

Supplementary Materials for Enhancing Numerical Prediction of MLLMs with Soft Labeling

Pei Wang, Zhaowei Cai, Hao Yang, Davide Modolo, Ashwin Swaminathan

Amazon AGI

{pwwng, zhaoweic, haoyng, dmodolo, swashwin}@amazon.com

In this supplement, we show some other additional experimental results and details that are not present in the main paper due to the page limitation.

A. Experiment

A.1. The Impact of Soft Labeling on Generic Tasks

In this section, we demonstrate that soft labeling does not degrade performance on tasks that do not rely on numerical tokens. Table A presents results on representative generic benchmarks for both weak and strong baselines, consistent with those used in Section 4.1. In this section, the model is always fine-tuned on LLaVA-Mix. As shown in Table A, incorporating soft labeling does not negatively affect performance across tasks. This is expected, as equation (5), when no numerical tokens are involved in training, the loss naturally reduces to standard one-hot cross-entropy.

A.2. Baseline Results of Soft Labeling

In this section, we present the baseline results of our method in comparison to state-of-the-art approaches in Section 4.3. These results are obtained using the same datasets and settings but replacing soft labels with one-hot hard labels. As shown in Tables B, C, and D, soft labeling consistently yields stable and significant improvements, even over strong baselines.

A.3. Visualization

More visualizations are provided in Tables E, F and G to highlight the effectiveness of soft labeling in improving numerical prediction.

B. Supplementary Experimental Settings

B.1. Pretraining Data

In this section, we describe the pretraining data used in the second stage.

Visual grounding For visual grounding, we sampled a 5M subset from the public GRIT-20M [17], referred to as

GRIT-5M. This data is used for an interleaved captioning task, following the example template below.

```
{
  'human': '<image>\n Please provide a
            description for this image along
            with the coordinates for every
            object.',
  'gpt': 'A man <bbox1> is sitting in a
          bench <bbox2>.'
```

where $\langle \text{bbox1} \rangle$ and $\langle \text{bbox2} \rangle$ represent bounding box coordinates $[x_{min}, y_{min}, x_{max}, y_{max}]$ normalized to an integer in the range $[0, 1000]$.

We use OFA [20] to generate pseudo captions for ground truth bounding box annotations from the Objects365 [18], LVIS [5], and COCO2017 [11] training sets. During this process, we exclude examples from the COCO2017 training set that belong to the RefCOCO, RefCOCO+, and RefCOCOG validation sets. A similar exclusion is applied to the LVIS training set. After generating pseudo captions for each bounding box, the resulting 2M-example dataset is used for visual grounding pretraining. We refer to this dataset as “OGC-2M” (OFA-generated Grounding Captions, 2M examples).

Additionally, based on another randomly sampled 5M subset of LAION images, we use ScaleDet [2] to generate pseudo bounding boxes and associated descriptions for each image, creating another synthetic visual grounding dataset, referred to as “LAION-VG-5M” (LAION-based Visual Grounding, 5M examples).

A sample template used for visual grounding is shown below:

```
{
  'human': '<image>\n Please provide the
            bounding box coordinate of the
            region this sentence describes: <
            region caption>.',
  'gpt': '<bbox>.'
```

Table A. The impact of soft labeling on generic tasks

Pretrain model	Loss	Finetune data	VQAv2 [4]	VisWiz [6]	ScienceQA [13]	TextVQA [19]	POPE [10]			MME [9]
							rand	pop	adv	
LLaVA-7B	Hard	LLaVA-Mix	78.6	49.7	66.8	62.1	87.3	86.1	84.2	1510.7
	Soft	LLaVA-Mix	78.8	50.6	69.9	65.4	87.1	86.2	84.1	1517.3
LLaVA-13B	Hard	LLaVA-Mix	80.0	53.6	71.6	61.3	87.1	86.2	84.5	1531.3
	Soft	LLaVA-Mix	79.7	53.2	72.9	67.2	87.5	86.4	85.0	1538.1

