

# Reg-DPO 深度解析： 从 DPO 的结构性缺陷到 SFT 正则化的完整推导

基于 “Reg-DPO: SFT-Regularized Direct Preference Optimization with GT-Pair for Improving Video Generation” 论文的详细分析

## Contents

<b>1 引言与问题背景</b>	<b>3</b>
<b>2 前置知识：从 RLHF 到 DPO 的推导</b>	<b>3</b>
2.1 PPO 目标函数 . . . . .	3
2.2 求解最优策略 . . . . .	3
2.3 隐式奖励函数 . . . . .	4
2.4 推导 DPO 目标函数 . . . . .	4
<b>3 扩散模型中的 DPO</b>	<b>5</b>
3.1 从语言模型到扩散模型的对应关系 . . . . .	5
3.2 扩散模型 DPO 损失函数 (Eq.2) . . . . .	5
3.3 优化方向分析 . . . . .	6
<b>4 DPO 梯度的完整推导 (从 Eq.2 到 Eq.3)</b>	<b>6</b>
4.1 第一步：对 $\log \sigma$ 求导 (链式法则外层) . . . . .	6
4.2 第二步：计算 $\nabla_{\theta} S$ . . . . .	7
4.3 第三步：计算单个平方范数项的梯度 . . . . .	7
4.4 第四步：代入正负样本 . . . . .	8
4.5 第五步：合并得到最终梯度公式 (Eq.3) . . . . .	8
<b>5 DPO 梯度比率 (DGR) 及其衰减机制</b>	<b>8</b>
5.1 DGR 的定义 (Eq.4) . . . . .	8
5.2 DGR 为什么会迅速衰减 . . . . .	9
5.2.1 训练目标决定了 $s(\theta)$ 必须变负 . . . . .	9
5.2.2 $s(\theta)$ 变负时 DGR 的变化 . . . . .	9
5.2.3 自毁循环 . . . . .	9
5.2.4 $\beta$ 越大问题越严重 . . . . .	9
<b>6 DPO 的结构性缺陷：缺乏对样本分布的直接监督</b>	<b>10</b>
6.1 问题诊断 . . . . .	10
6.2 理想路径 vs “作弊” 路径 . . . . .	10
6.3 高维空间中 “作弊” 路径为何被优先选择 . . . . .	10
<b>7 Reg-DPO：通过 SFT 正则化稳定 DPO</b>	<b>11</b>
7.1 设计思路：从梯度缺陷到最小修补 . . . . .	11
7.1.1 第一步：诊断梯度缺陷 . . . . .	11
7.1.2 第二步：最小修补——给正样本梯度加常数 . . . . .	11
7.2 从修改后的梯度反推损失函数 . . . . .	11
7.3 设计思路总结 . . . . .	12

<b>8 SFT 正则化如何防止过拟合与灾难性遗忘</b>	<b>12</b>
8.1 机制一：为正样本提供持续的梯度信号	12
8.2 机制二：间接约束负样本分布	13
8.3 机制三：控制分布偏移幅度	13
<b>9 与标准 DPO 中参考模型 KL 约束的对比</b>	<b>13</b>
9.1 约束层级不同	13
9.2 约束对象不同	13
9.3 失效模式不同	13
9.4 作用时机不同	14
<b>10 动态权重的选择：<math>r = \text{DGR}</math></b>	<b>14</b>
<b>11 全文公式索引与总结</b>	<b>15</b>

# 1 引言与问题背景

在视频生成领域，直接偏好优化（Direct Preference Optimization, DPO）因其简洁高效而成为提升生成质量的有力工具。DPO 直接应用于视频生成任务时，面临三大核心挑战：

- (1) **数据构建成本高**: 视频偏好对的标注需要多名专业标注员对每段视频进行多维度评分, 耗时且昂贵。
  - (2) **训练不稳定**: DPO 在视频的高维时空依赖场景下容易快速收敛并发生分布偏移, 导致模型坍塌。
  - (3) **显存消耗巨大**: 视频生成模型参数常超过 10B, DPO 训练还需要冻结的参考模型和成对数据, 显存需

Reg-DPO 框架针对这三个挑战分别提出了 GT-Pair（数据构建）、SFT 正则化（算法稳定性）和内存优化。DPO 在扩散模型中的损失函数和梯度公式，深入分析其结构性缺陷，并展示 Reg-DPO 如何通过引入 SFT 正则化来解决这些问题。

## 2 前置知识：从 RLHF 到 DPO 的推导

## 2.1 PPO 目标函数

DPO 源自 RLHF (Reinforcement Learning from Human Feedback) 框架。PPO (Proximal Policy Optimization) 的目标是最大化期望奖励，同时约束当前策略  $\pi_\theta$  不偏离参考策略  $\pi_{\text{ref}}$  太远：

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(y|x)} [r_\phi(x, y)] - \beta D_{\text{KL}}[\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)] \quad (1)$$

其中：

- $r_\phi(x, y)$ : 奖励函数, 评估  $(x, y)$  的质量;
  - $\beta > 0$ : 控制奖励最大化与分布稳定性之间的权衡;
  - $D_{\text{KL}}$ : KL 散度, 衡量两个分布的差异。

