

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

Abstract

The optimality of using the *de facto* cross-entropy loss with one-hot target distribution (hard labeling) is questioned when training (Multimodal) Large Language Models (LLMs/MLLMs). Although it is reasonable for language token prediction, which is a typical multi-class classification problem in discrete space, it is suboptimal for task like numerical prediction, which is a typical regression problem in continuous space. However, enabling regression in LLMs/MLLMs will complicate the training and next-token prediction paradigm at inference. Instead, to address this challenge, we propose a novel loss design, called soft labeling, which smooths the target probability distribution, enabling predictions to be penalized according to their distance to the target. This is similar to regression loss, which penalizes more on the further predictions in the continuous space, but will not change the model architecture and the next-token prediction paradigm of LLMs/MLLMs. We demonstrate the efficacy of soft labeling through extensive experiments on visual grounding, object counting, and chart understanding, achieving state-of-the-art performance on multiple benchmarks without bells and whistles. Soft labeling can be applied in any LLM/MLLM.

1. Introduction

In recent years, thanks to the success of Large Language Models (LLMs) [7, 52, 63], Multimodal Large Language Models (MLLMs) [2, 4, 31, 60] have achieved remarkable progresses across various tasks, such as image captioning, visual question answering, and many specialized domains like document analysis, chart interpretation, and visual grounding, etc. Their learning is to predict the next token, associated with the *de facto* cross-entropy loss across the token vocabulary, no matter the target is a word, digit, punctuation, or anything else. Although this unified mechanism simplifies the learning across various tasks and it becomes dominant, a natural question to ask is: is it optimal for any task? For example, the multi-class classification paradigm is reasonable for language word token predic-

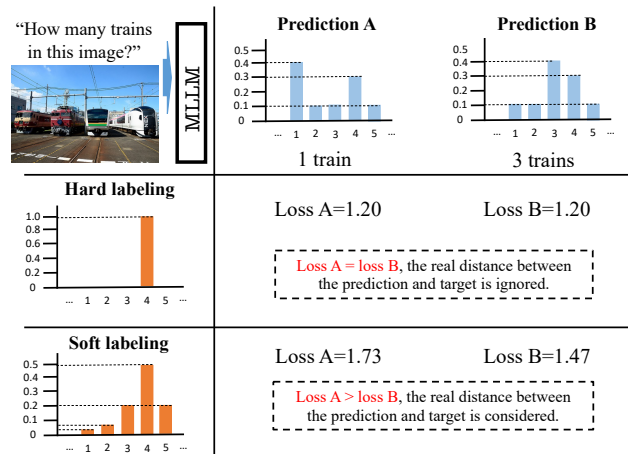


Figure 1. Given an image with 4 trains, the MLLMs generate probabilistic predictions (A and B) for the count. In standard one-hot *hard labeling*, only the ground-truth numerical token has probability of 1, leading to equal loss values (1.20) for those two incorrect predictions (1 and 3 trains). In contrast, *soft labeling* assigns probabilities to numerical tokens according to their distances to the target. This results in different loss values (1.73 for prediction A of 1 train but 1.47 for prediction B of 3 trains), enabling higher penalty for further prediction.

tion because language word is in discrete space, but it is no longer optimal for objects that are in continuous space. In this paper, we are investigating the limits of standard cross-entropy loss in an important case, *i.e.*, numerical prediction, which is in continuous space.

Numerical prediction is a critical capability for LLMs and MLLMs, with many applications in object counting, object localization, chart/figure understanding, etc. For instance, answering the gross income from a tax report, or answering how many trains in a traffic scene as in Figure 1. Although the gross income value (*e.g.*, 35421.7) and the number of trains (*e.g.*, 4) are in continuous space, they are converted to strings (*e.g.*, “35421.7” and “4”) and then to discrete tokens to be processed by LLMs/MLLMs as a multi-class classification problem with cross-entropy loss. Cross-entropy loss measures the distance between the target and prediction distributions over the token vocabulary, and the target distribution is usually a one-hot vector where

the probability on the target is 1 and the rest are 0. In this case, the wrong prediction to any non-target index can potentially have no difference. For example, in Figure 1, where there are 4 trains in a given image, predictions A and B have exactly the same cross-entropy loss because their probabilities on the target (*i.e.*, 4) are the same (*i.e.*, 0.3), although their overall distributions are quite different from each other. However, this is counter-intuitive. Prediction B of 3 trains should have lower loss than prediction A of 1 train, because the former is closer to the target. The cause of this issue is rooted in the one-hot representation of the target probability distribution, namely *hard labeling*, of the standard cross-entropy loss.

To address this problem, a new loss design is required, which can better reflect how wrong the prediction is for objects in continuous space. An ideal choice would be regression-style loss, *e.g.*, L1 or L2 loss. However, asking LLMs/MLLMs to output scalar values is nontrivial without extra layers or heuristics. In addition, this will break the loss/architecture unification of LLMs/MLLMs and make the next-token prediction paradigm very complicated. Instead, we propose *soft labeling* for LLM/MLLM training, in contrast with *hard labeling* of the standard cross-entropy loss. In *soft labeling*, the target probability is no longer a one-hot probability distribution (Figure 1 middle), but smoothed where closer (further) classes have higher (lower) probabilities (Figure 1 bottom). Different from *hard labeling*, where only the target probability contributes to the loss, *soft labeling* enables losses coming from nearby classes, and considers the relative distance between the predicted and target tokens, similar to the regression loss. For example, in Figure 1, prediction A and B have the same loss using *hard labeling*, but the loss of B is lower than that of A for *soft labeling*. This is reasonable because prediction B of 3 trains is closer to the target of 4. In addition, the design of *soft labeling* will not modify the model architecture and next-token generation paradigm, which can be used in any LLM/MLLM.

In our experiments, we demonstrate the effectiveness of the proposed *soft labeling* over *hard labeling* across three representative tasks of numerical prediction: visual grounding, object counting, and chart understanding on multiple benchmarks. *Soft labeling* consistently improves performance on all baselines. We also conduct comprehensive ablation studies on each component of *soft labeling* to gain deeper understanding of their impacts. By applying *soft labeling*, we achieve state-of-the-art or competitive performance on the RefCOCO/+g [25], ChartQA [40], and TallyQA [1] datasets, using only a standard LLaVA model [32] without any additional architectural modifications. This highlights the effectiveness of *soft labeling* in enhancing numerical prediction in MLLMs.