Table B. The improvement over visual grounding baseline in Table 5 of the paper

Models	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
Hard labeling -7B	91.0	93.9	87.0	86.1	91.5	79.5	89.0	89.4
Soft labeling -7B	91.8(+0.8)	94.7(+0.8)	88.9(+1.9)	87.0(+0.9)	92.7(+1.2)	80.0(+0.5)	89.6(+0.6)	89.5(+0.1)
Hard labeling -13B	92.1	94.6	88.7	86.9	92.5	82.1	89.7	89.6
Soft labeling -13B	92.7(+0.6)	95.0(+0.4)	89.0(+0.3)	87.6(+0.7)	92.7(+0.2)	82.3(+0.2)	89.8(+0.1)	90.0(+0.4)

Table C. The improvement over chart understanding baseline in Table 6 of the paper

Models	ChartQA
Hard labeling	80.2
Soft labeling	81.5 (+1.3)

Table D. The improvement over object counting baseline in Table 7 of the paper

Models	TallyQA simple		TallyQA complex	
	Acc.(↑)	RMSE(↓)	Acc.(↑)	RMSE(↓)
Hard labeling	86.2	0.73	76.4	1.33
Soft labeling	86.6(+0.4)	0.56(-0.17)	77.2(+0.8)	1.06(-0.27)

where $\langle \text{region caption} \rangle$ represents the generated pseudo caption.

Chart understanding For chart understanding, we use a 3M mixture of public datasets, including PlotQA [16], PMC2022 [3], UniChart [15], FigureQA [7], and ChartOCR [14], for pretraining. We refer to this dataset as Chart-public-3M.

Additionally, we crawl 17M tables from the web and convert each table into various charts or plots. We then prompt DeepSeek-V2 [12] to generate QA questions for pretraining. We refer to this dataset as Chart-synthesis-17M.

Object counting For object counting, we reformat the public object detection datasets, Objects365 [18] and OpenImages [8], for counting tasks. These datasets provide bounding box and class annotations for each object, making it straightforward to extract the count information for each class. We then create a naive counting dataset with 3M examples (Counting-3M) and train the model using the following template:

```
{
  'human': '<image>\n How many <class> in
           this image?',
  'gpt': '<num>.'
}
```

In addition to this naive counting approach, we also train the model using a Chain-of-Thought (CoT) style prompt, as exemplified below:

```
{
  'human': '<image>\n Please count the
           number of <class>. Also, please
           provide the bounding box coordinate
           for each object as evidence before
           giving the final count.',
  'gpt': '<bbox1>, <bbox2>. So, the total
           number is <num>.'
}
```

where $\langle \text{class} \rangle$ is the object category and $\langle \text{num} \rangle$ is the ground truth count. This forms another pretraining dataset, Counting-CoT-3M.

Finally, for the 16M mixture data used to train the strong LLaVA-13B baseline in Section 4.1, the dataset includes LAION-5M, GRIT-5M, Chart-public-3M, and Counting-3M. For the pretraining model used in Section 4.3, we utilize all the collected data for each task. Specifically, for visual grounding, we pretrain on a 12M mixture of GRIT-5M, OGC-2M, and LAION-VG-5M. For chart understanding, we pretrain on a 20M mixture of Chart-public-3M and Chart-synthesis-17M. For object counting, we pretrain on a 6M mixture of Counting-3M and Counting-CoT-3M.

B.2. Finetuning Data

For RefCOCO, RefCOCO+, and RefCOCOg training data, we train the model with grounding and expression referring