## 2.2 求解最优策略

下面对式(1)进行求解。首先展开 KL 散度：

$$\begin{aligned}
& \max_{\pi_\theta} \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta} [r_\phi(x, y)] - \beta \sum_{x,y} \pi_\theta(y|x) \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \\
&= \max_{\pi_\theta} \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta} [r_\phi(x, y)] - \beta \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta} \left[ \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\
&= \min_{\pi_\theta} \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta} \left[ \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r_\phi(x, y) \right]
\end{aligned} \tag{2}$$

为了将其转化为 KL 散度的标准形式，定义归一化常数：

$$Z(x) = \sum_y \pi_{\text{ref}}(y|x) e^{\frac{1}{\beta} r_\phi(x,y)} \quad (3)$$

以及最优策略：

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) e^{\frac{1}{\beta} r_\phi(x,y)} \quad (4)$$

将  $\pi^*(y|x)$  代入式 (2):

$$\begin{aligned}
& \min_{\pi_\theta} \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta} \left[ \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r_\phi(x, y) \right] \\
&= \min_{\pi_\theta} \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta} \left[ \log \frac{\pi_\theta(y|x)}{\frac{1}{Z(x)} \pi_{\text{ref}}(y|x) e^{\frac{1}{\beta} r_\phi(x, y)}} - \log Z(x) \right] \\
&= \min_{\pi_\theta} \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta} \left[ \log \frac{\pi_\theta(y|x)}{\pi^*(y|x)} - \log Z(x) \right] \\
&= \min_{\pi_\theta} \mathbb{E}_x [D_{\text{KL}}(\pi_\theta(y|x) \| \pi^*(y|x)) - \log Z(x)]
\end{aligned} \tag{5}$$

由于  $\log Z(x)$  不依赖  $\theta$ , 且 KL 散度非负, 当且仅当  $\pi_\theta = \pi^*$  时取得最小值。因此最优策略为:

$$\boxed{\pi_\theta(y|x) = \pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) e^{\frac{1}{\beta} r_\phi(x, y)}} \tag{6}$$

### 2.3 隐式奖励函数

从式 (6) 可以反解出奖励函数与策略之间的隐式关系。对两边取对数:

$$\log \pi_\theta(y|x) = -\log Z(x) + \log \pi_{\text{ref}}(y|x) + \frac{1}{\beta} r_\phi(x, y) \tag{7}$$

整理得到:

$$\boxed{r_\phi(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)} \tag{8}$$

这一等式揭示了一个关键洞察: 策略本身隐式地编码了奖励信号。

### 2.4 推导 DPO 目标函数

奖励模型的目标是最大化正样本与负样本之间的奖励差距:

$$\mathcal{L}_{\text{RM}} = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))] \tag{9}$$

其中  $\sigma(x) = \frac{1}{1+e^{-x}}$  是 sigmoid 函数,  $y_w$  是偏好的正样本,  $y_l$  是负样本。

将式 (8) 代入式 (9), 计算奖励差:

$$\begin{aligned}
r_\phi(x, y_w) - r_\phi(x, y_l) &= \left( \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} + \beta \log Z(x) \right) - \left( \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} + \beta \log Z(x) \right) \\
&= \beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}
\end{aligned} \tag{10}$$

注意  $\beta \log Z(x)$  项在相减时被消去。代入式 (9) 得到 DPO 的目标函数:

$$\boxed{\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(\beta \cdot s(\theta))]} \tag{11}$$

其中：

$$s(\theta) = \underbrace{[\log \pi_\theta(y_w|x) - \log \pi_\theta(y_l|x)]}_{\Delta_\theta(y_w, y_l)} - \underbrace{[\log \pi_{\text{ref}}(y_w|x) - \log \pi_{\text{ref}}(y_l|x)]}_{\Delta_{\text{ref}}(y_w, y_l)} \quad (12)$$

### DPO 的核心思想

DPO 将奖励模型训练和策略优化合二为一，直接从偏好对  $(y_w, y_l)$  中学习，无需显式的奖励模型。 $s(\theta)$  衡量的是当前模型相比参考模型在区分正负样本能力上的改进程度。

## 3 扩散模型中的 DPO

### 3.1 从语言模型到扩散模型的对应关系

在语言模型中， $\pi_\theta(y|x)$  直接表示生成概率。在扩散模型（如 DDPM、Rectified Flow）中，概率通过预测误差

$$\text{概率高} \iff \text{预测误差小} \iff \|y - y_\theta(x_t, t)\|^2 \text{ 小} \quad (13)$$

具体而言：

- 在 DDPM 中， $y$  是噪声目标， $y_\theta(x_t, t)$  是模型预测的噪声；
- 在 Rectified Flow 中， $y$  是向量场目标， $y_\theta(x_t, t)$  是模型预测的向量场；
- $x_t$  是时间步  $t$  处的中间噪声状态。

