A. Outline

We begin by presenting an overview of our Appendix.

- **Section B: Framework Datails.** We provide a detailed explanation on the connection between our UV-CoT and DPO.
- **Section C: Implement Details.** We describe the specifics of our dataset, and evaluation methodology.
- **Section D: Limitations.** We analysis the constraints and challenges of our approach.
- **Section E: Potential negative societal impacts.** We discuss possible negative consequences and ethical considerations associated with our work.

B. Connection between UV-CoT and DPO

B.1. Loss Function Formulation

To better captures the impact of key regions, we introduce a preference-weighted optimization approach. The loss function for UV-CoT is defined as follows:

$$\mathcal{L}_{sDPO}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)} - \big(g(s_w)-g(s_l)\big)\right)\right]. \tag{6}$$

where π_θ is the target policy model being optimized, π_{ref} is the frozen reference model, constraining deviations from the initial policy, $g:\mathbb{R}\to\mathbb{R}$ is a monotonically increasing function mapping preference scores s_w and s_l (for winning and losing responses y_w and y_l) into the logit space, β is a temperature parameter, $\mathcal{D}$ represents the dataset distribution over input-output pairs (x, y_w, y_l).

This formulation extends the DPO framework by incorporating preference differences $g(s_w)-g(s_l)$, which reflect the utility of key regions identified by UV-CoT.

B.2. DPO Background and Reparameterization

DPO reformulates reward model training as a policy optimization problem by reparameterizing the reward function from Proximal Policy Optimization (PPO) [34]:

$$r(x,y) = \beta\log\frac{\pi_\theta(y\mid x)}{\pi_{\text{ref}}(y\mid x)} + \beta\log Z(x), \tag{7}$$

where $Z(x)$ is the partition function. Substituting this into the Bradley-Terry preference model [6] yields:

$$\begin{aligned} p(y_w\succ y_l) &= \sigma\left(r(x,y_w)-r(x,y_l)\right)\\ &= \sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)}\right), \end{aligned} \tag{8}$$

where σ is the sigmoid function. Maximizing the log-likelihood of this preference model leads to the naive DPO loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \beta\log\frac{\pi_\theta(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)}\right)\right]. \tag{9}$$

B.3. Derivation with Gumbel Distribution

To incorporate preference-weighted optimization, we define $\Delta_r = g(s_w)-g(s_l)$ and introduce Gumbel-distributed random variables $R_w\sim\text{Gumbel}(r(x,y_w),1)$ and $R_l\sim\text{Gumbel}(r(x,y_l),1)$. The probability that the winning response is preferred, adjusted by the preference gap, is:

$$\begin{aligned} P(R_w-R_l>\Delta_r) &= \sigma\left(r(x,y_w)-r(x,y_l)-\Delta_r\right)\\ &= \sigma\left(\beta\log\frac{\pi^*(y_w\mid x)}{\pi_{\text{ref}}(y_w\mid x)} - \beta\log\frac{\pi^*(y_l\mid x)}{\pi_{\text{ref}}(y_l\mid x)} - \Delta_r\right). \end{aligned} \tag{10}$$

This result leverages the Gumbel-max trick [25], with a similar derivation found in ODTO [1].

B.4. Proof of Gumbel-Based Preference

We first prove the foundational probability $P(R_w-R_l>0)=\sigma(\Delta_{\hat{r}_\theta})$, where $\Delta_{\hat{r}_\theta}=r_\theta(x,y_w)-r_\theta(x,y_l)$.

Proof: Define the random variable $I = \arg\max_{l,w}\{R_l,R_w\}$. The goal is to show:

$$P(I=w) = \frac{\exp(\hat{r}_\theta(x,y_w))}{\exp(\hat{r}_\theta(x,y_w))+\exp(\hat{r}_\theta(x,y_l))}. \tag{11}$$

For notation simplicity, let $\hat{r}_w=\hat{r}_\theta(x,y_w)$, $\hat{r}_l=\hat{r}_\theta(x,y_l)$, and $g_{\hat{r}_w}\sim\text{Gumbel}(\hat{r}_w,1)$. Then:

$$P(I=w) = \mathbb{E}_{m\sim g_{\hat{r}_w}}\left[P(R_l<m)\right] \tag{12}$$

$$= \int_{-\infty}^{\infty}\exp(-m)\cdot\exp\left(-\exp(\hat{r}_l-m)\right)\cdot\exp(\hat{r}_w)\,dm, \tag{13}$$

where the integral accounts for the Gumbel CDF. Let $Z=\exp(\hat{r}_w)+\exp(\hat{r}_l)$. The expression simplifies to:

$$P(I=w) = \frac{\exp(\hat{r}_w)}{Z} = \frac{\exp(\hat{r}_\theta(x,y_w))}{\exp(\hat{r}_\theta(x,y_w))+\exp(\hat{r}_\theta(x,y_l))}, \tag{14}$$

proving Equation (11). Extending this to the preference gap Δ_r, we derive:

$$\begin{aligned} P(R_w-R_l>\Delta_r) &= 1-\mathcal{F}(\Delta_r)\\ &= \frac{1}{2}-\frac{1}{2}\tanh\left(\frac{\Delta_r-\Delta_{\hat{r}_\theta}}{2}\right)\\ &= \sigma(\Delta_{\hat{r}_\theta}-\Delta_r), \end{aligned} \tag{15}$$

Table 6. The details of the datasets, which spans five distinct domains and includes various source datasets.