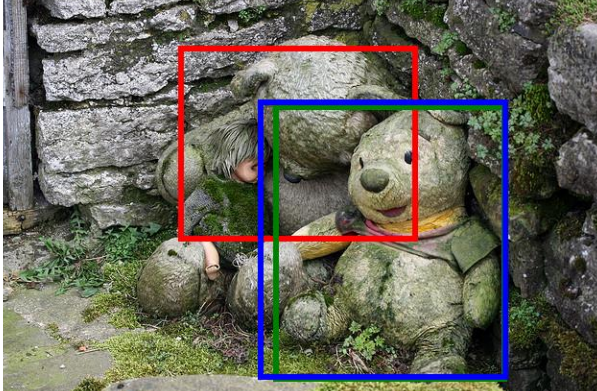
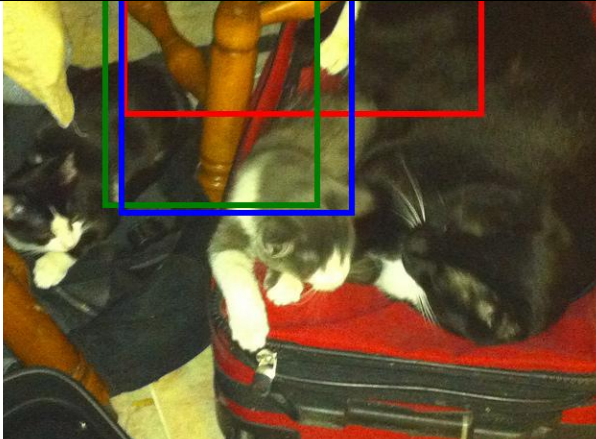
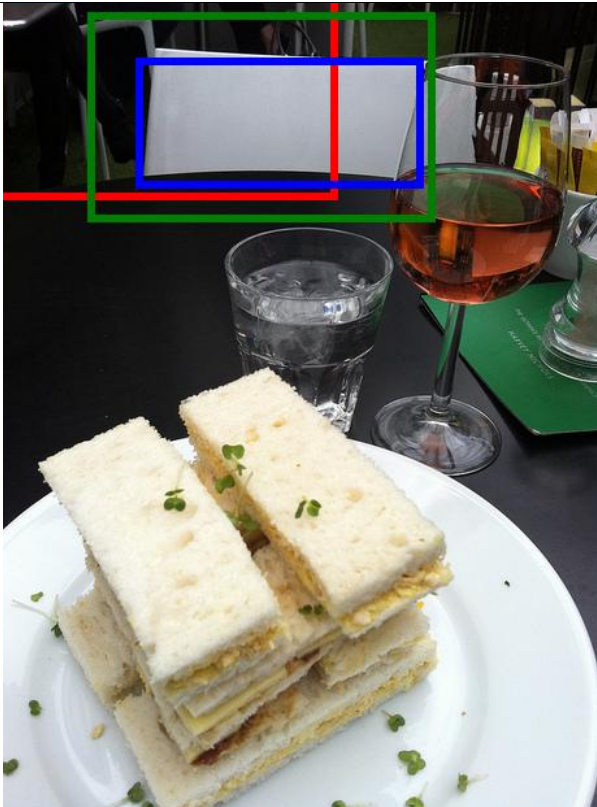

Image		
Prompt	Rocks that look like Winnie-the-poo, facing the camera	wooden stool leg between cats and suitcases
Hard labeling	red box	red box
Soft labeling	green box	green box
Ground truth	blue box	blue box
Image		
Prompt	white chair at table	the bottom end of a book with a leather book cover
Hard labeling	red box	red box
Soft labeling	green box	green box
Ground truth	blue box	blue box

Table E. Qualitative examples of soft labeling improvements in visual grounding

generation tasks. For ChartQA and TallyQA, we use the official training sets.

B.3. Balanced TallyQA Testing Sets

TallyQA [1] is a commonly used benchmark dataset for evaluating object counting. However, we identified at least two significant issues with it. First, the ground-truth label distribution is highly imbalanced, with counts of 1 and 2

