Unsupervised Visual Chain-of-Thought Reasoning via Preference Optimization

Kesen Zhao[1] **Beier Zhu**[1*] **Qianru Sun**[2] **Hanwang Zhang**[1]
[1]Nanyang Technological University, [2]Singapore Management University
kesen002@e.ntu.edu.sg, qianrusun@smu.edu.sg
{beier.zhu, hanwangzhang}@ntu.edu.sg

Abstract

Chain-of-thought (CoT) reasoning greatly improves the interpretability and problem-solving abilities of multimodal large language models (MLLMs). However, existing approaches focus on text CoT, limiting their ability to leverage visual cues. Visual CoT remains underexplored, and the only work [35] is based on supervised fine-tuning that relies on extensive labeled bounding-box data and is hard to generalize to unseen cases. In this paper, we introduce Unsupervised Visual CoT (UV-CoT), a novel framework for image-level CoT reasoning via preference optimization. UV-CoT performs preference comparisons between model-generated bounding boxes (one is preferred and the other is dis-preferred), eliminating the need for bounding-box annotations. We get such preference data by introducing an automatic data generation pipeline. Given an image, our target MLLM (e.g., LLaVA-1.5-7B) generates seed bounding boxes using a template prompt and then answers the question using each bounded region as input. An evaluator MLLM (e.g., OmniLLM-12B) ranks the responses, and these rankings serve as supervision to train the target MLLM with UV-CoT by minimizing negative log-likelihood losses. By emulating human perception–identifying key regions and reasoning based on them–UV-CoT can improve visual comprehension, particularly in spatial reasoning tasks where textual descriptions alone fall short. Our experiments on six datasets demonstrate the superiority of UV-CoT, compared to the state-of-the-art textual and visual CoT methods. Our zero-shot testing on four unseen datasets shows the strong generalization of UV-CoT.

1. Introduction

With the recent advancements in multimodal large language models (MLLMs) [3, 21, 22, 45], many efforts have been made to incorporate text CoT reasoning [9, 17, 40, 49] to handle complex vision-language tasks [42, 43, 50]. However, the visual understanding ability of MLLM is limited by fixed-granularity image processing, *i.e.*, the MLLM cannot dynamically adjust focus across different spatial regions of the input image, even when guided by text CoT prompts [11]. This underscores the critical need to explicitly integrate visual cues into the CoT process.

A very recent work, Visual-CoT [35], has made an initial attempt towards this goal. The model is trained using supervised fine-tuning (SFT) with human-labeled bounding boxes that indicate the key image regions relevant to the question. It performs the multimodal CoT approach with human-annotated reasoning steps by associating textual inputs with the detected regions. An overview of Visual-CoT is presented in Fig. 1. However, it has two key drawbacks: (1) it relies on large-scale, high-quality labeled data, making it costly and hard to scale; and (2) SFT learns only from positive examples (*i.e.*, the labeled data), limiting its ability to generalize to unseen or ambiguous scenarios where intermediate reasoning or dynamic interpretation is needed.

To address these issues, we introduce an Unsupervised approach to Visual CoT dubbed as UV-CoT. It has two key parts: data collection and model training. The data collection does not need human annotation, as it leverages the data generation and evaluation capabilities of pre-trained MLLMs. The model training is inspired by the idea of direct preference optimization (DPO). It is implemented with an adapted version of DPO to address specific limitations in capturing the degree of preference and fine-grained region-based reasoning when conducting visual CoT on MLLMs.

Our method UV-CoT, as shown in Fig. 1, differs from Visual-CoT [35] by adopting an unsupervised approach with contrastive preference data. We design an automatic two-step pipeline to generate this data: 1) Region Generation: Given an image, the target model generates multiple seed bounding boxes using a template prompt. Then it answers the question by processing each bounded region together with the question as input. 2) Quality Assessment: An evaluator MLLM scores the responses, using these scores as proxies to measure the quality of the regions.

*Corresponding author. The code is available in https://github.com/kesenzhao/UV-CoT.