因此对应关系为：

$$\log \pi_\theta(y|x) \text{ (高 = 概率大)} \iff \|y - y_\theta(x_t, t)\|^2 \text{ (低 = 概率大)} \quad (14)$$

注意方向是相反的：语言模型中对数概率越大越好，扩散模型中预测误差越小越好。

### 3.2 扩散模型 DPO 损失函数 (Eq.2)

将上述对应关系代入式 (11)，同时注意方向取反 ( $\sigma(\cdot)$  内加负号)，得到扩散模型版本的 DPO 损失：

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(c, x_w, x_l) \sim \mathcal{D}} [\log \sigma(-\beta \cdot s(\theta))] \quad (15)$$

其中  $s(\theta)$  定义为：

$$s(\theta) = \underbrace{\left( \|y^w - y_\theta(x_t^w, t)\|^2 - \|y^l - y_\theta(x_t^l, t)\|^2 \right)}_{\Delta_\theta(x^w, x^l)} - \underbrace{\left( \|y^w - y_{\text{ref}}(x_t^w, t)\|^2 - \|y^l - y_{\text{ref}}(x_t^l, t)\|^2 \right)}_{\Delta_{\text{ref}}(x^w, x^l)} \quad (16)$$

各符号含义如下：

- $c$ : 条件输入（文本提示、参考图像等）；
- $x^w, x^l$ : 正样本和负样本对应的数据；
- $x_t^w, x_t^l$ : 时间步  $t$  处的中间噪声状态；
- $y^w, y^l$ : 正样本和负样本对应的目标输出（噪声或向量场）；
- $y_\theta$ : 当前模型的预测； $y_{\text{ref}}$ : 冻结参考模型的预测。

## Win Gap 与 Lose Gap

$s(\theta)$  也可以分解为两个有直观含义的量：

**Win Gap** (正样本改进) :

$$\text{Win Gap} = \|y^w - y_\theta(x_t^w, t)\|^2 - \|y^w - y_{\text{ref}}(x_t^w, t)\|^2 \quad (17)$$

理想情况下为负值，表示当前模型在正样本上预测得比参考模型更准确。

**Lose Gap** (负样本变化) :

$$\text{Lose Gap} = \|y^l - y_\theta(x_t^l, t)\|^2 - \|y^l - y_{\text{ref}}(x_t^l, t)\|^2 \quad (18)$$

理想情况下为正值，表示当前模型在负样本上的预测精度下降（隐式降低其生成概率）。

两者的关系为  $s(\theta) = \text{Win Gap} - \text{Lose Gap}$ 。

### 3.3 优化方向分析

要最小化  $\mathcal{L}_{\text{DPO}}$ ，需要让  $\log \sigma(-\beta \cdot s(\theta))$  尽可能大。

由于  $\sigma$  是单调递增函数， $\log$  也是单调递增函数，所以需要  $-\beta \cdot s(\theta)$  尽可能大。又因为  $\beta > 0$ ，这等价于  $s(\theta)$  尽可能小（更负）。

而  $\Delta_{\text{ref}}$  是冻结的常数，所以：

$$\text{最小化 } \mathcal{L}_{\text{DPO}} \iff \text{最小化 } s(\theta) \iff \text{最小化 } \Delta_\theta(x^w, x^l) \quad (19)$$

其中：

$$\Delta_\theta(x^w, x^l) = \underbrace{\|y^w - y_\theta(x_t^w, t)\|^2}_A - \underbrace{\|y^l - y_\theta(x_t^l, t)\|^2}_B \quad (20)$$

最小化  $A - B$  意味着让  $A$  尽可能小、 $B$  尽可能大，即模型应更精确地预测正样本、同时降低对负样本的预测。

## 4 DPO 梯度的完整推导（从 Eq.2 到 Eq.3）

本节逐步推导 DPO 损失函数关于参数  $\theta$  的梯度。

### 4.1 第一步：对 $\log \sigma$ 求导（链式法则外层）

定义  $S = -\beta \cdot s(\theta)$ ，则  $\mathcal{L}_{\text{DPO}} = -\mathbb{E}[\log \sigma(S)]$ 。

对  $\theta$  求梯度：

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\mathbb{E}[\nabla_\theta \log \sigma(S)] \quad (21)$$

对  $\log \sigma(S)$  应用链式法则：

$$\nabla_\theta \log \sigma(S) = \frac{1}{\sigma(S)} \cdot \sigma'(S) \cdot \nabla_\theta S \quad (22)$$

利用 sigmoid 函数的导数性质。sigmoid 函数定义为：

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (23)$$

其导数为：

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)) \quad (24)$$

验证:

$$\begin{aligned}\sigma'(x) &= \frac{d}{dx} \left( \frac{1}{1+e^{-x}} \right) = \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} = \sigma(x) \cdot \frac{1+e^{-x}-1}{1+e^{-x}} = \sigma(x)(1-\sigma(x)) \quad \checkmark\end{aligned}\quad (25)$$