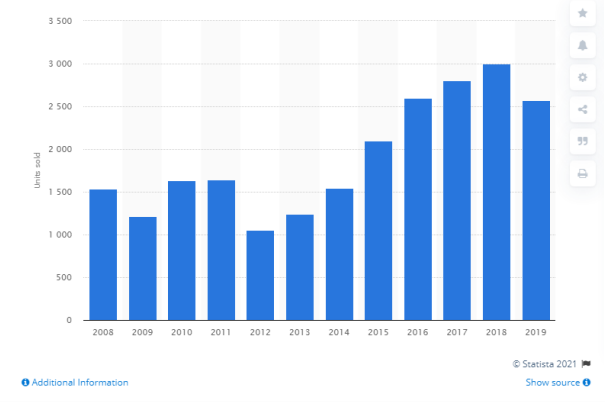
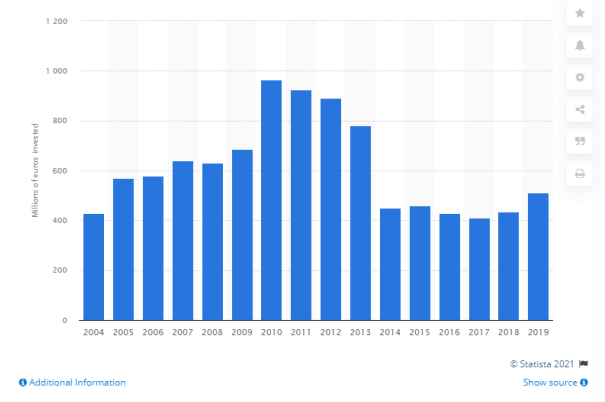
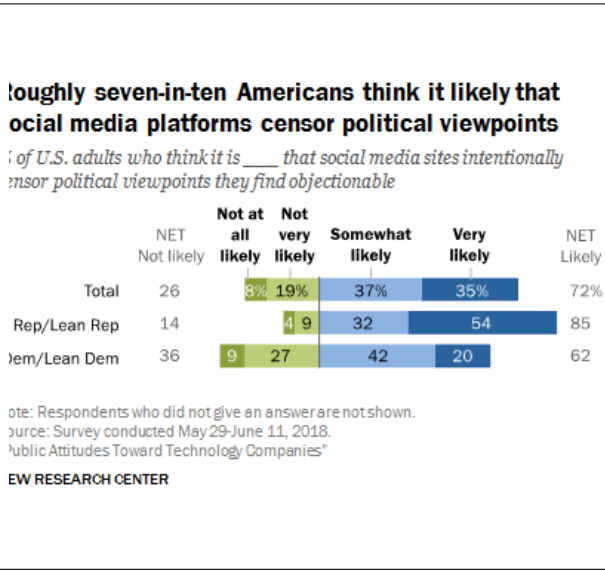
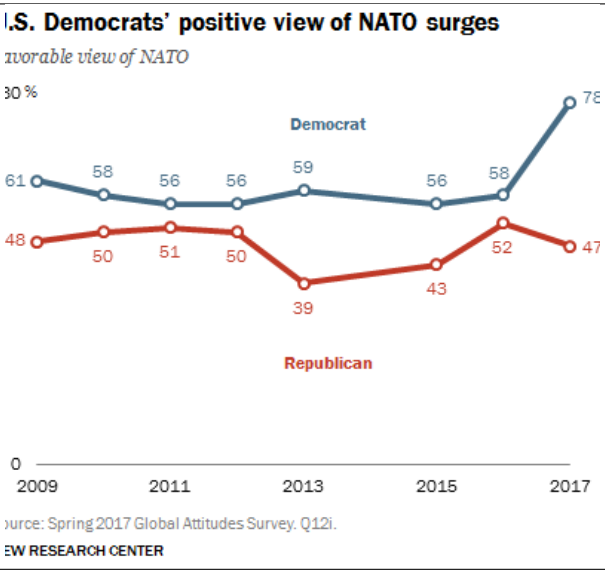
Image		
Prompt	How many cars did Mini sell in Portugal as of 2019?	What was the total amount of investments in sea port infrastructure in 2010?
Hard labeling	2301	800
Soft labeling	2701	930.5
Ground truth	2601	965
Image		
Prompt	How many colors are used to represent the bar graph?	How many points have 56 value in blue graph?
Hard labeling	3	2
Soft labeling	4	3
Ground truth	4	3

Table F. Qualitative examples of soft labeling improvements in chart understanding

dominating the testing samples (see Figure A). Second, the ground-truth labels are noisy, ambiguous, and unverified, as illustrated by the examples provided in Figure B. As a result, directly evaluating on TallyQA can lead to misleading conclusions.