Overall, this work makes three major contributions.

First, we question the optimality of using the *de facto* cross-entropy loss with one-hot target encoding (*hard labeling*) in any task of LLMs/MLLMs, and explain why standard cross-entropy loss leads to counter-intuitive loss assignment for a critical task, *i.e.*, numerical prediction, in continuous space by nature. Second, we introduce *soft labeling*, where target probability distributions are smoothed such that closer (further) tokens to the target receive higher (lower) probabilities. This design better reflects the degree of error in predictions while maintaining compatibility with the standard next-token prediction paradigm of LLMs/MLLMs. Third, extensive experiments on visual grounding, object counting, and chart understanding demonstrate that *soft labeling* consistently outperforms *hard labeling*. The proposed approach achieves state-of-the-art or competitive results on benchmarks like RefCOCO, ChartQA, and TallyQA, without any architectural modification.

2. Related Work

Numerical prediction is a core challenge across a wide range of tasks in MLLMs and LLMs, spanning visual grounding [12, 48, 70], object counting [1], chart understanding [40, 43], visual question answering [18, 56], mathematical reasoning [38, 67], temporal reasoning [14, 16, 51] and document analysis [42, 59]. In visual grounding, models must quantify spatial location of objects with numerical bounding box coordinates [12, 48, 70]. In object counting, the model is required to predict precise numerical numbers of instances [1, 53]. Chart understanding relies on extracting and reasoning over numerical data in visual form [40, 43]. General VQA often involves answering questions that require numerical reasoning, such as counting, comparing quantities, or understanding scales [18, 56]. Mathematical reasoning, including arithmetic and algebraic problem-solving, tests models’ ability to manipulate numerical values symbolically [38, 67]. Temporal reasoning in multimodal setups involves understanding and predicting time intervals or event sequences from textual and visual data [14, 16, 51]. Document analysis further highlights numerical prediction challenges, where models must extract and interpret numerical information from structured and unstructured data and make reasoning [42, 59]. These diverse applications underscore the pervasive role of numerical prediction in MLLMs and LLMs, making it a critical area for improvement.

Ordinal regression (also called ordinal classification) is the problems where the target variable exhibits a relative ordering on an arbitrary scale, *e.g.* categories such as “bad”, “good” and “very good”. It has been proven useful in different research areas, including medical research [5, 35], monocular depth estimation [17], age estimation [46], pose estimation [21], etc. Ordinal regression has been well studied in conventional machine learn-

ing methods [19]. In the scenarios of deep learning, there are some typical methods. Class distance weighted loss relies on a coefficient for the loss of each class, which utilizes the distance to the ground-truth class and increases in relation to that distance [49]. The Earth Mover’s Distance (EMD) loss uses the predicted probabilities of all classes and penalizes the miss-predictions according to a ground distance matrix that quantifies the dissimilarities between classes [9, 20]. Soft labeling converts ground truth data labels into soft probability distributions and the loss is computed on them [6, 15, 36, 65]. However, these methods have only shown success on simple multi category classification problem. For LLMs, the probability space of the output has tens to hundreds of thousands of tokens and they are mixture of numerical and non-numerical. The problem becomes difficulty. In this work, we will show how to apply ordinal regression to LLMs.

Soft labeling uses probability distributions over classes as labels instead of hard, one-hot labels. A typical type is label smoothing which has been found more robust to noisy annotations for classification problem [39, 61, 71] and has been applied to many fields such as know distillation [44, 73], network calibration [30], and some applications like image segmentation [66], text mining [34]. On the other hand, soft loss terms have been useful for domain and task transfer [64] in order to avoid dataset biases. Elaborate loss functions are defined in [68] to take into account the subjective scenicness of outdoor pictures, by trying to predict the same rating distribution of human annotations. Age estimation is another particular niche where soft labels have become popular [15, 58]. However, there is few research of applying soft labeling on numerical prediction of LLMs. In this paper, we will show that the label smoothing is just a specific format of our soft labeling and our proposal performs better.

3. Method

In traditional regression models, numerical prediction is typically handled using L1/L2-style loss functions, which penalize errors based on their distance to the target. Further distance incurs greater penalties and vice versa, encouraging the model to make predictions that are closer to the target. However, when training LLMs/MLLMs, the de facto cross-entropy loss with one-hot encoded target fails to account for the distance between predicted and target tokens *properly*. It does so in a limited and imprecise manner, a misclassification is a misclassification, regardless of how explicitly far the predicted token is from the target token. This limitation hinders the model’s ability to predict accurate numerical values. To address this, a natural idea is to apply a regression-style loss to LLMs. However, in the context of LLMs, it is not trivial to directly convert the LLM’s output into a scalar without introducing additional layers or heuristics. For this reason, we opt for a new loss de-

sign which can be directly applied to the intrinsic output of LLM, similar to the standard cross-entropy loss.

We begin by reviewing the LLM training loss. LLMs typically generate text token by token, treating the next-token prediction as a classification problem over a vocabulary of size V . Given a model output logit $\mathbf{z} \in R^V$ from the last layer before the final prediction, the softmax function is applied to convert these logits into probability distribution $\mathbf{p} \in R^V$ such that $\sum_{i=1}^V p_i = 1$. This distribution is then compared to the target token using cross-entropy loss. Typically, the target token is encoded into a one-hot vector $\mathbf{q} \in R^V$ where it is a binary vector consisting of a value 1 at the correct token index, and 0s everywhere else. The cross-entropy loss is computed by

$$L(\mathbf{p}, \mathbf{q}) = - \sum_{i=1}^V q_i \log p_i. \quad (1)$$

Since \mathbf{q} is a one-hot vector, the final loss simplifies to

$$L(\mathbf{p}, \mathbf{q}) = - \log p_t, \quad (2)$$

where t is the index of the target token. As shown, the loss function depends solely on the posterior probability p_t of the target token. This ignores other rich information inherited in the full probability distribution \mathbf{p} .