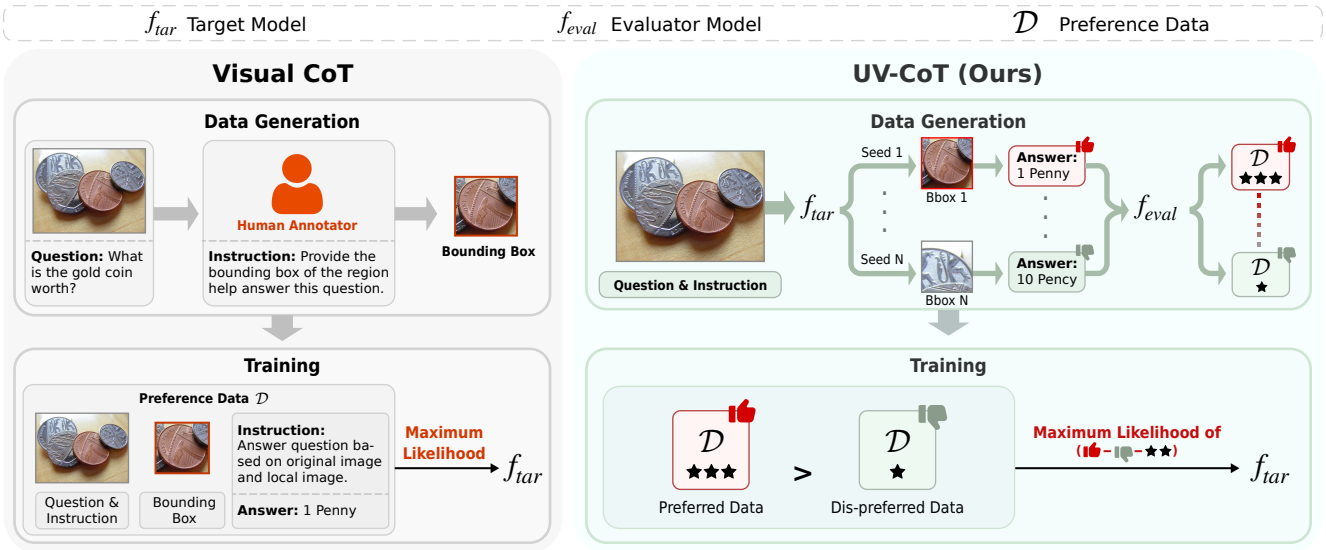

Figure 1. Comparison of Visual-CoT [35] and our UV-CoT. **Left**: Visual-CoT relies on human-annotated bounding boxes to identify key regions. The model is trained via supervised fine-tuning to maximize the likelihood of the labeled data. **Right**: UV-CoT eliminates the need for human annotation. Given an image, the target model generates seed bounding boxes and answers questions based on these regions, respectively. An evaluator MLLM then scores the responses as a proxy for assessing region quality. Lastly, the target model is optimized via preference optimization by maximizing the likelihood of preferred regions over dis-preferred ones.

Unlike traditional DPO, we propose Score-DPO (sDPO), which not only ranks preference data (i.e., preferred and dis-preferred responses shown in Fig. 1) but also assigns preference scores. This scoring enables more precise optimization based on score differences. During UV-CoT training, the rankings of the preference data act as supervision by minimizing negative log-likelihood loss, while the scores define the decision margin. By mimicking human perception–first identifying key regions, then reasoning over them–UV-CoT significantly improves visual comprehension, especially in spatial reasoning tasks where text-based methods fall short. By leveraging unsupervised data in a contrastive way, UV-CoT also shows strong generalization ability when tested on unseen datasets.

Our contributions in this paper include: 1) an automatic pipeline for generating high-quality preference data, enabling robust and scalable preference learning of UV-CoT; 2) an improved version of DPO by integrating the degree of preference for visual regions, allowing the model to distinguish key regions more precisely; and 3) extensive experiments on multiple challenging datasets, demonstrating state-of-the-art performance of UV-CoT on four benchmarks and strong generalization to four unseen test datasets.

2. Related Works

Chain-of-thought in LLMs and MLLMs. LLMs [3, 10, 14, 24, 29, 38] with CoT [5, 8, 40] show strong inferential abilities by introducing intermediate reasoning steps. Both manually designed [40] and self-generated [17, 49] reasoning approaches have proven effective. In contrast, MLLMs rely on image encoders [15, 32, 52–54] to extract visual features but often struggle with reasoning due to inherent differences in how textual and visual data are processed [23, 48, 51]. Multimodal CoT methods [42, 43, 50] attempt to address this by transforming multimodal inputs into a unified textual format, enabling LLMs to perform CoT at the text level. However, this transformation introduces significant information loss and prevents the models from capturing key visual details [50]. For example, LLaVA-CoT [42] leverages GPT-4o to summarize questions and image captions but suffers from weak optical character recognition and sometimes hallucinations.

A very recent work, Visual-CoT [35], improves the MLLM reasoning by introducing CoT methods at the image level. This approach involves scanning the entire image, identifying key references, and then focusing the model on specific regions for reasoning. Despite its improvements, Visual-CoT is heavily based on costly human-annotated data. In contrast, our UV-CoT framework utilizes unsupervised preference optimization with auto-generated preference data, eliminating the need for manual annotations.

Preference learning in LLMs and MLLMs. RLHF [56] aligns LLMs with human preferences by training a reward model via contrastive response evaluations. To reduce reliance on human annotations, RLAIF [19] leverages pre-trained LLMs for preference label generation. However, RL-based fine-tuning faces stability and efficiency challenges. Direct Preference Optimization (DPO) [33] ad-

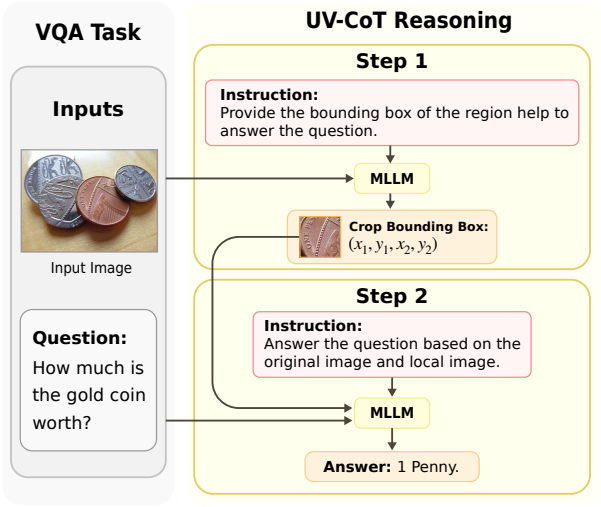

Figure 2. Illustration of UV-CoT reasoning.

dresses this by directly linking reward functions to optimal policies, eliminating reward model fine-tuning. Further improvements include IPO [2], which mitigates overfitting with a bounded preference function, and KTO [7], which removes the need for paired preference data, relying instead on single examples labeled as either 'good' or 'bad'.