将  $\sigma'(S) = \sigma(S)(1-\sigma(S))$  代入:

$$\nabla_{\theta} \log \sigma(S) = \frac{1}{\sigma(S)} \cdot \sigma(S)(1-\sigma(S)) \cdot \nabla_{\theta} S = (1-\sigma(S)) \cdot \nabla_{\theta} S \quad (26)$$

因此:

$$\boxed{\nabla_{\theta} \mathcal{L}_{\text{DPO}} = -\mathbb{E}[(1-\sigma(S)) \cdot \nabla_{\theta} S]} \quad (27)$$

## 4.2 第二步：计算 $\nabla_{\theta} S$

由  $S = -\beta \cdot s(\theta)$ , 所以  $\nabla_{\theta} S = -\beta \cdot \nabla_{\theta} s(\theta)$ 。

展开  $s(\theta)$  (式 (16)) :

$$s(\theta) = \Delta_{\theta}(x^w, x^l) - \Delta_{\text{ref}}(x^w, x^l) \quad (28)$$

由于参考模型  $y_{\text{ref}}$  是冻结的, 其输出不依赖于  $\theta$ , 因此:

$$\nabla_{\theta} \Delta_{\text{ref}}(x^w, x^l) = 0 \quad (29)$$

所以:

$$\nabla_{\theta} s(\theta) = \nabla_{\theta} \Delta_{\theta}(x^w, x^l) = \nabla_{\theta} \|y^w - y_{\theta}(x_t^w, t)\|^2 - \nabla_{\theta} \|y^l - y_{\theta}(x_t^l, t)\|^2 \quad (30)$$

## 4.3 第三步：计算单个平方范数项的梯度

对于一般形式  $\|y - y_{\theta}(x_t, t)\|^2$ , 将其展开为向量内积:

$$\|y - y_{\theta}(x_t, t)\|^2 = (y - y_{\theta}(x_t, t))^{\top} (y - y_{\theta}(x_t, t)) \quad (31)$$

定义  $f(\theta) = y - y_{\theta}(x_t, t)$ , 则需要计算  $\nabla_{\theta} \|f(\theta)\|^2 = \nabla_{\theta} [f(\theta)^{\top} f(\theta)]$ 。  
利用向量函数的链式法则。设  $g = f^{\top} f$  (标量), 则:

$$\nabla_{\theta} g = \nabla_{\theta} (f^{\top} f) = 2f(\theta)^{\top} \cdot \nabla_{\theta} f(\theta) \quad (32)$$

推导细节: 对  $f^{\top} f = \sum_i f_i^2$  求导, 得  $\frac{\partial}{\partial \theta_j} (f^{\top} f) = 2 \sum_i f_i \frac{\partial f_i}{\partial \theta_j} = 2f^{\top} \frac{\partial f}{\partial \theta_j}$ 。

现在计算  $\nabla_{\theta} f(\theta)$ :

$$\nabla_{\theta} f(\theta) = \nabla_{\theta} (y - y_{\theta}(x_t, t)) = 0 - \nabla_{\theta} y_{\theta}(x_t, t) = -\nabla_{\theta} y_{\theta}(x_t, t) \quad (33)$$

这里  $y$  是固定的目标值 (不依赖  $\theta$ ), 所以其梯度为零。

代入得:

$$\begin{aligned}\nabla_{\theta} \|y - y_{\theta}(x_t, t)\|^2 &= 2(y - y_{\theta}(x_t, t))^{\top} \cdot (-\nabla_{\theta} y_{\theta}(x_t, t)) \\ &= -2(y - y_{\theta}(x_t, t))^{\top} \nabla_{\theta} y_{\theta}(x_t, t)\end{aligned}\quad (34)$$

$$\boxed{\nabla_{\theta} \|y - y_{\theta}(x_t, t)\|^2 = -2(y - y_{\theta}(x_t, t))^{\top} \nabla_{\theta} y_{\theta}(x_t, t)} \quad (35)$$

#### 4.4 第四步：代入正负样本

将式 (35) 分别应用于正样本和负样本：

正样本项：

$$\nabla_{\theta} \|y^w - y_{\theta}(x_t^w, t)\|^2 = -2(y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) \quad (36)$$

负样本项：

$$\nabla_{\theta} \|y^l - y_{\theta}(x_t^l, t)\|^2 = -2(y^l - y_{\theta}(x_t^l, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^l, t) \quad (37)$$

将式 (36) 和式 (37) 代入式 (30)：

$$\nabla_{\theta} s(\theta) = -2(y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) + 2(y^l - y_{\theta}(x_t^l, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^l, t) \quad (38)$$

进而：

$$\begin{aligned} \nabla_{\theta} S &= -\beta \cdot \nabla_{\theta} s(\theta) \\ &= 2\beta \left[ (y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) - (y^l - y_{\theta}(x_t^l, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^l, t) \right] \end{aligned} \quad (39)$$

#### 4.5 第五步：合并得到最终梯度公式 (Eq.3)

将式 (39) 代入式 (27)：

$$\boxed{\nabla_{\theta} \mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[ 2\beta(1 - \sigma(S)) \cdot \left( (y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) - (y^l - y_{\theta}(x_t^l, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^l, t) \right) \right]} \quad (40)$$