In fact, each probability p_i reflects the model confidence on each possible token. When a prediction is incorrect, it is crucial to have a metric that precisely measures how “wrong” the prediction is by explicitly considering the model’s confidences on all tokens, not just the target token. This becomes important when the target token is numerical and the numerical prediction is done by LLMs. Numerical tokens have relative distance between each other by nature, and when a prediction is incorrect, we should measure how wrong it is by assessing the distance from the target, rather than treating all wrong predictions equally as long as the confidence on the target token is the same, as is done in hard labeling. For example, in the scenario illustrated in Figure 1 to answer how many trains given an image, when we train an MLLM with cross-entropy loss and one-hot encoding, there are two possible incorrect predictions, A and B, during MLLM training. If we apply hard labeling to encode the supervision “4”, both predictions incur the same cross-entropy loss, as both have a softmax probability of 0.3 for the target token “4”. This approach fails to account for numerical proximity since prediction B, which assigns a maximum probability of 0.4 to the incorrect token “3”, is closer to the correct answer “4” than prediction A, which assigns a maximum probability of 0.4 to the incorrect token “1”. Intuitively, prediction B should receive a lower loss from the regression perspective. This limitation arises from the one-hot encoding in hard labeling which only retains the probability p_t of the target token in the final loss computation

of (2). As a result, the final loss function does not consider the probabilities p_i for other numerical tokens, which contain valuable information about inter-token relationship and numerical distance. The absence of this information makes the hard labeling loss suboptimal, and we need a new loss function that can better account for such relationships.

In order to address this issue, we propose *soft labeling* cross-entropy loss. We first reformulate the cross-entropy loss of (1) into a more general form,

$$L_{CE}(\mathbf{p}, \mathbf{q}(t)) = \sum_{i=1}^V q_i(t)(-\log p_i), \quad (3)$$

where $\mathbf{q}(t)$ is the distribution in which the target token with index of t is encoded, and $q_i(t) \in [0, 1]$ s.t. $\sum_i q_i(t) = 1$. In the standard one-hot encoding for cross-entropy, $\mathbf{q}(t) = \delta(t)$ is a Dirac delta distribution where $q_i(t)$ equals to 1 for $i = t$, and 0 otherwise. In this case, (3) is simplified to (2), which corresponds to *hard labeling*. In fact, \mathbf{q} is not necessarily a one-hot distribution, but can be any distribution. The primary objective of \mathbf{q} is to guide the learning of the LLM model by minimizing the distance between two distributions. Instead of forcing the model to learn from a strict, hard label with harsh penalization, soft labeling relaxes this constraint by using a smooth distribution \mathbf{q} over all tokens.

In this work, we focus on numerical prediction, where numerical values are typically tokenized digit-wise [26]. That is, each digit (0–9) is treated as an individual token, and multi-digit numbers are split accordingly. For example, in the LLaMA tokenizer, number “123” is tokenized into the token IDs “29896”, “29906”, and “29941”. Therefore, we restrict our *soft labeling* discussion to the 10 digit tokens from “0” to “9”, leaving the soft labeling of regular tokens for future work. For instance, we do not consider the soft labeling for text tokens such as “two”, “eight”, etc. Following [6, 65], we define soft label for the 10 digit tokens from “0” to “9” as

$$\mathbf{q}^{SL}(t) = (1 - \eta)\delta(t) + \eta\psi(t) \quad (4)$$

where t is the index of the 10 digit tokens $\{“0”, “1”, \dots, “9”\}$. For simplicity, we assume that these indices are consecutive and form a set S ; if not, a straightforward mapping can be constructed to make them consecutive. $\mathbf{q}^{SL}(t)$ is a soft distribution to represent the token with index of t . $\psi(t)$ is a probability distribution that has a support set on S and typically centers at the token with index of t . It can be Poisson, Binomial, Triangular, Uniform distributions, etc. $\eta \in [0, 1]$ is a fixed hyperparameter that controls the mixture ratio of $\delta(t)$ and $\psi(t)$.

By smoothing the one-hot $\mathbf{q}(t)$ of hard labeling, soft labeling allows $q_i^{SL}(t)$ to be greater than 0 for $i \neq t$ and to

vary for different i . This enables the assessment of relationship between the target token and other digit tokens. The relationship reflects the similarity between token indexed by i and the token indexed by t , with $q_i^{SL}(t)$ being higher when i is closer to t and lower otherwise.

The final loss for each example is given by

$$L = \frac{1}{N_r + N_n} \left(\sum_{(x_r, y_r) \in (\mathbf{x}_r, \mathbf{y}_r)} L_{CE}(f(x_r), \mathbf{q}(y_r)) + \lambda \sum_{(x_n, y_n) \in (\mathbf{x}_n, \mathbf{y}_n)} L_{CE}(f(x_n), \mathbf{q}^{SL}(y_n)) \right), \quad (5)$$

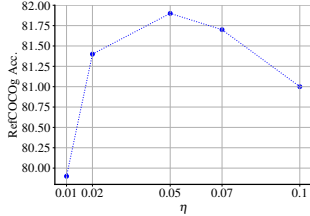
where N_r, N_n denote the numbers of regular tokens and numerical tokens within a sequence example, respectively. \mathbf{x}_r (\mathbf{x}_n), \mathbf{y}_r (\mathbf{y}_n) represent the sets of regular (numerical) tokens where \mathbf{x}_r (\mathbf{x}_n) is the input set and \mathbf{y}_r (\mathbf{y}_n) is the index set of ground truth. $f(\cdot)$ is the LLM with softmax output for each token prediction. λ controls the balance between the numerical token loss and the regular token loss in the final loss computation.