These preference learning techniques are applied to MLLMs, with RLHF-V and RLAIF-V [46, 47] refining behavior alignments using human and LLM-generated labels. To mitigate reward hacking, Sun et al. [37] enhance reward models with additional factual details, such as image captions and verified choices, further improving the performance of MLLM. Some works have attempted to apply DPO in the CoT process [18, 30]. However, these methods are designed for only text-level CoT and do not effectively handle visual features or cues. In contrast, in this paper, we propose UV-CoT, a specialized framework for image-level CoT reasoning inspired by the idea of DPO. Unlike traditional DPO, UV-CoT not only ranks preference data (i.e., preferred and dis-preferred data) but also assigns preference scores. This scoring enables more precise optimization of the MLLMs based on score differences.

3. Method

Fig. 2 illustrates the pipeline of UV-CoT reasoning. Given the original image and the question, we append a CoT prompt to guide the target MLLM in identifying the most informative image region and specifying its location via bounding box coordinates. A visual sampler then extracts the bounded region from the image. The MLLM subsequently integrates visual tokens from original and cropped images to generate more precise and comprehensive answers. In Sec. 3.1, we detail the automatic preference data generation pipeline. In Sec. 3.2, we describe our Score-DPO (sDPO) for image-level CoT reasoning.

Algorithm 1 Preference data generation for a query x

Input: Target model f_{tar}, evaluator f_{eval}, an image-question query x, number of seeds n, and number of preference pairs k.
Output: Preference data $\mathcal{D}$
Initialize $y_0 = x$
for $t = 1$ to T **do**
  $\{y_t^i\}_{i=1}^n \leftarrow \text{Generate}(f_{\text{tar}}, y_{0:t-1}, n)$
  $\{s^i\}_{i=1}^n \leftarrow \text{Evaluate}(f_{\text{eval}}, y_{0:t-1}, \{y_t^i\}_{i=1}^n)$
  $\mathcal{D}_t \leftarrow \text{ConstructPairs}(y_{0:t-1}, \{y_t^i\}_{i=1}^n, \{s^i\}_{i=1}^n, k)$
  $y_t \leftarrow \text{Select}(y_{0:t-1}, \{y_t^i\}_{i=1}^n, \{s^i\}_{i=1}^n)$
end for
return $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$

3.1. Preference Data Generation

Given a target model f_{tar}, an evaluator model f_{eval} (the target MLLM can also serve as the evaluator, validated in Tab. 5), and an image-question pair x, we illustrate how to construct n preference data points. Assuming there are T steps in CoT reasoning process, we generate preference data for T times on the way, as described in Algorithm 1. At each timestep t (*i.e.*, a reasoning step t), the process includes four stages: Response Generation, Response Evaluation, Pair Construction, and Response Selection.

Response generation. The goal of this stage is to generate seed bounding boxes and produce intermediate responses to the question using the target model. Here, a 'response' refers to any model output at an intermediate step, not necessarily the final answer to the question. We denote the model's response at timestep t as y_t, with the initial input x represented as y_0. To encourage diversity in the bounding boxes and subsequent responses, we apply stochastic decoding to the target model f_{tar} with n different random seeds, resulting in a set of responses $\{y_t^i\}_{i=1}^n$.

Response evaluation. This stage evaluates the quality of all generated responses. The evaluator model not only scores each individual response but also counts how this response influences the quality of its next response in the chain. This cumulative evaluation approach helps quantify the impact of each bounded region on the overall reasoning process. Below, we elaborate on the formulation. At timestep t, the evaluator assigns scores for y_t^i as follows:

$$\begin{aligned} s_{\text{cur}}^i &= f_{\text{eval}}(y_t^i \mid y_{0:t-1}), \\ s_{\text{nxt}}^i &= \mathbb{E}[f_{\text{eval}}(y_{t+1}^i \mid y_{0:t-1}, y_t^i)], \\ s^i &= s_{\text{cur}}^i + \gamma s_{\text{nxt}}^i, \end{aligned} \tag{1}$$

where s_{nxt}^i reflects the impact on the next response and $\gamma > 0$ is a hyperparameter to combine the current and next

response scores, with $\gamma = 0$ at the last step. We estimate the expectation $\mathbb{E}[\cdot]$ by randomly sampling next responses.

Pair construction. At each timestep t, we randomly select k (preferred and dis-preferred) pairs from $\{y_t^i\}_{i=1}^n$. For a single pair, we concatenate the preferred response with the past response chain $y_{0:t-1}$ (preserved after $t-1$ timesteps) and get a "preferred chain" denoted as y_t^w, and then we concatenate the dis-preferred response in the same way to get a 'dis-preferred chain' denoted as y_t^l. The pair of chains also includes their respective scores, and they are represented as $\{y_w, s_w, y_l, s_l\}$. The overall k pairs of chains compose the preference dataset $\mathcal{D}_t$.

Response selection. The abovementioned 'past response chain $y_{0:t-1}$' is unique and is concatenated by the highest-scoring response at timestep $t-1$ and the preserved chain at timestep $t=2$, *i.e.*, $y_{0:t-2}$. In other words, when finishing each timestep process, we preserve only the best chain and use it for the next step.

3.2. Unsupervised Learning of UV-CoT

Assume the preference dataset $\mathcal{D}$ has been generated across t timesteps, we then optimize the target model with our UV-CoT via preference optimization on $\mathcal{D}$. DPO [1] is widely used in preference learning, and it ranks responses without quantifying preference intensity. In our case of image-level reasoning, key regions vary in influence, necessitating finer differentiation between responses. Therefore, we refine DPO by adjusting the margin to capture the key region's influence. We name our loss Score-DPO, abbreviated as **sDPO**, as it incorporates the preference score into the optimization. The loss is formulated as:

$$\mathcal{L}_{\text{sDPO}}(\theta) = -\mathop{\mathbb{E}}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} - (g(s_w) - g(s_l))\right)\right], \tag{2}$$

where π_θ is the target model, and π_{ref} is its frozen initialization, serving as a reference to constrain deviation from the original model. $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is a monotonically increasing function that maps preference scores into the logit space of the DPO objective.

