Uncertainty-Aware Exploratory Direct Preference Optimization for Multimodal Large Language Models

Supplementary Material

A. The Implicit Advantage View of DPO

The learning objective in the KL-regularized RL setting, such as DPO, is to find a policy π_θ that maximizes the expected sum of rewards while penalizing its deviation from a fixed reference policy π_{ref}, which can be formally expressed as the maximization of the objective:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\left(r\left(s_t,a_t\right) - \beta D_{\text{KL}}\left(\pi\left(\cdot \mid s_t\right)\|\pi_{\text{ref}}\left(\cdot \mid s_t\right)\right)\right)\right], \tag{16}$$

where τ represents a trajectory, $r(s_t, a_t)$ is the reward function, and $\beta > 0$ is a temperature parameter that controls the strength of the KL-divergence penalty.

Connecting the trajectory-level optimization problem to single-step decision-making, the KL-regularized Bellman relations between the optimal state-value V^* and optimal action-value Q^* functions are as follows:

$$V^*(s) = \max_{\pi(\cdot|s)}\left\{\mathbb{E}_{a\sim\pi}\left[Q^*(s,a)\right] - \beta D_{\text{KL}}(\pi\|\pi_{\text{ref}})\right\}, \tag{17}$$

$$Q^*(s,a) = r(s,a) + \gamma\mathbb{E}_{s'}\left[V^*\left(s'\right)\right], \tag{18}$$

where s' denotes the subsequent state reached after state s receives token a. Eq. 17 and 18 defines the local objective at state s: to find a policy $\pi(\cdot|s)$ that maximizes the expected Q*-value while remaining close to the reference policy π_{ref}. The discount factor γ is typically set to 1 in large language model settings. Moreover, since the state transition in language modeling is deterministic, Eq. 18 can be simplified as $Q^*(s,a) = r(s,a) + V^*\left(s'\right)$.

From the perspective of dynamic programming, global optimality is guaranteed only if every state s satisfies the local optimality condition. Equivalently, violations at any single state imply sub-optimality at the global level. This principle formalizes the transition from the global optimization of $J(\pi)$ to the hierarchy of per-state optimizations. To solve the per-state optimization:

$$\max_{\pi(\cdot|s)}\left\{\mathbb{E}_{a\sim\pi}\left[Q^*(s,a)\right] - \beta D_{KL}(\pi\|\pi_{\text{ref}})\right\}, \tag{19}$$

subject to the constraint that $\pi(\cdot|s)$ is a valid probability distribution, *i.e.*:

$$\sum_a \pi(a|s) = 1. \tag{20}$$

The method of Lagrange multipliers can be used to solve this constrained problem:

$$\mathcal{L}(\pi,\eta) = \left(\sum_a \pi(a|s)Q^*(s,a) - \beta\sum_a \pi(a|s)\log\frac{\pi(a|s)}{\pi_{\text{ref}}(a|s)}\right) - \eta\left(\sum_a \pi(a|s) - 1\right). \tag{21}$$

To find the optimal policy, taking the partial derivative of $\mathcal{L}$ with respect to $\pi(a|s)$ and setting it to 0:

$$Q^*(s,a) - \beta\left(\log\frac{\pi^*(a|s)}{\pi_{\text{ref}}(a|s)} + 1\right) - \eta = 0. \tag{22}$$

Solving for $\pi^*(a|s)$ yields:

$$\pi^*(a|s) = \pi_{\text{ref}}(a|s)\exp\left(\frac{Q^*(s,a)}{\beta}\right)\exp\left(\frac{-\eta(s)-\beta}{\beta}\right). \tag{23}$$

