token (b) Object (c) Co-occurrence

[1]Song, Mingyang, et al. "From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration." CVPR 2025.

https://vlmlt.github.io/

Method

2. Data Rebalancing Stage

We construct entity-wise sampling probabilities to downsample redundant data and select a core dataset based on entity frequency and coverage.

3. Data Synthesis Stage

1) Language Data Synthesis

We rewrite head-biased data by replacing dominant concepts with tail synonyms via LLM-guided paraphrasing.

2) Visual Data Synthesis

a. given a tail instance $d_t = (I_t, C_t)$our objective is to generate an image similar to I and produce corresponding instruction data

b. We leverage ControlNet to generate images that retain the original content while exhibiting a similar visual style$I_{t,syn} = G(I_t, p_{def})$.

c. We use an off-the-shelf vision captioner to generate captions for the synthetic image, which are then expanded into full conversations using LLMs for visual instruction tuning.

Algorithm 1 Pseudo Code for **D**ata **R**esampling

```
# D: raw training set;
# C: target perspectives list
# tau: the threshold for entities;
# D_bal: the rebalanced data, a.k.a. D*;
# n_p, alpha: hyperparameters
D_bal=[]
for pers in C:      # build prob dict
    entity_dist =
        entity_distribution_construction
        (D,pers)
    prob_dict[pers] = {ent:tau[pers]/
        entry_dist[ent] for ent in
        entry_dict.keys()}
for instance in D: #  data rebalancing
    pass_cnt = 0
    for pers in C:
        for entity in instance['entity'
            ][pers]:
            if random.random() <
                prob_dict[pers][entity]:
                pass_cnt += 1
                break
    if pass_cnt > n_p and random.random
        () < alpha:
        D_bal.append(instance)
```

[1]Song, Mingyang, et al. "From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration." CVPR 2025.

Experiments

Quantitive Analysis of Different Methods on Popluar Benchmarks

Method	IT*	VQA^{OK}	$SEED^{2}$	QB^{2}	MMS	MME^{P}	SQA^{I}	MMMU	VQA^{T}	GQA	MMB	VQA^{v2}
LLaVA 1.5	665.0K	53.2	48.7	47.3	33.5	1510.7	69.3	35.3	46.0	61.9	64.3	76.6
+DR	581.0K	55.3	57.2	46.8	33.8	1470.6	69.5	34.8	46.0	62.8	**65.5**	76.9
+DR +DS	665.0K	**57.4**	**57.4**	**49.6**	**35.5**	**1512.8**	**70.4**	**36.7**	**47.2**	**62.9**	65.0	**76.9**
ShareGPT4V	1246.0K	54.0	59.6	44.2	34.7	1560.4	68.9	35.1	50.2	63.3	68.0	78.6
+DR	1168.0K	56.7	59.6	44.9	35.0	1542.3	68.6	35.7	**50.9**	**63.9**	67.9	78.7
+DR +DS	1246.0K	**57.9**	**59.9**	**45.7**	**35.5**	**1564.9**	**69.4**	**36.1**	50.9	63.7	**68.8**	**78.7**

Comparison with models trained with different methods on different benchmarks.

Quantitive Analysis among Different Data Rebalancing Methods

Method	IT	GQA	SEED	$SEED^{v2}$	POPE	MMB
Baseline	665K	62.0	61.0	57.2	87.2	65.5
EL2N	581K	62.5	53.6	47.4	87.2	65.2
Perplexity	581K	62.3	53.4	47.4	86.8	63.7
CLIP Score	581K	62.5	53.0	47.0	87.0	64.5
COINCIDE	133K	59.8	-	-	86.1	63.1
Ours-DR	581K	62.8	61.0	57.2	87.2	**65.5**
Ours	665K	**62.9**	**61.3**	**57.4**	**87.4**	65.0

Popular Data Rebalancing Methods.