To provide a deeper understanding of our sDPO loss, we establish its connection to the standard DPO loss. DPO reformulates reward model training as policy optimization by reparameterizing the reward function in PPO [34]:

$$r(x, y) = \beta\log\frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta\log Z(x) \tag{3}$$

Using the Bradley-Terry [4] model, the standard DPO aims to minimize the negative log-likelihood of the difference of rewards between paired responses:

$$\begin{aligned} P\left(y_w - y_l > 0\right) &= \sigma\left(r\left(x, y_w\right) - r\left(x, y_l\right)\right) \\ \mathcal{L}_{\text{DPO}} &= -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log P\left(y_w - y_l > 0\right)\right], \end{aligned} \tag{4}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the sigmoid function.

Let $\Delta_r = g(s_w) - g(s_l)$ and define the Gumbel-distributed random variables $R_w \sim \text{Gumbel}\left(r\left(x, y_w\right), 1\right)$ and $R_l \sim \text{Gumbel}\left(r\left(x, y_l\right), 1\right)$. Then, we derive:

$$\begin{aligned} P\left(R_w - R_l > \Delta_r\right) &= \sigma\left(r\left(x, y_w\right) - r\left(x, y_l\right) - \Delta r\right) \\ &= \sigma\left(\beta\log\frac{\pi^*\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \beta\log\frac{\pi^*\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)} - \Delta r\right) \end{aligned} \tag{5}$$

This result follows from the definition of Gumbel random variables [1] and the Gumbel-max trick [25]. We provide a detailed proof in Appendix B. By maximizing the log-likelihood, we obtain our proposed loss function. The Gumbel distribution models the extreme values of a variable, while Δ_r quantifies the degree of difference between preference pairs. Thus, our loss function explicitly optimizes preference learning by distinguishing not only the order but also the magnitude of preference differences.

Iterative learning. Standard DPO relies on static preference data during training, which can lead to distributional mismatch between training data and the model's generated outputs. To mitigate this issue, we adopt an iterative learning approach inspired by [47]. Algorithm 2 outlines the iterative learning process of UV-CoT, which incrementally refines a target model through preference learning. The iteration repeats m times, and the total image-question query set $\mathcal{X}$ is evenly divided into m subsets, $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$, each assigned to one iteration. The process starts with an initial target model f_{tar}, an evaluator model f_{eval}, and $\mathcal{X}$. In each iteration i, the algorithm first generates preference data $\mathcal{D}_i$ using the current target model f_{tar}^i and the subset $\mathcal{X}_i$, using the procedure in Algorithm 1. This newly generated preference data is then used to train the next iteration of the target model f_{tar}^{i+1} using our UV-CoT loss of Eq. (2). The process continues until the final model f_{tar}^m is obtained. By dynamically updating the preference data, our approach

Algorithm 2 Iterative learning of UV-CoT

Input: Initial target model f_{tar}^1, evaluator model f_{eval}, $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$, each $\mathcal{X}_i$ is a subset of image-question query
Output: f_{tar}^m
for $i = 1$ to m **do**
  $\mathcal{D}_i \leftarrow \text{GenerateData}(f_{\text{tar}}^i, f_{\text{eval}}, \mathcal{X}_i)$ ▷ Algorithm 1
  $f_{\text{tar}}^{i+1} \leftarrow \text{Train}(f_{\text{tar}}^i, \mathcal{D}_i)$ ▷ Eq. (2)
end for
return f_{tar}^m

MLLM	DocVQA	TextVQA	InfographicsVQA	Flickr30k	GQA	VSR	Average
LLaVA-1.5-7B	0.198	0.507	0.131	0.539	0.480	0.504	0.393
LLaVA-1.5-13B	0.225	0.543	0.169	0.607	0.506	0.512	0.427
MiniCPM-o-8B	0.232	0.529	0.175	0.618	0.495	0.521	0.428
OmniLMM-12B	0.254	0.578	0.172	0.621	0.509	0.523	0.443
Visual-CoT-7B (100% label)	**0.294**	0.673	0.194	**0.652**	0.546	0.532	0.482
UV-CoT (0% label)	0.265	0.686	0.173	0.632	0.536	0.548	0.473
UV-CoT (10% label)	0.283	**0.711**	**0.198**	0.649	**0.568**	**0.553**	**0.494**

Table 1. **Overall comparison** of different models on six evaluation benchmarks. The **best** result is bold, the second-best is underlined. '%' indicates the percentage of supervised data used in Visual-CoT. By default, our UV-CoT uses only unsupervised data.

MLLM	DUDE	SROIE	Visual7w	Average
LLaVA-1.5-7B	0.165	0.147	0.340	0.217
LLaVA-1.5-13B	0.174	0.159	0.352	0.228
MiniCPM-o-8B	0.182	0.165	0.341	0.229
OmniLMM-12B	0.194	0.166	0.357	0.239
Visual-CoT-7B	0.206	0.181	0.397	0.261
UV-CoT	0.241	0.184	0.432	0.286
UV-CoT*	**0.253**	**0.227**	**0.455**	**0.312**

Table 2. **Zero-shot experiments** on DUDE, SROIE and Visual7w. 'UV-CoT*' denotes our model trained with additional unlabeled preference data from the three zero-shot datasets.