其中  $S = -\beta \cdot s(\theta)$ 。

**推导链路总结：**

$$\mathcal{L}_{\text{DPO}} \xrightarrow[\text{第1步}]{\frac{d}{d\theta} \log \sigma} (1 - \sigma(S)) \cdot \nabla_{\theta} S \xrightarrow[\text{第2步}]{S = -\beta s(\theta)} -\beta \cdot \nabla_{\theta} s(\theta) \xrightarrow[\text{第3-4步}]{\frac{d}{d\theta} \|y - y_{\theta}\|^2} \text{式 (40)}$$

### 5 DPO 梯度比率 (DGR) 及其衰减机制

#### 5.1 DGR 的定义 (Eq.4)

从式 (40) 中提取公共系数，定义**DPO 梯度比率** (DPO Gradient Ratio, DGR) :

$$\boxed{\text{DGR} = \beta(1 - \sigma(S)) = \beta \left( 1 - \frac{1}{1 + e^{\beta s(\theta)}} \right) = \frac{\beta}{1 + e^{-\beta s(\theta)}}} \quad (41)$$

利用 DGR，梯度 (40) 可以简写为：

$$\nabla_{\theta} \mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[ 2 \text{DGR} \cdot \left( (y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) - (y^l - y_{\theta}(x_t^l, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^l, t) \right) \right] \quad (42)$$

DGR 的关键性质：

- DGR 的上界为  $\beta$  (当  $s(\theta) \rightarrow +\infty$  时取到)；
- DGR 的下界趋近于 0 (当  $s(\theta) \rightarrow -\infty$  时取到)；
- 正样本和负样本的梯度项共享相同的系数 2 DGR。

## 5.2 DGR 为什么会迅速衰减

### 5.2.1 训练目标决定了 $s(\theta)$ 必须变负

从式(15), 最小化  $\mathcal{L}_{\text{DPO}} = -\mathbb{E}[\log \sigma(-\beta \cdot s(\theta))]$ 。

函数  $\sigma$  单调递增  $\Rightarrow \log \sigma$  单调递增  $\Rightarrow$  前面有负号。

因此要让损失减小, 需要  $-\beta \cdot s(\theta)$  增大, 即  $s(\theta)$  减小 (更负)。

$$\text{训练成功} \iff s(\theta) \text{ 不断变负} \quad (43)$$

### 5.2.2 $s(\theta)$ 变负时 DGR 的变化

将 DGR 的等价形式  $\text{DGR} = \frac{\beta}{1+e^{-\beta s(\theta)}}$  代入具体数值。取论文默认值  $\beta = 1000$ :

Table 1:  $s(\theta)$  与 DGR 的对应关系 ( $\beta = 1000$ )

$s(\theta)$	$\beta s(\theta)$	$e^{-\beta s(\theta)}$	$\text{DGR} = \frac{1000}{1+e^{-\beta s(\theta)}}$
0 (初始)	0	1	500.0
-0.001	-1	$\approx 2.72$	$\approx 269$
-0.005	-5	$\approx 148$	$\approx 6.7$
-0.01	-10	$\approx 22026$	$\approx 0.045$
-0.02	-20	$\approx 4.9 \times 10^8$	$\approx 0$

#### 关键观察

$s(\theta)$  仅仅从 0 变到 -0.01——一个极其微小的变化——DGR 就从 500 暴跌到 0.045, 衰减了超过 10000 倍。这是因为指数函数  $e^{-\beta s(\theta)}$  在  $\beta$  很大时具有极强的放大效应。

### 5.2.3 自毁循环

训练开始时,  $\theta$  从参考模型初始化, 因此  $y_\theta \approx y_{\text{ref}}$ ,  $\Delta_\theta \approx \Delta_{\text{ref}}$ ,  $s(\theta) \approx 0$ 。此时  $\text{DGR} \approx \beta/2$  (很大), 梯度很强。这形成了一个自毁循环:

1. 训练开始:  $s(\theta) \approx 0$ , DGR 很大, 梯度很强;
2. 模型快速拉开正负样本预测误差的差距,  $\Delta_\theta$  迅速减小;
3.  $s(\theta)$  迅速变负;
4. 由于  $e^{-\beta s(\theta)}$  的指数放大力量, DGR 急剧衰减到接近零;
5. 梯度消失, 训练停滞 (过早收敛)。

整个过程可能在几百步内完成。

### 5.2.4 $\beta$ 越大问题越严重

从表 1 可以看出, DGR 对  $\beta s(\theta)$  的乘积极其敏感:

- $\beta = 100$  时,  $s(\theta) = -0.01$  对应  $\beta s(\theta) = -1$ ,  $\text{DGR} \approx 269$  (仍有有效梯度);
- $\beta = 1000$  时, 同样的  $s(\theta) = -0.01$  对应  $\beta s(\theta) = -10$ ,  $\text{DGR} \approx 0.045$  (梯度几乎消失)。