When the target is a digit token, compared to *hard labeling* (one-hot encoding) which solely rely on the probability of target token in the loss function, as indicated in (2), *soft labeling* incorporates the prediction probabilities of all incorrect digit tokens into loss computation. This allows the loss value to reflect the distance between the predicted token and target token. It therefore can be understood as an implicit way of introducing distance-based penalties into the classification loss, similar to how loss functions operate in regression. As illustrated in Figure 1, an incorrect prediction of “1” (Prediction A) has a greater distributional distance from the soft ground truth than an incorrect prediction of “3” (Prediction B). Consequently, Prediction A incurs a larger loss than Prediction B, aligning with the behavior of regression losses. This property arises naturally from the inter-digit token similarity encoded in soft labels where $q_i^{SL}(t)$ reflects the similarity ranking to the ground truth token. The closer a digit token is to the ground truth, the higher its value, enabling the model to account for the severity of prediction errors rather than treating all mistakes equally. In contrast, under *hard labeling*, predictions A and B receive the same loss due to the underlying assumption of one-hot encoding that all incorrect tokens are equally dissimilar to the ground truth.

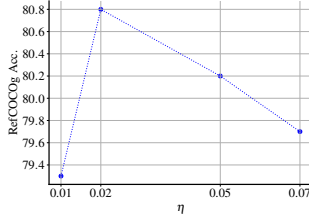
From another perspective, in (3), $q_i(t)$ can be viewed as a weight for each component loss $-\log p_i$. During training, a soft label with $q_i^{SL}(t) > 0$ encourages the model to predict higher probabilities on tokens closer to the target token, and lower probabilities on further tokens. This also aligns with the goal of regression-style loss, which aims to make predictions as close as possible to the target in continuous space.

Table 1. The enhancement by soft labeling on visual grounding.

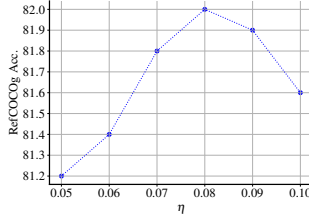
Pretrain model	Model type	Finetune data	Loss	RefCOCO			RefCOCO+			RefCOCOg	
				val	testA	testB	val	testA	testB	val	test
LLaVA-7B	specialist	RefCOCO/+ /g	hard	81.7	87.3	74.4	73.6	82.0	63.3	79.0	78.8
			soft	85.0(+3.3)	89.3(+2.0)	78.8(+4.4)	76.7(+3.1)	83.4(+1.4)	67.0(+3.7)	82.0(+3.0)	81.8(+3.0)
	generalist	LLaVA-Mix + RefCOCO/+ /g	hard	72.6	80.2	62.6	66.6	76.1	55.0	72.1	66.6
			soft	77.8(+5.2)	83.7(+3.5)	68.8(+6.2)	72.0(+5.4)	80.2(+4.1)	60.4(+5.4)	78.5(+6.4)	71.5(+4.9)
LLaVA-13B	specialist	RefCOCO/+ /g	hard	88.9	91.7	84.9	82.5	88.1	75.1	86.0	86.7
			soft	90.0(+1.1)	92.7(+1.0)	85.3(+0.4)	83.5(+1.0)	89.3(+1.2)	76.7(+1.6)	86.7(+0.7)	87.4(+0.7)
	generalist	LLaVA-Mix + RefCOCO/+ /g	hard	79.3	85.0	71.4	74.7	82.9	64.8	78.0	73.0
			soft	81.0(+1.7)	86.4(+1.4)	72.6(+1.2)	76.4(+1.7)	83.3(+0.4)	66.1(+1.3)	79.2(+1.2)	74.3(+1.3)



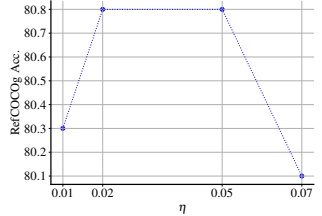
(a) RefCOCOg Acc. v.s. η for Binomial distrib.



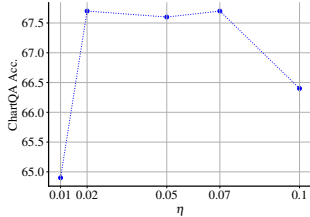
(b) RefCOCOg Acc. v.s. η for Poisson distrib.



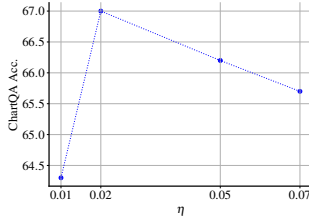
(c) RefCOCOg Acc. v.s. η for Triangular distrib.



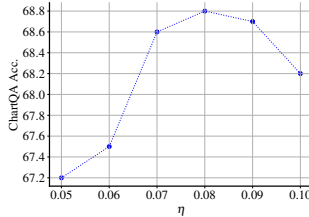
(d) RefCOCOg Acc. v.s. η for Uniform distrib.



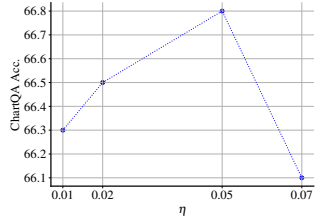
(e) ChartQA Acc. v.s. η for Binomial distrib.



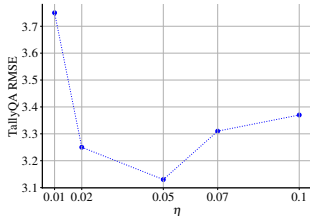
(f) ChartQA Acc. v.s. η for Poisson distrib.



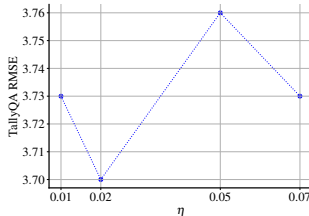
(g) ChartQA Acc. v.s. η for Triangular distrib.



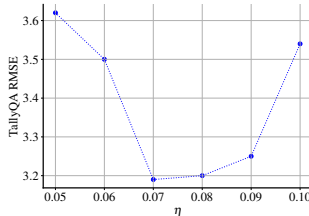
(h) ChartQA Acc. v.s. η for Uniform distrib.



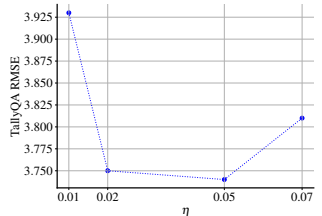
(i) TallyQA RMSE v.s. η for Binomial distrib.



(j) TallyQA RMSE v.s. η for Poisson distrib.