The term $\exp((-\eta(s)-\beta)/\beta)$ is constant with respect to any token (action) a. We can thus absorb it into a state-dependent normalization term, $1/Z(s)$, which ensures that the policy sums to 1 over all tokens. This gives the closed-form expression for the optimal policy $\pi^*(a|s)$:

$$\pi^*(a|s) = \frac{1}{Z(s)}\pi_{\text{ref}}(a|s)\exp\left(\frac{Q^*(s,a)}{\beta}\right), \tag{24}$$

where the partition function $Z(s)$ is defined by the normalization constraint:

$$Z(s) = \sum_{a'}\pi_{\text{ref}}(a'|s)\exp(Q^*(s,a')/\beta). \tag{25}$$

From the derivation of the optimal policy, we can rearrange Eq. 24 to express $\log(\pi^*(a|s)/\pi_{\text{ref}}(a|s))$:

$$\log\frac{\pi^*(a|s)}{\pi_{\text{ref}}(a|s)} = \frac{Q^*(s,a)}{\beta} - \log Z(s). \tag{26}$$

By substituting Eq. 26 back into the Bellman equation for $V^*(s)$, the expression can be simplified to:

$$V^*(s) = \beta\log Z(s). \tag{27}$$

Finally, we can derive an explicit expression for the optimal advantage function, defined as $A^*(s,a) \triangleq Q^*(s,a) - V^*(s)$. By substituting $V^*(s) = \beta\log Z(s)$ into the right-hand side of Eq. 26, we arrive at the key relationship:

$$\beta\log\left(\frac{\pi^*(a|s)}{\pi_{\text{ref}}(a|s)}\right) = Q^*(s,a) - V^*(s) \triangleq A^*(s,a) \tag{28}$$

Algorithm 1 Uncertainty-aware Exploratory Direct Preference Optimization (UE-DPO)

Require: Dataset $\mathcal{D} = \{(v, x, y_w, y_l)\}$, policy model π_θ, reference model π_{ref}
Require: Hyperparameters: noise level ξ, quantile τ, intensity scale α, learning rate η
while not converged **do**
  Sample a batch $\{(v, x, y_w, y_l)\}$ from $\mathcal{D}$
  // Uncertainty Awareness (Sec. 4.1)
  Generate blurred image v' via diffusion noise: $v'(k) \leftarrow \sqrt{\bar{\xi}_k} \cdot v + \sqrt{1 - \bar{\xi}_k} \cdot \epsilon$
  Compute token logit variation $\Delta(a_t, s_t) \leftarrow \text{logit}_\theta(a_t|v, x, y_{<t}) - \text{logit}_\theta(a_t|v', x, y_{<t})$
  Compute token epistemic uncertainty $\text{u}(a_t, s_t) \leftarrow \text{logit}_\theta(\hat{a}_t(v')|v, x, y_{<t}) - \text{logit}_\theta(a_t|v, x, y_{<t})$
  // The Control of Exploration Intensity (Sec. 4.2)
  Preferred branch (y_w):
  Identify visually insensitive mask: $I_w \leftarrow 1\{\Delta(a_t, s_t) \le q_\tau(\Delta)\}$
  Compute exploration intensity: $\lambda_w(a_t, s_t) \leftarrow 1 + \alpha \cdot I_w \cdot \sigma(\text{Normalize}(\text{u}))$
  Dispreferred branch (y_l):
  Identify visually sensitive mask: $I_l \leftarrow 1\{\Delta(a_t, s_t) \ge q_{1-\tau}(\Delta)\}$
  Compute exploration intensity: $\lambda_l(a_t, s_t) \leftarrow 1 - \alpha \cdot I_l \cdot \sigma(\text{Normalize}(\text{u}))$
  // Training Objective (Sec. 4.3)
  Compute UE-DPO loss:

$$\mathcal{L}_{\text{UE-DPO}} \leftarrow -\mathbb{E}\left[\log\sigma\left(\beta\sum_t \log\frac{\pi_\theta(a_t^w|s_t^w)^{\lambda_w}}{\pi_{\text{ref}}(a_t^w|s_t^w)} - \beta\sum_t \log\frac{\pi_\theta(a_t^l|s_t^l)^{\lambda_l}}{\pi_{\text{ref}}(a_t^l|s_t^l)}\right)\right]$$

  Update $\theta \leftarrow \theta - \eta\nabla_\theta\mathcal{L}_{\text{UE-DPO}}$
end while

The quantity $A^*(s, a)$ measures the improvement in expected performance of token (action) a over the reference baseline at state s, that is, how much better token a performs compared to the baseline performance under reference policy π_{ref}. This relationship is central to the DPO framework, where the advantage is parameterized as an implicit immediate reward, and policy improvement is conducted through the supervised alignment on preference pairs.