$\beta$  越大, DGR 的初始值越高, 但下降也越陡峭, 训练越不稳定。

## 6 DPO 的结构性缺陷：缺乏对样本分布的直接监督

### 6.1 问题诊断

从式(15)和(16)可以看出，DPO的损失函数只约束正负样本预测误差的差值 $\Delta_\theta(x^w, x^l)$ ，而不直接约束各自的 $\|y^w - y_\theta(x_t^w, t)\|^2$ 和 $\|y^l - y_\theta(x_t^l, t)\|^2$ 。设：

$$A = \|y^w - y_\theta(x_t^w, t)\|^2 \quad (\text{正样本预测误差}) \quad (44)$$

$$B = \|y^l - y_\theta(x_t^l, t)\|^2 \quad (\text{负样本预测误差}) \quad (45)$$

则训练目标是最小化 $A - B$ 。

### 6.2 理想路径 vs “作弊” 路径

**理想路径：** $A$ 下降（更好地重建正样本）， $B$ 上升（抑制负样本）。

Table 2: 理想的优化路径

阶段	$A$ (正样本误差)	$B$ (负样本误差)	$A - B$
初始	10	10	0
更新后	5	15	-10

**“作弊” 路径：** $A$ 和 $B$ 同时增大，但 $B$ 涨得更快。

Table 3: “作弊” 优化路径——损失值相同，但模型已退化

阶段	$A$ (正样本误差)	$B$ (负样本误差)	$A - B$
初始	10	10	0
更新后	50	60	-10

两条路径产生完全相同的 $A - B = -10$ ，因此损失函数值完全一样。但在“作弊”路径中，正样本的预测误差从10涨到50，模型的生成质量严重退化。

更极端的情况：

Table 4: 极端“作弊”——模型完全坍塌

阶段	$A$ (正样本误差)	$B$ (负样本误差)	$A - B$
初始	10	10	0
更新后	500	510	-10

损失函数对此完全“无感”，但模型对所有输入的预测都已崩溃。

### 6.3 高维空间中“作弊”路径为何被优先选择

**命题 6.1** (高维球体的体积集中性质). 考虑 $n$ 维欧氏空间中半径为 $R$ 的闭球 $B^n(A, R) = \{x \in \mathbb{R}^n : \|x - A\| \leq R\}$ 。设随机点 $y$ 在 $B^n(A, R)$ 内均匀分布，则对任意 $0 < r < R$ :

$$P(\|y - A\| \leq r) = \left(\frac{r}{R}\right)^n \quad (46)$$

当 $r$ 减小、 $R$ 增大或 $n$ 增大时， $P(\|y - A\| \leq r) \rightarrow 0$ 。

证明：由均匀分布的性质，概率等于体积之比：

$$P(\|y - A\| \leq r) = \frac{V(B^n(r))}{V(B^n(R))} = \frac{\frac{\pi^{n/2}}{\Gamma(n/2+1)} r^n}{\frac{\pi^{n/2}}{\Gamma(n/2+1)} R^n} = \left(\frac{r}{R}\right)^n \quad \square \quad (47)$$

**直觉：**在高维球体中，绝大部分体积集中在球的**外壳**附近。如果  $y_\theta$  当前距离目标  $y^w$  为  $r$ ，对  $y_\theta$  做一次无约束的随机扰动后，新点落在半径  $r$  以内的概率为  $(r/R)^n$ ——当维度  $n$  很大时，这趋近于零。

**在视频生成中，**  $y^w$  的维度极高（多帧  $\times$  高分辨率  $\times$  通道数），因此  $A = \|y^w - y_\theta(x_t^w, t)\|^2$  和  $B = \|y^l - y_\theta(x_t^l, t)\|^2$  都有**天然的膨胀趋势**。DPO 的梯度只关心  $A - B$  的差值方向，不会主动抵抗这种膨胀——甚至会利用它，因为让两个误差都膨胀再控制差值，比精确降低  $A$  要容易得多。

## 7 Reg-DPO：通过 SFT 正则化稳定 DPO

### 7.1 设计思路：从梯度缺陷到最小修补

作者的推理过程是先改梯度，再反推损失函数，而非先设计损失再求梯度。

#### 7.1.1 第一步：诊断梯度缺陷

从式 (40)，用 DGR 简写后，DPO 梯度的结构为：

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[ \underbrace{2 \text{DGR}}_{\text{正样本系数}} \cdot (y^w - y_\theta(x_t^w, t))^T \nabla_\theta y_\theta(x_t^w, t) - \underbrace{2 \text{DGR}}_{\text{负样本系数}} \cdot (y^l - y_\theta(x_t^l, t))^T \nabla_\theta y_\theta(x_t^l, t) \right] \quad (48)$$