(k) TallyQA RMSE v.s. η for Triangular distrib.



(l) TallyQA RMSE v.s. η for Uniform distrib.

Figure 2. Ablation study of η of different distributions across three datasets, RefCOCOg val, ChartQA and TallyQA. On ChartQA, we evaluate on numerical subset. On TallyQA, we use the RMSE as metric and evaluate on simple balanced set.

Table 2. The enhancement by soft labeling on chart understanding.

Pretrain model	Model type	Finetune data	Loss	ChartQA	
				All	Numerical
LLaVA-7B	specialist	ChartQA	hard	67.9	66.3
			soft	69.6(+1.7)	68.6(+2.3)
	generalist	LLaVA-Mix + ChartQA	hard	67.0	66.5
			soft	68.3(+1.3)	68.6(+2.1)
LLaVA-13B	specialist	ChartQA	hard	72.8	72.0
			soft	73.4(+0.6)	73.6(+1.6)
	generalist	LLaVA-Mix + ChartQA	hard	71.0	69.7
			soft	72.9 (+1.9)	72.6(+2.9)

4. Experiments

In this work, we evaluate our approach on three representative numerical prediction tasks: visual grounding, object counting, and chart understanding. For visual grounding, we assess performance on three benchmarks RefCOCO (val/testA/testB), RefCOCO+ (val/testA/testB), and RefCOCOg (val/test) [25] using the standard Acc@0.5 metric. For object counting, we evaluate on TallyQA’s simple and complex test sets [1]. However, the original test sets are count-imbalanced, with ground truth counts ranging from 0 to 15 but the combined counts of 1 and 2 account for 78% of the simple testing set and 60% of the complex testing

Table 3. The enhancement by soft labeling on object counting.

Pretrain model	Model type	Finetune data	Loss	TallyQA simple		TallyQA complex		TallyQA simple balanced		TallyQA complex balanced	
				Acc. (↑)	RMSE(↓)	Acc.(↑)	RMSE(↓)	Acc. (↑)	RMSE(↓)	Acc. (↑)	RMSE(↓)
LLaVA-7B	specialist	TallyQA	hard	78.5	1.00	66.9	1.30	38.5	3.84	16.7	6.68
			soft	79.0(+0.5)	0.97(-0.03)	67.2(+0.3)	1.29(-0.01)	39.4(+0.9)	3.20(-0.64)	18.8(+2.1)	6.32(-0.36)
	generalist	LLaVA-Mix + TallyQA	hard	79.8	1.43	68.2	2.54	39.4	2.33	21.8	4.96
			soft	80.5(+0.7)	1.35(-0.08)	69.4(+1.2)	1.28(-1.26)	42.2(+2.8)	2.24(-0.09)	24.0(+2.2)	4.41(-0.55)
LLaVA-13B	specialist	TallyQA	hard	81.9	0.85	67.2	1.27	40.4	2.84	22.9	5.98
			soft	82.7(+0.8)	0.67(-0.18)	69.3(+2.1)	1.20(-0.07)	42.2(+1.8)	2.02(-0.82)	27.0(+4.1)	5.16(-0.82)
	generalist	LLaVA-Mix + TallyQA	hard	82.0	0.81	70.9	1.30	45.9	2.37	23.9	4.87
			soft	82.5(+0.5)	0.76(-0.05)	71.5(+0.6)	1.24(-0.06)	47.8(+1.9)	1.88(-0.49)	29.1(+5.2)	4.45(-0.42)

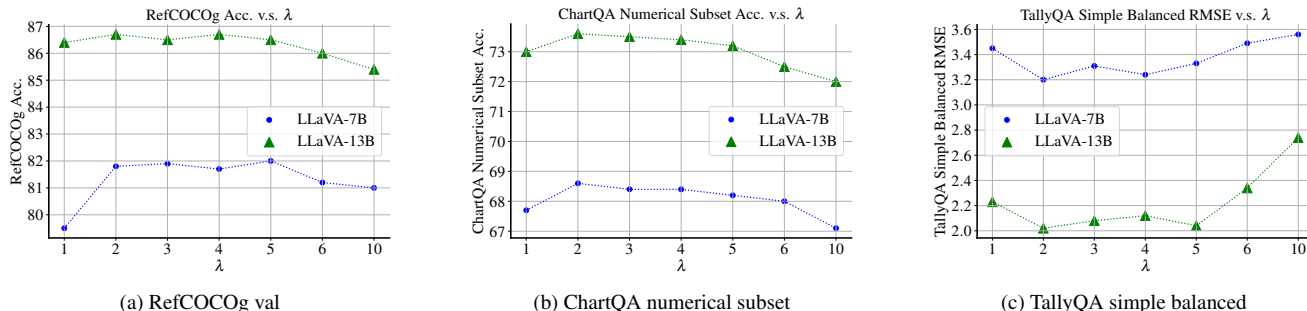


Figure 3. Ablation study of λ on three numerical prediction datasets.

Table 4. The effect of different distributions. On TallyQA, RMSE is the metric.

Distribution	RefCOCO val			TallyQA balanced		ChartQA Numerical
	*	*+	*g	Simple	Complex	
Hard Label	81.7	73.6	79.0	3.84	6.68	66.3
Binomial	84.7	76.8	81.9	3.13	6.28	67.6
Posisson	84.1	76.3	80.8	3.70	6.49	67.0
Triangular	85.0	76.7	82.0	3.20	6.32	68.6
Uniform	83.0	76.6	80.8	3.74	6.56	66.8

set. In addition, the testing sets also contain noisy labels. These make the evaluation results misleading. To address this, we also construct a balanced subset for reference by sampling a count-balanced simple and complex testing sets from TallyQA’s test sets respectively, referred to as “TallyQA simple balanced” and “TallyQA complex balanced”. The details are in the Supp. Accuracy and RMSE are the chosen metrics following [1]. For chart understanding, we use ChartQA benchmark [40], where approximately half of the questions have numerical answers. We report accuracy on both the full test set and a subset containing only numerical answers.