MLLM	Attributes	GPT4V-hard	OCR	Average
LLaVA-1.5-7B	0.317	0	0.1	0.139
LLaVA-1.5-13B	0.326	0.118	0.133	0.192
MiniCPM-o-8B	0.322	0.118	0.1	0.180
OmniLMM-12B	0.326	0.118	0.167	0.204
Visual-CoT-7B	0.330	0.118	0.593	0.347
UV-CoT	**0.352**	**0.176**	**0.677**	**0.402**

Table 3. **Zero-shot experiments** on V* Bench (high-resolution image reasoning task, average resolution 2246×1582).

ensures that the learned model adapts to its evolving distribution, enhancing training robustness.

4. Experiments

4.1. Setup

Datasets. For a comprehensive evaluation, we adopt ten datasets spanning five domains: (1) Text/Document: DocVQA [26], TextVQA [36], DUDE [39], and SROIE [12]. (2) Chart: InfographicsVQA [27]. (3) General VQA: Flickr30k [31] and Visual7W [55]. (4) Relation Reasoning: VSR [20] and GQA [13]. (5) High-Resolution Image Reasoning: V* Bench [41]. Notably, we also provide a model trained on data excluding DUDE, SROIE, Visual7W, and V* Bench, which is used to evaluate zero-shot performance. See Appendix C.1 for details.

Evaluation. Following prior work [22], we prompt GPT-4o [28] to assign a score between 0 and 1, with higher scores indicating better prediction. See details in Appendix C.2.

Baselines. We compare UV-CoT with five baselines. LLaVA-1.5-{7B, 13B} [21] and OmniLMM-12B are strong general baselines. MiniCPM-o-8B [44] adopts adaptive visual encoding for fine-grained understanding and Visual-CoT-7B [35] learns image-level CoT via SFT.

Implementation details. We use LLaVA-1.5-7B as the target model and OmniLMM-12B as the evaluator. To ensure scalability, we avoid using GPT-4 as the evaluator, preventing constraints imposed by high API costs. We implement iterative learning of UV-CoT over four iterations, utilizing a total of 249K preference data pairs. Notably, UV-CoT achieves higher data efficiency than Visual-CoT, which uses 376K data pairs. For each iteration, we train the model with AdamW optimizer for 4 epochs with a learning rate of 5×10^{-7}, $\beta = 0.1$, and a batch size of 8. In total, data generation takes 80 hours and training requires 60 hours, both conducted on an 8×A100 40GB machine. Additionally, we provide a variant UV-CoT trained with extra SFT on 10% of the labeled Visual-CoT data, denoted as UV-CoT (10%).

4.2. Comparison with State-of-the-Art Methods

The overall performance comparisons are reported in Tab. 1, leading to the following key observations:

Explicitly incorporating visual cues proves beneficial for multimodal reasoning. For instance, MiniCPM-o-8B, which uses rule-based cropping to focus on salient regions, outperforms LLaVA-1.5-7B by an average of 3.5%. However, its reliance on heuristics limits adaptability. In contrast, image-level CoT models like Visual-CoT and our UV-CoT achieve greater performance gains, even surpassing the larger LLaVA-1.5-13B, by leveraging MLLMs to adaptively generate key visual regions. This highlights the superior effectiveness of image-level CoT reasoning.

Model	DocVQA	TextVQA	InfographicsVQA	Flickr30k	GQA	VSR	Average
UV-CoT (10% labels)	0.283	0.711	0.198	0.649	0.568	0.553	0.494
IF: w/o UV-CoT	0.149	0.574	0.160	0.585	0.509	0.522	0.417
IF: UV-CoT w/ G.T. BBox	0.528	0.769	0.504	0.655	0.664	0.585	0.618
Tr: w/ naive DPO	0.258	0.695	0.189	0.623	0.552	0.534	0.475
Tr: w/o iterative learning	0.256	0.671	0.162	0.609	0.523	0.531	0.459
Ge: w/o γ	0.247	0.539	0.152	0.551	0.484	0.460	0.406

Table 4. **Ablation study** on key components of UV-CoT (10% labeled data). 'w/o UV-CoT' denotes standard inference without CoT reasoning. 'UV-CoT w/ G.T. BBox' uses annotated ground truth bounding boxes. 'w/ naive DPO' applies the standard DPO loss. 'w/o iterative learning' generates preference pairs for the entire set of question queries $\mathcal{X}$ in a single pass and trains once. 'w/o γ' evaluates responses with $\gamma = 0$. IF, Tr, and Ge indicate ablations on the inference process, training loss, and data generation, respectively.

UV-CoT fundamentally differs from distillation/pseudo-labeling. Unlike distillation, where student performance is typically bounded by the teacher, UV-CoT outperforms its evaluator OmniLMM-12B by 5.1% on average. This suggests that UV-CoT goes beyond simply mimicking a larger model. Direct generation of accurate bounding boxes remains a challenge for MLLMs due to the need for precise spatial localization. Instead, UV-CoT reformulates the task as ranking—an inherently simpler and more tractable problem—which leads to better performance.

UV-CoT outperforms the supervised Visual-CoT. Despite using significantly less data (249K unlabeled vs. 376K labeled), UV-CoT outperforms Visual-CoT-7B on TextVQA (+1.3%) and VSR (+1.6%). Moreover, UV-CoT(10%) surpasses Visual-CoT-7B by 2.1% on average, with notable gains on TextVQA (+2.5%), GQA (+2.2%), and VSR (+2.1%) and achieves comparable or better performance on the remaining datasets. This validates the effectiveness of our high-quality preference data generation and enhanced preference optimization method.