DPO parameterizes this advantage as an implicit immediate reward and performs preference learning by fitting pairwise preference data $(x, y_w, y_l) \sim \mathcal{D}$:

$$L_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log\sigma\left(\beta\sum_{t=0}^{T_w}\log\frac{\pi_\theta(a_t^w|s_t)}{\pi_{\text{ref}}(a_t^w|s_t)} - \beta\sum_{t=0}^{T_l}\log\frac{\pi_\theta(a_t^l|s_t)}{\pi_{\text{ref}}(a_t^l|s_t)}\right). \tag{29}$$

B. Algorithmic Outline

We summarize our method in Alg. 1.

C. The Influence of Exploration Intensity on the Optimal Policy

The integration of the uncertainty-aware exploration can be interpreted as introducing a token-wise entropy regularization factor $\lambda(s, a)$ into the standard reverse-KL regularized RL objective in Eq. 16: (Note that, for notational simplicity, we henceforth denote $\lambda(s, a)$ and $\lambda(s, a')$ as λ and λ', respectively, in the subsequent equations.)

$$\begin{aligned}&\mathbb{E}_{a\sim\pi}\left[\log\pi(a|s) - \log\pi_{\text{ref}}(a|s)\right]\\ &\rightarrow \mathbb{E}_{a\sim\pi}\left[\lambda\log\pi(a|s) - \log\pi_{\text{ref}}(a|s)\right].\end{aligned} \tag{30}$$

This factor dynamically modulates the KL regularization strength in accordance with the model's epistemic uncertainty across tokens, leading to the following reformulated optimal value function:

$$\max_{\pi(\cdot|s)}\{\mathbb{E}_{a\sim\pi}\left[Q^*(s, a) - \beta(\lambda\log\pi(a|s) - \log\pi_{\text{ref}}(a|s))\right]\}. \tag{31}$$

Subsequently, the Lagrangian function is constructed as follows:

$$\begin{aligned}L(\pi, \eta) = &\sum_a \pi(a|s)[Q^*(s, a) - \beta(\lambda\log\pi(a|s) - \log\pi_{\text{ref}}(a|s))]\\ &+ \eta\left(\sum_a \pi(a|s) - 1\right).\end{aligned} \tag{32}$$

Taking the partial derivative with respect to $\pi(a|s)$ and equating it to zero yields:

$$\frac{\partial L}{\partial\pi(a|s)} = Q^* - \beta\lambda(1 + \log\pi) + \beta\log\pi_{\text{ref}} + \eta = 0 \tag{33}$$

The optimal policy can be derived as:

$$\pi^*(a|s) = \pi_{\text{ref}}(a|s)^{1/\lambda} \exp\left(\frac{Q^*(s,a)+\eta(s)}{\beta\lambda}\right)/Z(s), \tag{34}$$

where $Z(s)$ is partition function, and η is Lagrange multiplier. $\eta(s)$ indicates that the multiplier η depends only on the state s and is independent of token a.

The partial derivative of $\log \pi^*$ with respect to $\lambda(s,a)$ is:

$$\frac{\partial}{\partial\lambda}\log\pi^* \propto -\frac{\log\left(\pi_{\text{ref}}(a|s)\right)}{\lambda^2} - \frac{\left(Q^*(s,a)+\eta(s)\right)}{(\beta\lambda^2)}, \tag{35}$$