Following [4], we adopt a three-stage pipeline, where the first stage is to learn the vision-language (VL) adapter, the second stage is to pretrain LLM and VL adapter on a large set of pretraining data, and the third stage is to fine-tune the pretrained model on some small scale finetuning data. To evaluate the broad applicability of our approach, we conduct experiments on various pretraining and finetuning settings. For the pretraining of generalist models,

we use the publicly available LLaVA-7B architecture [32], pretrained on a sampled LAION-5M dataset [54], as our base model. For a stronger baseline, we use LLaVA-13B, pretrained on a 16M data mixture including LAION-5M for image captioning, GRIT-5M [47] for grounding, Objects365 [55] and OpenImages [27] of 3M in total for object counting, and 3M plot or chart related datasets such as PlotQA [43], UniChart [41], FigureQA [24] for chart understanding. For specialist model pretraining, we use three domain-specific data sources: a 12M mixture of grounding and detection data for visual grounding, a 6M counting data mixture with multiple tasks for object counting, and a 20M chart analysis data mixture for chart understanding, respectively. More details for pretraining data are given in Supp. After pretraining, the specialist models are finetuned on the task-specific training sets: RefCOCO, RefCOCO+, and RefCOCOg [25] for visual grounding; TallyQA [1] for object counting; and ChartQA [40] for chart understanding, and generalist models are finetuned on the same task-specific training sets but with the additional public 665K LLaVA-Mix tuning dataset [28]. The other settings are exactly the same as LLaVA’s [32].

4.1. The Effectiveness of Soft Labeling

We first compare soft labeling and one-hot hard labeling across various pretrained models, including LLaVA-7B (pretrained on LAION-5M) and LLaVA-13B (pretrained on a 16M multi-task mixture). They are fine-tuned as specialist and generalist models and evaluated on multiple tasks and benchmarks. Tables 1, 2 and 3 report the performance comparisons for visual grounding, chart understanding and

Table 5. Performance comparison to the state-of-the-art models on visual grounding (Acc@0.5).

Models	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
Shikra-7B [12]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
Ferret-7B [70]	87.5	91.4	82.5	80.8	87.4	73.1	83.9	84.8
Qwen-VL-7B [4]	88.6	92.3	84.5	82.8	88.6	76.8	86.0	86.3
InternVL2-8B [13]	87.1	91.1	80.7	79.8	87.9	71.4	82.7	82.7
VistaLLM [50]	88.1	91.5	83.0	82.9	89.8	74.8	83.6	84.4
MiniGPT-v2 [11]	88.7	91.7	85.3	80.0	85.1	74.5	84.4	84.7
Qwen2-VL-7B [4]	91.7	93.6	87.3	85.8	90.5	79.5	87.3	87.8
Ferret-v2-7B [72]	92.8	94.7	88.7	87.4	92.8	79.3	89.4	89.3
SoftLabeling-7B	91.8	94.7	88.9	87.0	92.7	80.0	89.6	89.5
LION-12B [10]	89.8	93.0	85.6	84.0	89.2	78.1	85.5	85.7
Shikra-13B [12]	87.8	91.1	81.8	82.9	87.8	74.4	82.6	83.2
Ferret-13B [70]	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3
Ferret-v2-13B [72]	92.6	95.0	88.9	87.4	92.1	81.4	89.4	90.0
SoftLabeling-13B	92.7	95.0	89.0	87.6	92.7	82.3	89.8	90.0

object counting tasks, respectively. Some observations are available. First, soft labeling consistently improves performance across all model and benchmark combinations, with most gains larger than 2 points, demonstrating its effectiveness and robustness. Second, the absolute gains are smaller for the stronger 13B model compared to the 7B model. This is expected due to performance saturation. Nonetheless, notable improvements are remained, e.g., a 2.9 points gain in accuracy on ChartQA numerical subset (generalist model) in Table 2 and 4.1/5.2 points accuracy gains on TallyQA’s simple/complex balanced test sets in Table 3. Third, on ChartQA, the gain on numerical subset is higher than the full set. This aligns with the design of soft labeling which is proposed specifically for enhancing the numerical prediction ability. Finally, on object counting, gains are more pronounced on the more reasonable balanced test sets, particularly in RMSE, a preferable metric reflecting error distance to the ground truth, where the relative improvements exceed 10% in most cases. Notably, while enhancing numerical prediction, soft labeling does not degrade performance on generic tasks (please see details in Supp.)

4.2. Ablation Study

Soft Distribution The distribution ψ in (4) is crucial because it determines the degree of softness in soft labeling and the measurement of inter-digital token similarity. We investigate four widely used distributions, Binomial [6], Poisson [6], Triangular [65] and Uniform (label smoothing) [57]. The optimal η in (4) is exhaustively searched for each distribution and used for comparison. We assess performance using some representative benchmarks and metrics on the base LLaVA-7B specialist model, as reported in Table 4. Regardless of the chosen distribution, soft labeling consistently outperforms hard labeling, demonstrating

Table 6. Performance comparison to the state-of-the-art models on ChartQA.

Models	ChartQA
ChartPaLI-5B [8]	77.3
ScreenAI 5B [3]	76.7
MatCha4096 + LaMenDa [29]	72.6
SMoLA-PaLI-X [69]	74.6
Qwen-VL-Chat [4]	66.3
Qwen-VL-Max [4]	79.8
LLaVA-OV-7B [28]	80.0
Cambrian-34B [62]	73.8
Gemini Ultra [60]	80.8
SoftLabeling	81.5

Table 7. Performance comparison to the state-of-the-art models on TallyQA. * indicates that the numbers are copied from a third-party paper [37].

Models	TallyQA simple		TallyQA complex	
	Acc.(\uparrow)	RMSE(\downarrow)	Acc.(\uparrow)	RMSE(\downarrow)
SMoLA-PaLI-X [69]	86.3	-	77.1	-
PaLI-X-VPD [22]	86.2	-	76.6	-
MoVie-ResNeXt [45]	74.9	1.00	56.8	1.43
RCN [1]	71.8	1.13	56.2	1.43
LLaVA-NEXT [33]	79.8*	0.70*	67.9*	1.76*
GPT-4V [2]	73.6*	0.86*	62.6*	1.58*
GPT-4o [23]	81.5*	0.60*	71.7*	1.21*
SoftLabeling	86.6	0.56	77.2	1.06

its robustness. Among the distributions, Binomial and Triangular yield the best results, followed by Poisson, while Uniform performs the worst. This suggests that sharper distributions are more effective, as they provide more discriminative token similarities. In contrast, smoother distributions like Uniform treat all incorrect tokens as equally distant from the target, disregarding the relative distance information. Throughout this paper, we use Triangular distribution as default, though Binomial is comparable.