4.3. Zero-Shot Generalization

We evaluate the zero-shot performance of UV-CoT on the test sets of SROIE, DUDE, Visual7W, and V* Bench [41], *without any training exposure to these datasets*. V* Bench is a high-resolution benchmark (avg. size: 2246×1582) covering diverse tasks; we focus on three representative ones: Attributes (object attribute recognition), GPT4V-Hard (complex visual reasoning), and OCR. Additionally, we train a variant of our model, UV-CoT*, using preference data generated from the training splits of SROIE, DUDE, and Visual7W. The results are reported in Tab. 2 and Tab. 3, leading to the following observations:

UV-CoT exhibits stronger zero-shot performance. Supervised CoT learning relies on labeled data and often overfits to specific annotation distributions, limiting its generalization to unseen tasks. In contrast, UV-CoT uses preference optimization based on relative comparisons, avoiding reliance on absolute labels and enhancing generalization. As a result, UV-CoT outperforms all baselines across zero-shot datasets (+2.5% on average), with notable gains on DUDE and Visual7W (both +3.5%). Furthermore, UV-CoT* achieves greater gains (5.1% on average), showing that our model can effectively learn the image-level CoT process without the need for costly human annotation.

UV-CoT excels in high-resolution image reasoning. Image-level CoT methods—UV-CoT (+19.8%) and Visual-CoT-7B (+14.3%)—significantly outperform non-CoT baselines, with over 50% gains on OCR tasks. UV-CoT further surpasses Visual-CoT-7B by 5.5% on average, achieving the best performance across all tasks. This large performance gap underscores the advantage of unsupervised image-level CoT for high-resolution visual reasoning.

4.4. Ablation Studies

In Tab. 4, we present the ablations on key components.

Image-level CoT: We design two variants to evaluate the impact of image-level CoT: (1) 'w/o UV-CoT' removes intermediate reasoning and directly outputs answers; (2) 'UV-CoT w/ G.T. BBox' replaces predicted regions with ground truth bounding boxes to assess localization quality. Results show that removing CoT leads to a significant drop (–7.7% on average), confirming its necessity. On Flickr30k, UV-CoT matches the G.T. variant, suggesting accurate region selection. However, in DocVQA and InfographicsVQA, using G.T. boxes yields better performance, revealing the difficulty of precise localization. This highlights the potential of our method and suggests future work for improving region accuracy for complex tasks.

Score-DPO (sDPO): 'w/ naive DPO' applies standard DPO loss and shows degraded performance across all datasets (1.9% on average), especially on DocVQA (2.5%) and Flickr30k (2.6%), revealing its limitations for CoT learning. In contrast, our sDPO incorporates preference scores to better differentiate choices, yielding consistent gains.

Iterative learning: 'w/o iterative learning' generates all preference pairs in a single pass and trains only once, leading to a significant performance drop (3.5% on average).

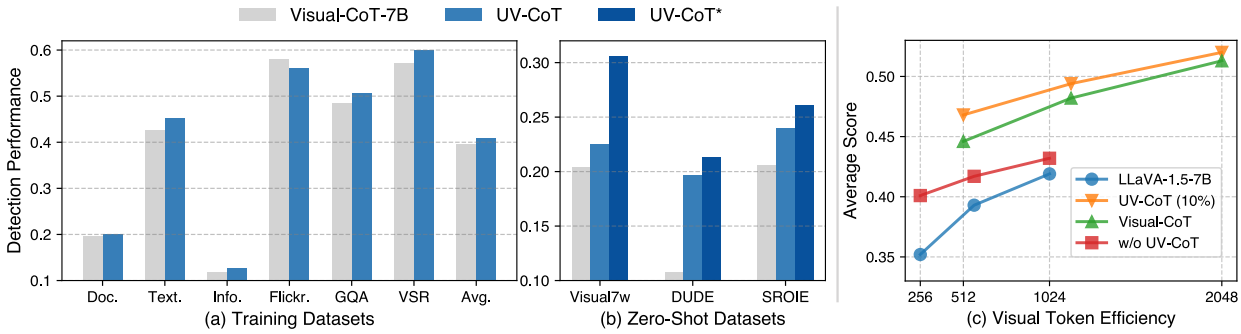

Figure 3. **(a&b) Bounding box evaluation** on (a) training datasets and (b) zero-shot datasets. Our `UV-CoT` performs better than Visual-CoT. **(c) Model performance** under varying visual token sizes. Our `UV-CoT` demonstrates better token efficiency.

Model	DocVQA	TextVQA	InfographicsVQA	Flickr30k	GQA	VSR	Average
OmniLMM-12B	0.254	0.578	0.172	0.621	0.509	0.523	0.443
`IF`: OmniLMM-12B + CoT	0.305	0.722	0.217	0.675	0.580	0.595	0.516
LLaVA-1.5-7B	0.198	0.507	0.131	0.539	0.480	0.504	0.393
`IF`: LLaVA-1.5-7B + CoT	0.245	0.577	0.149	0.606	0.530	0.533	0.440
`Tr`: `UV-CoT` (Evaluator: self-evaluated)	0.242	0.609	0.153	0.598	0.517	0.526	0.441
`Tr`: `UV-CoT` (Evaluator: OmniLMM-12B)	0.265	0.686	0.173	0.632	0.536	0.548	0.473

Table 5. Analysis of evaluator model. '+ CoT' refers to inference assisted by bounding boxes generated by `UV-CoT`. Self-evaluated denotes using target model as the evaluator during training. `IF` and `Tr` indicate experiments conducted on the inference and training process.

This underscores the importance of iterative learning in continuously aligning the preference data distribution with the model's evolving policy throughout the training process.

Response evaluation: 'w/o γ model' sets $\gamma = 0$ during response evaluation stage, ignoring the next response when generating preference scores. We observe a significant performance drop (8.8% on average), particularly on TextVQA (17.2%), highlighting the difficulty MLLMs face in directly evaluating bounding boxes. This underscores the necessity of incorporating next response in our evaluation method.