where **the first term** $-\log\left(\pi_{\text{ref}}(a|s)\right)/\lambda^2$ is always positive, since the prior probability π_{ref} takes value in the interval between 0 and 1. This term, for tokens a with a small π_{ref}, exerts a relatively strong positive driving force, thereby endowing the factor $\pi_{\text{ref}}(a|s)^{1/\lambda}$ in Eq. 34 with a prior-corrective exploratory effect. **The second term** $-\left(Q^*(s,a)+\eta(s)\right)/(\beta\lambda^2)$ is negative for the valuable tokens of interest, that is, those with large $Q^*(s,a)$. This term exerts a pulling force that attempts to reduce their probability through reducing the $\exp(\cdot)$ factor in Eq. 34, such that the $\exp(\cdot)$ factor contributes to a smoothing effect on π^*. This smoothing effect questions the authority of high-Q^* tokens and induces a conservative, broad exploratory behavior. **The overall exploration mechanism** of λ is precisely the outcome of the dynamic interplay between these two effects. The ultimate sign of $\frac{\partial}{\partial\lambda}\log\pi^*$ depends on the relative magnitudes of $-\log\left(\pi_{\text{ref}}(a|s)\right)/\lambda^2$ and $\left(Q^*(s,a)+\eta(s)\right)/(\beta\lambda^2)$.

For the "forgotten correct tokens" of interest in our UE-DPO framework, namely, those that are correct but have been omitted by the reference policy, with small π_{ref} yet large Q^*, the term $-\log\left(\pi_{\text{ref}}(a|s)\right)/\lambda^2$ dominates $\left(Q^*(s,a)+\eta(s)\right)/(\beta\lambda^2)$, that is, $\pi_{\text{ref}}(a|s) < \exp\left(-(Q^*(s,a)+\eta(s))/\beta\right)$. In this case, since the derivative is positive, increasing λ raises the probability of these tokens in the optimized policy π^*, embodying an error-correcting mode of exploration.

D. Generalized Exploratory Advantage

It follows that the optimal policy $\pi^*(a|s)$ (in Eq. 34) satisfies:

$$\beta\log\frac{\pi^*(a|s)^\lambda}{\pi_{\text{ref}}(a|s)} = Q^*(s,a) - \beta\lambda(s,a) + \eta(s), \tag{36}$$

Recall that the optimal value function $V^*(s)$ in Eq. 31:

$$V^* = \mathbb{E}_{a\sim\pi^*}\left[Q^*(s,a) - \beta(\lambda\log\pi^*(a|s) - \log\pi_{\text{ref}}(a|s))\right]. \tag{37}$$

Substituting Eq. 36 into Eq. 37, we have:

$$V^*(s) = \beta\mathbb{E}_{a'\sim\pi^*}\left[\lambda(s,a')\right] - \eta(s). \tag{38}$$

Then, substituting Eq. 38 into Eq. 36, we obtain:

$$\begin{aligned}&\beta\log\frac{\pi^*(a|s)^\lambda}{\pi_{\text{ref}}(a|s)}\\ &= Q^*(s,a) - V^*(s) - \beta(\lambda - \mathbb{E}_{a'\sim\pi^*}\left[\lambda'\right])\\ &\triangleq A_{\text{e}}^*(s,a),\end{aligned} \tag{39}$$

which we refer to as the generalized exploratory advantage A_{e}^*. This generalized advantage A_{e}^* consists of two components: the **standard advantage function** $Q^*(s,a) - V^*(s)$ and a novel **exploration cost advantage** $-\beta(\lambda - \mathbb{E}_{a'\sim\pi^*}\left[\lambda'\right])$. If $\lambda(a,s)$ is higher than the average cost $\mathbb{E}_{a'\sim\pi^*}\left[\lambda'\right]$, *i.e.*, $\lambda - \mathbb{E}_{a'\sim\pi^*}\left[\lambda'\right] > 0$, which implies that the token a receives a stronger exploration cost penalty, thereby reducing its total advantage $A_{\text{e}}^*(s,a)$. As the intensity $\lambda(s,a)$ diminishes in preference learning, the exploratory advantage A_{e}^* increases.

By substituting this generalized exploratory advantage A_{e}^* into DPO framework, our proposed UE-DPO method can fit the pair-wise preference ordering while effectively mitigating the under-cognition of visual information through the λ-weighted gradients for active exploration, as discussed in the main paper.