#### 核心缺陷

正样本和负样本的梯度项共享**相同的系数** 2 DGR。当 DGR 衰减到接近零时，**两个梯度项同时消失**，不论此时正样本的预测质量如何，模型都停止更新。

#### 7.1.2 第二步：最小修补——给正样本梯度加常数

问题很清楚：DGR 衰减后，正样本失去梯度信号。最直接的修补就是：

给正样本的梯度系数加一个不会衰减的常数  $r > 0$ ，负样本保持不变。

将正样本的系数从 2 DGR 改为  $2(\text{DGR} + r)$ ：

$$\nabla_\theta \mathcal{L}_{\text{Reg-DPO}} = -\mathbb{E} \left[ \underbrace{2(\text{DGR} + r)}_{\text{不会衰减到零}} \cdot (y^w - y_\theta(x_t^w, t))^T \nabla_\theta y_\theta(x_t^w, t) - \underbrace{2 \text{DGR}}_{\text{可以衰减}} \cdot (y^l - y_\theta(x_t^l, t))^T \nabla_\theta y_\theta(x_t^l, t) \right] \quad (49)$$

即使  $\text{DGR} \rightarrow 0$ ，正样本仍有系数  $2r > 0$ ，模型被迫持续优化正样本的重建。

## 7.2 从修改后的梯度反推损失函数

现在验证：是否存在一个损失函数，其梯度恰好等于式 (49)？

将式 (49) 拆分为两部分：

$$\begin{aligned}\nabla_{\theta} \mathcal{L}_{\text{Reg-DPO}} &= \underbrace{-\mathbb{E} \left[ 2 \text{DGR} \cdot \left( (y^w - y_{\theta})^{\top} \nabla_{\theta} y_{\theta} - (y^l - y_{\theta})^{\top} \nabla_{\theta} y_{\theta} \right) \right]}_{\text{第一部分: 原始 DPO 梯度 } \nabla_{\theta} \mathcal{L}_{\text{DPO}}} \\ &\quad + \underbrace{\left( -\mathbb{E} \left[ 2r(y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) \right] \right)}_{\text{第二部分: 多出来的项}}\end{aligned}\quad (50)$$

**第一部分**已知对应损失  $\mathcal{L}_{\text{DPO}} = -\log \sigma(-\beta \cdot s(\theta))$ 。

**第二部分:** 利用式 (35) (第四节第三步的结论) :

$$\nabla_{\theta} \|y^w - y_{\theta}(x_t^w, t)\|^2 = -2(y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) \quad (51)$$

因此:

$$\begin{aligned}-2r(y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) &= r \cdot \left( -2(y^w - y_{\theta}(x_t^w, t))^{\top} \nabla_{\theta} y_{\theta}(x_t^w, t) \right) \\ &= r \cdot \nabla_{\theta} \|y^w - y_{\theta}(x_t^w, t)\|^2\end{aligned}\quad (52)$$

这正好是损失函数  $r \cdot \|y^w - y_{\theta}(x_t^w, t)\|^2$  对  $\theta$  求梯度的结果!

而  $\|y^w - y_{\theta}(x_t^w, t)\|^2$  就是标准的SFT (监督微调) 损失——让模型预测尽量接近正样本的目标输出。

### 反推完成

将两部分合并, 得到 Reg-DPO 的损失函数:

$$\mathcal{L}_{\text{Reg-DPO}} = \mathbb{E}_{(c, x_w, x_l) \sim \mathcal{D}} \left[ \underbrace{-\log \sigma(-\beta \cdot s(\theta))}_{\text{原始 DPO 损失}} + \underbrace{r \cdot \|y^w - y_{\theta}(x_t^w, t)\|^2}_{\text{SFT 正则化项}} \right] \quad (53)$$

## 7.3 设计思路总结

作者的完整推理链路为:

**步骤1: 观察问题**—DGR  $\rightarrow 0$  导致梯度消失, 正负样本都失去监督;

**步骤2: 最小改动**—给正样本梯度系数加不衰减的常数  $r$ , 负样本保持不变;

**步骤3: 写出修改后的梯度**—式 (49) (Eq.5) ;

**步骤4: 拆分梯度**—原 DPO 梯度 + 多出来的项;

**步骤5: 识别多出来的项**—发现它等于  $r \cdot \nabla_{\theta} \|y^w - y_{\theta}\|^2$ ;

**步骤6: 反推损失**—多出来的损失  $= r \cdot \|y^w - y_{\theta}\|^2 = \text{SFT 损失}$ 。

本质上, 作者并非凭空想到“加一个 SFT 项”, 而是从梯度层面诊断问题, 做了最小修补, 再反推发现这 SFT 正则化。

## 8 SFT 正则化如何防止过拟合与灾难性遗忘

### 8.1 机制一: 为正样本提供持续的梯度信号

从式 (49), 正样本的梯度系数从  $2 \text{DGR}$  变为  $2(\text{DGR} + r)$ 。

**在标准 DPO 中:** 当  $\text{DGR} \rightarrow 0$  时, 正样本梯度系数  $\rightarrow 0$ , 模型停止学习。

**在 Reg-DPO 中:** 即使  $\text{DGR} \rightarrow 0$ , 正样本仍有系数  $2r > 0$ , 确保模型持续朝着准确重建正样本的方向优化。

## 8.2 机制二：间接约束负样本分布

虽然 SFT 项  $r \cdot \|y^w - y_\theta(x_t^w, t)\|^2$  只显式作用于正样本，但它对负样本的分布也产生了间接约束。

原理：加入 SFT 项后，模型收敛需要同时满足两个条件：

- (a) DGR 衰减（DPO 部分的收敛要求）；
- (b)  $\|y^w - y_\theta(x_t^w, t)\|^2$  足够小（SFT 部分的收敛要求）。