4.5. Bounding Box Evaluation

We compare the quality of bounding boxes learned from *supervised* and *unsupervised* strategies using GPT-4o as a scorer, on both training datasets (Fig. 3 (a)) and zero-shot datasets (Fig. 3 (b)). Our main observations are:

(1) Our `UV-CoT` outperforms Visual-CoT-7B, achieving higher scores in five of six datasets, supporting its superior performance in generating helpful bounding box.

(2) The bounding box quality is closely related to the final performance. Our model exhibits lower scores (below 0.210) for bounding box generation in DocVQA and InfographicsVQA, correlating with its reduced final scores in these datasets (below 0.290 in Tab. 1). It underscores the validity of evaluating bounding box quality through its impact on subsequent answers.

(3) Both `UV-CoT` and `UV-CoT`* outperform Visual-CoT-7B across all zero-shot datasets, which illustrates the strong generalization of our method in bounding box generation.

4.6. Insight of Evaluator Model

We further perform studies to better understand the role of evaluator model. Key findings from Tab. 5 are:

(1) We compare `UV-CoT` with its self-evaluated variant, where the evaluator model is the same as the target model (initialed with LLaVA-1.5-7B). Although the self-evaluated version exhibits a performance decrease of 3.2% compared to the original `UV-CoT`, it still outperforms LLaVA-1.5-7B (+4.8% on average) across all evaluated datasets and achieves performance comparable to the larger OmniLMM-12B model (-0.2% on average). This demonstrates that `UV-CoT` maintains robust performance even under self-evaluation, highlighting its efficiency without requiring larger model scales.

(2) For OmniLMM-12B and LLaVA-1.5-7B, we incorporate a CoT process using bounding boxes generated by `UV-CoT`. The CoT-enhanced versions significantly outperform their original counterparts, achieving average performance gains of 4.7% for LLaVA-1.5-7B and 7.4% for OmniLMM-12B. Remarkably, these models were not fine-tuned for the CoT process, indicating that the bounding box information alone substantially improves performance. This finding underscores that our evaluating process simplifies the task of generating complex spatial annotations,

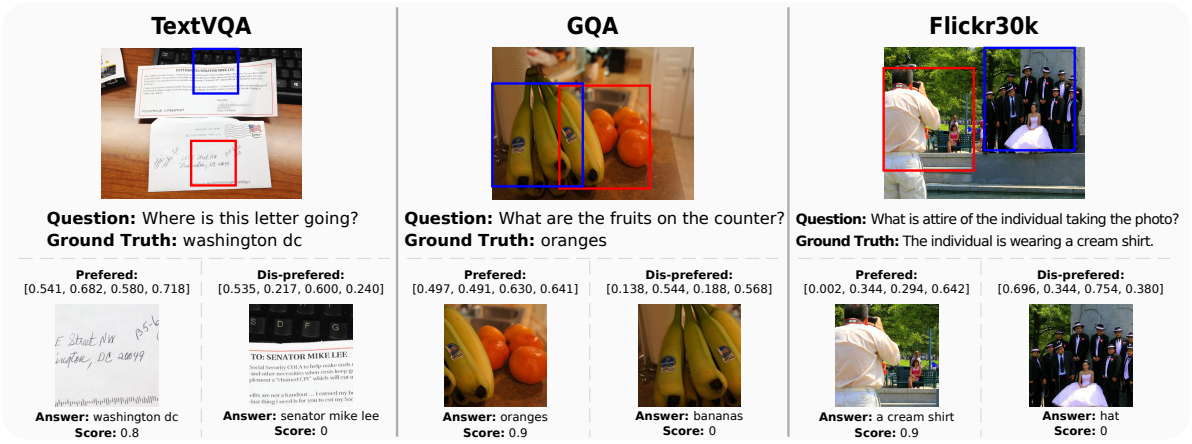

Figure 4. Visualization of preference data generated by Algorithm 1. Preferred BBoxes are in **red**. Dis-preferred BBoxes are in **blue**.

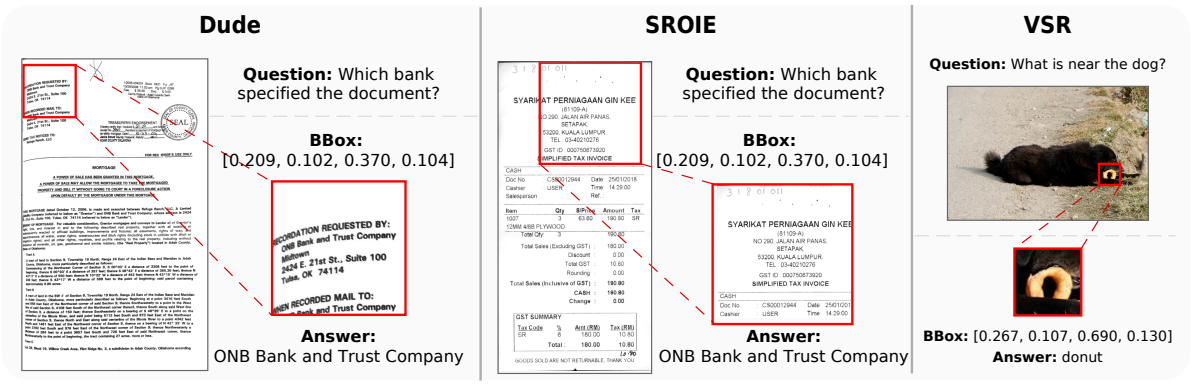

Figure 5. Visualization of our UV-CoT inference. Model-generated bounding boxes are shown in **red**.

enabling MLLMs to focus on evaluating final answers.