In order to evaluate and develop models on a trustworthy validation set, we sampled two balanced subsets from TallyQA’s simple and complex testing sets respectively, referred to as “TallyQA simple balanced” and “TallyQA complex balanced”. TallyQA simple balanced contains 109 images with 7 examples per count from 0 to 15 (with 3 counts having 6 examples). TallyQA complex balanced contains 96 images with 6 examples per count from 0 to 15. Each

image and annotation was manually checked to ensure accuracy and clarity.

C. Potential Negative Societal Impacts

While our work focuses on improving numerical prediction in MLLMs through soft labeling, its broader implications warrant consideration. Enhanced numerical reasoning could be leveraged in applications that inadvertently contribute to misinformation, such as AI-generated content misrepresenting statistical data or object counts in critical domains like journalism or forensic analysis. To mitigate these risks, we suggest to implement robust verifi-





Image		
Prompt	How many cars did Mini sell in Portugal as of 2019?	What was the total amount of investments in sea port infrastructure in 2010?
Hard labeling	7	0
Soft labeling	11	5
Ground truth	13	5
Image		
Prompt	How many surf boards are there?	How many donuts are visible?
Hard labeling	7	7
Soft labeling	5	14
Ground truth	5	15

Table G. Qualitative examples of soft labeling improvements in object counting

cation mechanisms that cross-check numerical predictions with trusted external sources before they are used in applications like journalism or forensic analysis. This could involve integrating fact-checking systems and leveraging

domain-specific knowledge to ensure predictions align with established data.

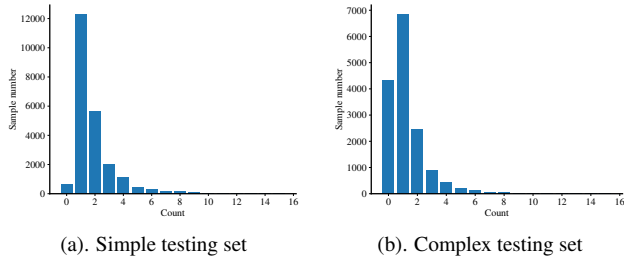


Figure A. Count distribution on TallyQA simple and complex testing sets

References

- [1] Manoj Acharya, Kushal Kaffle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *AAAI*, pages 8076–8084, 2019. [3](#)
- [2] Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu, Paolo Favaro, Joseph Tighe, and Davide Modolo. Scaledet: A scalable multi-dataset object detector. In *CVPR*, pages 7288–7297, 2023. [1](#)
- [3] Kenny Davila, Fei Xu, Saleem Ahmed, David A Mendoza, Srirangaraj Setlur, and Venu Govindaraju. Icp2022: Competition on harvesting raw tables from infographics (chart-infographics). In *Pie*, page 242. [2](#)
- [4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. [2](#)
- [5] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, pages 5356–5364, 2019. [1](#)
- [6] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. [2](#)
- [7] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. [2](#)
- [8] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 128(7):1956–1981, 2020. [2](#)
- [9] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *CVPR*, pages 13299–13308, 2024. [2](#)
- [10] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. [2](#)
- [11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. [1](#)
- [12] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. [2](#)
- [13] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35: 2507–2521, 2022. [2](#)
- [14] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In *WACV*, pages 1917–1925, 2021. [2](#)
- [15] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. [2](#)
- [16] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020. [2](#)
- [17] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. [1](#)
- [18] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*, pages 8430–8439, 2019. [1](#), [2](#)
- [19] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. [2](#)
- [20] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR, 2022. [1](#)



Question: How many orange leaves are there?
Answer: 14



Question: How many water drops are in the picture?
Answer: 14



Question: How many roads are there?
Answer: 6

Figure B. Representative examples of TallyQA's testing set with confusing or noisy annotations.