Domain	Source Dataset	Size	Dataset Description
Text/Doc	TextVQA	16k	Images with text
	DocVQA	35k	Doc images
	DUDE	15k	Doc images
	SROIE	4k	Invoice images
Chart	InfographicsVQA	15k	Infographic images
General VQA	Visual7W	43k	Images
	Flickr30k	136k	Images
Relation Reasoning	VSR	3k	Images
	GQA	88k	Images
High-Resolution	V^* Bench	238	Images

where $\Delta_{\hat{r}_\theta} = r_\theta(x, y_w) - r_\theta(x, y_l)$. This completes the derivation of the UV-CoT loss in Equation (6).

C. Implement Details

C.1. Datasets

To generate diverse and comprehensive image-level Chain-of-Thought (CoT) data for training Multimodal Large Language Models (MLLMs), we select nine source datasets spanning four distinct domains: Text/Doc, General Visual Question Answering (VQA), Charts, and Relation Reasoning. These domains are chosen to ensure a broad representation of visual reasoning tasks, enabling the model to develop robust CoT capabilities across varied contexts.

Before performing preference optimization, we conduct Supervised Fine-Tuning (SFT) using 10% of the labeled Visual-CoT dataset, which corresponds to approximately 25k samples, as detailed in Tab. 6. This subset is chosen to balance computational efficiency with sufficient exposure to diverse reasoning patterns, resulting in a model we denote as `UV-CoT (10%)`. Following SFT, we perform preference optimization using a total of 249k preference data points, curated from the same nine datasets. The preference data is generated by ranking model outputs for each dataset, ensuring that the distribution across domains mirrors that of Visual-CoT [35]. Specifically, for each dataset, we maintain roughly the same proportion of preference pairs as in Visual-CoT (e.g., Text/Doc datasets contribute approximately 50% of the data, consistent with their representation in our dataset). This approach ensures that `UV-CoT` benefits from a balanced and comprehensive preference optimization process, enhancing its ability to prioritize key regions in visual reasoning tasks.

As shown in Tab. 6, we provide a detailed introduction to the datasets we used. For the Text/Doc domain, we include four datasets focusing on text recognition and comprehension in diverse document and image formats: DocVQA [26], TextVQA [36], DUDE [39], and SROIE [12]. These datasets provide rich text-based reasoning scenarios, such as extracting information from invoices (SROIE) or answering questions about document content (DocVQA), which are crucial for generating CoT data that involves step-by-step text interpretation.

In the General VQA domain, we select Flickr30k [31] and Visual7W [55]. These datasets are well-suited for general visual question answering, as they contain diverse images paired with questions that require understanding both visual content and textual prompts, facilitating the creation of CoT data for general reasoning tasks.

For the Charts domain, we use the InfographicsVQA dataset [27], which consists of high-resolution infographics. This dataset is particularly advantageous for training MLLMs to localize and interpret specific regions in charts, enabling the generation of CoT data that involves reasoning about data visualization elements such as legends, labels, and trends.

In the Relation Reasoning domain, we select VSR [20] and GQA [13]. These datasets are rich in spatial relational information among objects in images, making them ideal for constructing CoT data that focus on reasoning about object relationships, such as identifying relative positions or dependencies in a scene.

For high-resolution image reasoning, we use V^* Bench [41], which comprises 238 images from the SA-1B dataset [16] with an average resolution of 2246×1582.

C.2. Evaluation

We utilize GPT-4o to assess the performance of our model due to its superior reasoning capabilities and adopt an evaluation prompt to. The prompt template is like:

Evaluating Prompt Template

> *You are a helpful and precise assistant for checking the quality of the answer, you need to give a score to the model's answer by referring to the standard answer, based on the given question. The full score is 1 point and the minimum score is 0 points. Please output the score in the form 'score: <score >'. The evaluation criteria require that the closer the model's answer is to the standard answer, the higher the score.*

The meaning score, ranging from 0 to 1, reflects the semantic relevance of the model's responses to the given prompts.

D. Limitations

While our UV-CoT model demonstrates high performance across most evaluated datasets, it encounters challenges in accurately identifying anchor boxes on certain datasets, notably DocVQA [26] and InfographicsVQA [27]. These difficulties may arise due to the complex layouts, variable text densities, and noisy annotations prevalent in these datasets, which complicate the localization of relevant regions. In contrast, the ground truth (GT) boxes achieve exceptional performance on these datasets, suggesting that our model has significant untapped potential. Future research could explore advanced anchor box detection algorithms, such as incorporating adaptive thresholding or multi-scale feature extraction, to address these limitations and enhance the model's robustness across diverse visual domains.

E. Potential negative societal impacts

The potential negative societal impacts of our work align with those of other MLLMs and LLMs. While the development of UV-CoT and MLLMs advances AI capabilities, it also introduces several risks. These include heightened privacy concerns, the reinforcement of existing biases, the spread of misinformation, job displacement due to automation, and ethical challenges related to accountability, transparency, and informed consent. Addressing these issues requires responsible deployment, continuous monitoring, and the implementation of safeguards to mitigate unintended consequences.