4.7. Other Detailed Analyses

Visual token efficiency. Compared to standard MLLM generation, image-level CoT doubles the number of visual tokens by processing additional local image regions. To evaluate token efficiency, *i.e.*, performance under the same visual token budget, we resize the input image to different resolutions ($224^2, 336^2$ and 448^2) and report the average score of different models in Fig. 3(c). Our key findings are:
(1) MLLMs with image-level reasoning (Visual-CoT-7B and our UV-CoT) demonstrate better token efficiency than the standard answer generation pipeline. *E.g.*, they achieve higher performance with 512 visual tokens than the standard pipeline does with 1024 tokens.
(2) Our UV-CoT consistently outperforms Visual-CoT-7B across all scales, achieving higher average scores. This highlights the token efficiency of our method.
Visualization. Fig. 4 visualizes some preference data from our UV-CoT inference process. Given different local regions and their corresponding answers, the evaluator MLLM assigns reasonable scores, validating the effectiveness of our automatic generation and labeling process. In Fig. 5, we present reasoning cases with model-generated bounding boxes overlaid. The precision of bounding box detection and the depth of understanding play a crucial role in determining the quality of generated answers.

5. Conclusion

In this work, we propose UV-CoT, a framework that enables image-level CoT reasoning in MLLMs via preference optimization. Unlike previous methods that rely on SFT needing large amounts of labeled data, our approach leverages unsupervised learning to refine the model's ability with image-level CoT using model-generated preference data (which are rough but useful). We address key challenges in preference data generation and effective optimization, ensuring a more adaptive and interpretable reasoning process. Extensive experiments demonstrate that UV-CoT achieves state-of-the-art performance, significantly improving visual comprehension in MLLMs on ten reasoning datasets. Our findings highlight the potential of preference learning as a scalable alternative to traditional SFT, enabling more robust and data-efficient multimodal reasoning.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-030). Thanks to Zichen Tian for his assistance with figure visualization throughout this work. The authors also thank the reviewers for their valuable comments and suggestions.

References

[1] Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024. 4, 11

[2] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 2024. 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[4] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 4

[5] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025. 2

[6] Herbert Aron David. *The method of paired comparisons*. London, 1963. 11

[7] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. 3

[8] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *NeurIPS*, 2023. 2

[9] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. In *NeurIPS*, 2024. 1

[10] Jinpeng Hu, Tengteng Dong, Luo Gang, Hui Ma, Peng Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang. Psycollm: Enhancing llm for psychological understanding and evaluation. *IEEE Transactions on Computational Social Systems*, 2024. 2

[11] Mingxin Huang, Yuliang Liu, Dingkang Liang, Lianwen Jin, and Xiang Bai. Mini-monkey: Alleviating the semantic sawtooth effect for lightweight mllms via complementary image pyramid. *arXiv preprint arXiv:2408.02034*, 2024. 1

[12] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *ICDAR*, 2019. 5, 12

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 5, 12

[14] Pengyue Jia, Yiding Liu, Xiaopeng Li, Xiangyu Zhao, Yuhao Wang, Yantong Du, Xiao Han, Xuetao Wei, Shuaiqiang Wang, and Dawei Yin. G3: an effective and adaptive framework for worldwide geolocalization using large multi-modality models. In *NeurIPS*, 2024. 2

[15] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *CVPR*, 2025. 2

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 12

[17] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. 1, 2

[18] Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024. 3

[19] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. 2023. 2

[20] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 5, 12

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 1, 5

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1, 5

[23] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):1–17, 2024. 2

[24] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 2

[25] Chris J. Maddison and Danny Tarlow. Gumbel machinery, 2017. Available online. 4, 11

[26] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021. 5, 12, 13

[27] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022. 5, 12, 13

[28] OpenAI. Chatgpt, 2023. Accessed: Mar. 4, 2025. 5

[29] Haowen Pan, Xiaozhi Wang, Yixin Cao, Zenglin Shi, Xun Yang, Juanzi Li, and Meng Wang. Precise localization of memories: A fine-grained neuron-level knowledge editing technique for llms. In *ICLR*, 2025. 2

[30] Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In *NeurIPS*, 2025. 3

[31] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 5, 12

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023. 2

[34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 4, 11

[35] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *NeurIPS*, 2025. 1, 2, 5, 12

[36] 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*, 2019. 5, 12

[37] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 3

[38] 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. 2

[39] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *ICCV*, 2023. 5, 12

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 1, 2

[41] Penghao Wu and Saining Xie. V?: Guided visual search as a core mechanism in multimodal llms. In *CVPR*, 2024. 5, 6, 12

[42] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 1, 2

[43] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. 1, 2

[44] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 5

[45] Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*, 2023. 1

[46] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, 2024. 3

[47] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 3, 4

[48] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Xiangyu Zhao, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. *arXiv preprint arXiv:2405.16789*, 2024. 2

[49] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 1, 2

[50] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 1, 2

[51] Sheng Zhou, Junbin Xiao, Qingyun Li, Yicong Li, Xun Yang, Dan Guo, Meng Wang, Tat-Seng Chua, and Angela Yao. Egotextvqa: Towards egocentric scene-text aware video question answering. In *CVPR*, 2025. 2

[52] Xingyu Zhu, Shuo Wang, Jinda Lu, Yanbin Hao, Haifeng Liu, and Xiangnan He. Boosting few-shot learning via attentive feature regularization. In *AAAI*, 2024. 2

[53] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting. *NeurIPS*, 2024.

[54] Xingyu Zhu, Beier Zhu, Yi Tan, Shuo Wang, Yanbin Hao, and Hanwang Zhang. Selective vision-language subspace projection for few-shot clip. In *ACMMM*, 2024. 2

[55] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, 2016. 5, 12

[56] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 2