Hyperparameter There are two primary hyperparameters influencing the performances, η of (4) and λ of (5). η controls the sharpness of the soft labeling. It is selected through exhaustive search based on the performance of the base LLaVA-7B specialist model on RefCOCOg val, ChartQA and TallyQA. For ChartQA, we evaluate accuracy on numerical subset. On TallyQA, we use RMSE as metric and evaluate on simple balanced set. Figure 2 shows the ablation study results. Our analysis shows that the model favors sharper soft labeling distributions with smaller η . However, if it is too small, the distribution is towards a hard one-hot encoding, leading to performance drop. On the other hand, a large η over smooths the distribution, reducing discrimination among numerical tokens. In our experiments, η is set to 0.05/0.02/0.08/0.05 for Binomial/Poisson/Triangular/U-

Image				
Prompt	first sandwich on the left just beneath the fork	the man on the closest motorcycle	What was the internet penetration rate in Africa in 2019?	How many books did Blurb publish in 2018?
Hard labeling	red box	red box	20	16000
Soft labeling	green box	green box	27.4	19053
Ground truth	blue box	blue box	28.6	19098

Table 8. Qualitative examples of soft labeling improvement in visual grounding and chart understanding.

niform distribution, respectively, by default.

As for λ of (5), it balances the contribution of numerical token loss and regular token loss in the final loss. We ablate its effect using the base 7B and stronger 13B specialist models across three datasets and present the results in Figure 3. It can be found that the models perform well and remain stable when λ ranges from 2 to 5, beyond which the performance degrades. In our experiments, we set λ as 2 by default.

4.3. Comparison to the State-of-the-Art

We evaluate soft labeling and compare it with other methods across visual grounding, chart understanding, and object counting using public benchmarks, including RefCOCO (val/testA/testB), RefCOCO+ (val/testA/testB), RefCOCOg (val/test), ChartQA (test), and TallyQA (simple/complex). The model is pretrained on scaled-up domain- and task-specific data and fine-tuned as specialist models (details in Section B of the Supp.). By default, our results are based on the 13B model, but for grounding, we follow standard practice and compare 7B and 13B models separately. As shown in Table 5, 6, and 7, our Soft-Labeling, without any architectural modifications, achieves the state-of-the-art performance on RefCOCOg among 7B scale models, RefCOCO, RefCOCO+, RefCOCOg among 13B scale models, ChartQA, and TallyQA. It also delivers competitive results on RefCOCO and RefCOCO+ in 7B scale model group. These results highlight the effectiveness of soft labeling.

4.4. Visualization

Table 8 presents examples randomly selected from visual RefCOCOg and ChartQA test sets, demonstrating the improvement achieved by soft labeling. Compared to hard labeling, soft labeling produces more accurate numerical predictions for both bounding box coordinates and chart-based values, bringing them closer to the ground truth. For instance, in the motorcycle image, the bounding box predicted by soft labeling (green) better aligns with the ground truth (blue) in both position and scale. Similarly, in the chart-

based QA task, soft labeling provides more accurate numerical estimation. These examples highlight the effectiveness of soft labeling in improving numerical prediction across both spatial and quantitative tasks.

5. Limitation

In this work, we limit our scope to digit tokens. However, other tokens also exhibit semantic similarity, such as natural language numerals (“one/two/three”) or categorical tokens (“easy/medium/hard”). Incorporating these into the loss computation is an avenue for future research. Another limitation is that our current loss function considers only single digits. Many numerical predictions involve multi-digit numbers, where digit-wise comparisons may not accurately reflect actual numerical distances. For instance, comparing “21” v.s. “32” digit-wise results in a distance of $|2-3|+|1-2|=2$ which is smaller than $|2-1|+|2-9|=8$ for “21” v.s. “29”. However, their actual numerical distances are $|21-32|=11$ and $|21-29|=8$, showing a discrepancy in digit-wise calculations. Addressing this limitation remains an open challenge.

6. Conclusion

In this paper, we have identified the limitations of cross-entropy loss with one-hot encoding for numerical prediction in LLMs/MLLMs, where treating numbers as discrete tokens fails to reflect their continuous nature. To address this, we have proposed soft labeling, which smooths target distributions to account for numerical proximity, ensuring that closer predictions incur lower loss. This method retains compatibility with the next-token prediction paradigm without requiring architectural modifications. Our experiments across visual grounding, object counting, and chart understanding have shown that soft labeling consistently improves numerical prediction performance. We have achieved state-of-the-art or competitive results on RefCOCO+/g, ChartQA, and TallyQA using a basic LLaVA model. Our findings highlight the importance of adapting appropriate loss functions for numerical tasks in LLMs/M-LLMs.