The Impact of ADR on Tail concepts

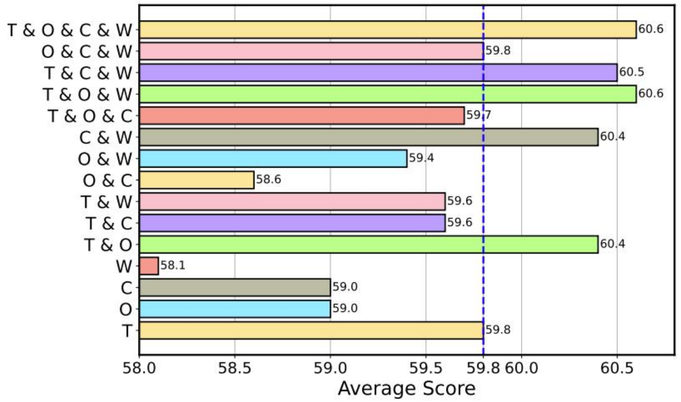
Methods	IT	ScienceQA					
		@5	@10	@15	@20	H@80	Overall
LLaVA 1.5	665.0K	67.9	70.0	67.9	68.5	74.6	69.3
+DR	581.0K	69.2	69.7	67.8	68.5	76.2	69.5
+DR +DS	665.0K	**70.1**	**70.5**	**68.3**	**69.0**	**78.6**	**70.2**

Tail concept prediction accuracy on ScienceQA-IMG

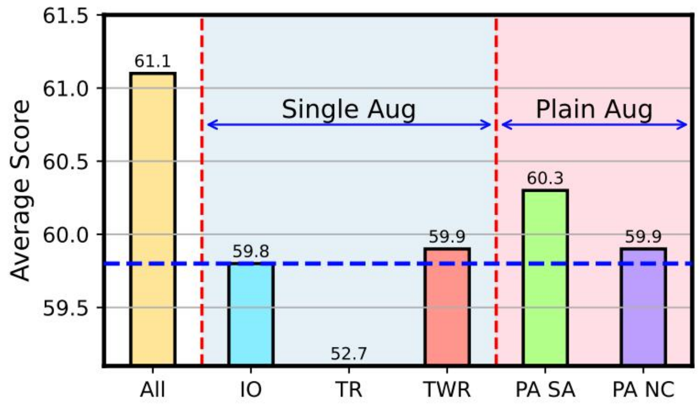
[1]Song, Mingyang, et al. "From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration." CVPR 2025.

Experiments

Ablation Study

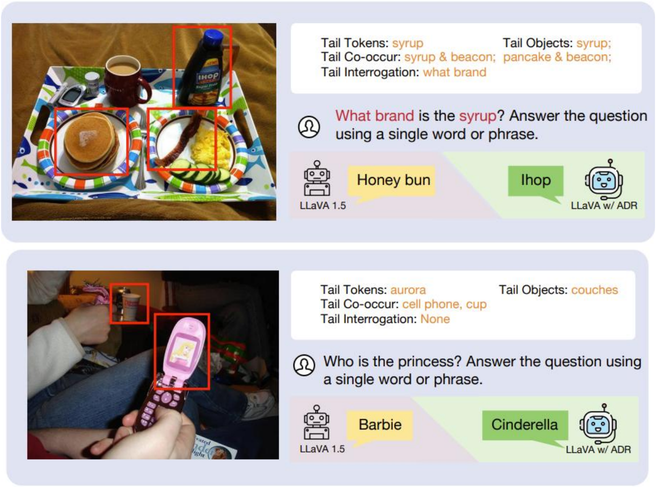


Ablation study on data rebalancing combinations.

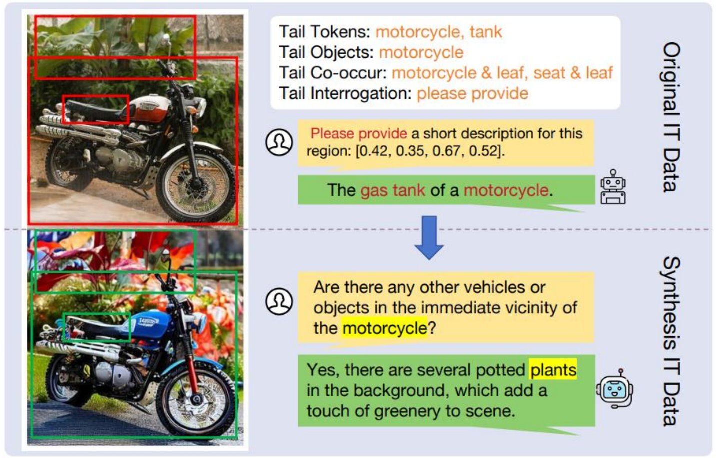


Ablation study on data synthesis methods.

Visualization



Qualitative comparison between LLaVA 1.5 and LLaVA 1.5 w/ ADR



Comparison between the original instruction-tuning (IT) data and our synthesized IT data.

[1]Song, Mingyang, et al. "From head to tail: Towards balanced representation in large vision-language models through adaptive data calibration." CVPR 2025.

Thanks!

Poster: Jun 13th, 16:00-18:00 ExHall D Poster #387