References

- [1] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tal-lyqa: Answering complex counting questions. In *AAAI*, pages 8076–8084, 2019. 2, 5, 6, 7
- [2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 7
- [3] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*, 2024. 7
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(8), 2023. 1, 6, 7
- [5] Javier Barbero-Gómez, Pedro-Antonio Gutiérrez, Víctor-Manuel Vargas, Juan-Antonio Vallejo-Casas, and César Hervás-Martínez. An ordinal cnn approach for the assessment of neurological damage in parkinson’s disease patients. *Expert Systems with Applications*, 182:115271, 2021. 2
- [6] Christopher Beckham and Christopher Pal. Unimodal probability distributions for deep ordinal classification. In *ICML*, pages 411–419. PMLR, 2017. 3, 4, 7
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 1
- [8] Victor Carbune, Hassan Mansoor, Fangyu Liu, Rahul Aralikatte, Gilles Baechler, Jindong Chen, and Abhanshu Sharma. Chart-based reasoning: Transferring capabilities from llms to vlms. *arXiv preprint arXiv:2403.12596*, 2024. 7
- [9] Alberto Castaño, Pablo González, Jaime Alonso González, and Juan Jose Del Coz. Matching distributions algorithms based on the earth mover’s distance for ordinal quantification. *IEEE T-NNLS*, 35(1):1050–1061, 2022. 3
- [10] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *CVPR*, pages 26540–26550, 2024. 7
- [11] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 7
- [12] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2, 7
- [13] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, pages 24185–24198, 2024. 7
- [14] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. *arXiv preprint arXiv:2311.17667*, 2023. 2
- [15] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, pages 4738–4747, 2019. 3
- [16] Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*, 2024. 2
- [17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 2
- [18] 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
- [19] Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervás-Martínez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015. 3
- [20] Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared earth mover’s distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016. 3
- [21] Heng-Wei Hsu, Tung-Yu Wu, Sheng Wan, Wing Hung Wong, and Chen-Yi Lee. Quatnet: Quaternion-based head pose estimation with multiregression loss. *IEEE Transactions on Multimedia*, 21(4):1035–1046, 2018. 2
- [22] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *CVPR*, pages 9590–9601, 2024. 7
- [23] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7
- [24] 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. 6
- [25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 2, 5, 6
- [26] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. 4

- [27] 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. 6
- [28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7
- [29] Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. Synthesize step-by-step: Tools templates and llms as data generators for reasoning-based chart vqa. In *CVPR*, pages 13613–13623, 2024. 7
- [30] Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. The devil is in the margin: Margin-based label smoothing for network calibration. In *CVPR*, pages 80–88, 2022. 3
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023. 1
- [32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024. 2, 6
- [33] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7
- [34] Peiyang Liu, Xiangyu Xi, Wei Ye, and Shikun Zhang. Label smoothing for text mining. In *ICCL*, pages 2210–2219, 2022. 3
- [35] Xiaofeng Liu, Yang Zou, Yuhang Song, Chao Yang, Jane You, and BV K Vijaya Kumar. Ordinal regression with neuron stick-breaking for medical diagnosis. In *ECCV Workshops*, pages 0–0, 2018. 2
- [36] Xiaofeng Liu, Fangfang Fan, Lingsheng Kong, Zhihui Diao, Wanqing Xie, Jun Lu, and Jane You. Unimodal regularized neuron stick-breaking for ordinal classification. *Neurocomputing*, 388:34–44, 2020. 3
- [37] Jian Lu, Shikhar Srivastava, Junyu Chen, Robik Shrestha, Manoj Acharya, Kushal Kafle, and Christopher Kanan. Revisiting multi-modal llm evaluation. *arXiv preprint arXiv:2408.05334*, 2024. 7
- [38] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2
- [39] Michal Lukasik, Srinadh Bhojanapalli, Aditya Menon, and Sanjiv Kumar. Does label smoothing mitigate label noise? In *ICML*, pages 6448–6458. PMLR, 2020. 3
- [40] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022. 2, 6
- [41] 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. 6
- [42] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209, 2021. 2
- [43] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *WACV*, pages 1527–1536, 2020. 2, 6
- [44] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *NeurIPS*, 32, 2019. 3
- [45] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. Movie: Revisiting modulated convolutions for visual counting and beyond. *arXiv preprint arXiv:2004.11883*, 2020. 7
- [46] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, pages 4920–4928, 2016. 2
- [47] 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. 6
- [48] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, 2024. 2
- [49] Gorkem Polat, Ilkay Ergenc, Haluk Tarik Kani, Yesim Ozen Alahdab, Ozlen Atug, and Alptekin Temizel. Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. In *Annual Conference on Medical Image Understanding and Analysis*, pages 157–171. Springer, 2022. 3
- [50] Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chelappa, and Amjad Almahairi. Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In *CVPR*, pages 14076–14088, 2024. 7
- [51] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momen-tor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*, 2024. 2
- [52] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1
- [53] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *CVPR*, pages 3394–3403, 2021. 2
- [54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 6
- [55] 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. 6
- [56] 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

- [57] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 7
- [58] Zichang Tan, Shuai Zhou, Jun Wan, Zhen Lei, and Stan Z Li. Age estimation based on a single network with soft softmax of aging modeling. In *ACCV*, pages 203–216. Springer, 2016. 3
- [59] Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. Visualmrc: Machine reading comprehension on document images. In *AAAI*, pages 13878–13888, 2021. 2
- [60] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 7
- [61] Christian Thiel. Classification on soft labels is robust against label noise. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 65–73. Springer, 2008. 3
- [62] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024. 7
- [63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [64] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *ICCV*, pages 4068–4076, 2015. 3
- [65] Víctor Manuel Vargas, Pedro Antonio Gutiérrez, Javier Barbero-Gómez, and César Hervás-Martínez. Soft labelling based on triangular distributions for ordinal classification. *Information Fusion*, 93:258–267, 2023. 3, 4, 7
- [66] Sukesh Adiga Vasudeva, Jose Dolz, and Herve Lombaert. Geols: geodesic label smoothing for image segmentation. In *Medical Imaging with Deep Learning*, pages 468–478. PMLR, 2024. 3
- [67] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *NeurIPS*, 37:95095–95169, 2024. 2
- [68] Scott Workman, Richard Souvenir, and Nathan Jacobs. Understanding and mapping natural beauty. In *ICCV*, pages 5589–5598, 2017. 3
- [69] Jialin Wu, Xia Hu, Yaqing Wang, Bo Pang, and Radu Soricut. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. In *CVPR*, pages 14205–14215, 2024. 7
- [70] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023. 2, 7
- [71] Chang-Bin Zhang, Peng-Tao Jiang, Qibin Hou, Yunchao Wei, Qi Han, Zhen Li, and Ming-Ming Cheng. Delving deep into label smoothing. *IEEE T-IP*, 30:5984–5996, 2021. 3
- [72] Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024. 7
- [73] Helong Zhou, Liangchen Song, Jiajie Chen, Ye Zhou, Guoli Wang, Junsong Yuan, and Qian Zhang. Rethinking soft labels for knowledge distillation: A bias-variance tradeoff perspective. *arXiv preprint arXiv:2102.00650*, 2021. 3