条件 (b) 直接阻止了正样本误差  $A$  的膨胀。而一旦  $A$  被控制，模型就无法再通过“同时增大  $A$  和  $B$ ”来走“作弊”路径。要满足 DPO 的偏好学习目标 ( $A - B$  变负)，模型只能老老实实地让  $A$  下降并让  $B$  上升——这正是我们期望的理想行为。

用之前的数值例子：

Table 5: Reg-DPO 下“作弊”路径被惩罚				
路径	$A$	$B$	$A - B$	SFT 惩罚 $r \cdot A$
理想	5	15	-10	$5r$
作弊	500	510	-10	$500r$

虽然  $A - B$  相同，但“作弊”路径的 SFT 惩罚是理想路径的 100 倍，总损失大幅增加，因此会被优化器拒绝。

## 8.3 机制三：控制分布偏移幅度

SFT 项本质上要求  $y_\theta$  在正样本上的预测不能偏离目标太远。这等价于一种“锚定”效果：模型在学习偏好的同时

实验中观察到：当  $r$  足够大时 ( $\geq 1 \times \text{DGR}$ )，Win Gap 变为负值（正样本预测优于参考模型），Lose Gap 的异常上升也得到控制，说明正负样本的输出分布都被有效约束了。

## 9 与标准 DPO 中参考模型 KL 约束的对比

标准 DPO 的参考模型约束源自 PPO 目标 (式 (1)) 中的 KL 散度项：

$$\beta D_{\text{KL}}[\pi_\theta(y|x) \parallel \pi_{\text{ref}}(y|x)] \quad (54)$$

这个 KL 约束在推导后被隐式地编码进了 DPO 损失中的  $\Delta_{\text{ref}}(y_w, y_l)$  项。以下从四个维度对比两种约束。

### 9.1 约束层级不同

**参考模型 KL 约束（隐式）**：只关心  $\pi_\theta$  和  $\pi_{\text{ref}}$  在正负样本上的概率差值之差（即  $s(\theta)$ ），不直接约束模型在某

**SFT 正则化（显式）**：直接约束模型在正样本上的重建误差  $\|y^w - y_\theta(x_t^w, t)\|^2$ ，是对单个样本输出的硬性要

### 9.2 约束对象不同

KL 约束是对策略分布整体的软约束，在 DPO 中被简化为冻结参考模型的输出——一个常数项  $\Delta_{\text{ref}}$ 。由于  $\Delta_{\text{ref}}$  不依赖  $\theta$ ，它在训练中不产生关于当前模型的梯度信号。

SFT 项  $r \cdot \|y^w - y_\theta(x_t^w, t)\|^2$  直接参与梯度计算，对当前模型的每一步更新都施加实际的拉力。

### 9.3 失效模式不同

DPO 中参考模型的隐式 KL 约束在实践中不够强：因为  $\Delta_{\text{ref}}$  是常数，模型只需保持  $\Delta_\theta$  相对于  $\Delta_{\text{ref}}$  的差值即可满足约束，同时可以自由地让正负样本的绝对误差同时膨胀。

SFT 项直接惩罚绝对误差  $A$  的增长，堵住了这条“逃逸路径”。

## 9.4 作用时机不同

KL 约束的效力随 DGR 一起消失。这是因为整个 DPO 梯度都乘以 DGR 作为系数（式(48)），当  $DGR \rightarrow 0$  时，KL 约束也同时失效。

SFT 项的梯度系数包含额外的  $r$ （式(49)），在 DGR 接近零时仍然有效。这相当于给训练后期提供了一个

### 一句话总结

标准 DPO 的参考模型约束是“不要离起点太远”的隐式软约束，但在扩散模型的高维空间中容易被架空。DPO 的 SFT 正则化是“必须能准确重建好样本”的显式硬约束，直接绑定了模型的生成质量下限。

## 10 动态权重的选择： $r = DGR$

论文实验中测试了多种  $r$  的设置，包括固定值 ( $r = 25, 50, 125, 300, 500$ ) 和动态值 ( $r = DGR/5, DGR/2, DGR/2$ )。

实验发现：

- 当  $r$  过小 ( $\ll DGR$ ) 时，SFT 正则化不足以稳定训练，性能仍然下降；
- 当  $r$  增大到约  $1 \times DGR$  时，训练变得稳定，性能持续提升；
- 当  $r$  进一步增大 ( $\gg DGR$ ) 时，性能趋于饱和，但不再下降。

选择  $r = DGR$  作为默认设置的原因：

- (1) 它自适应地跟随训练动态变化，无需手动调节；
- (2) 它在正则化强度和偏好学习之间取得了良好平衡；
- (3) 由于 DGR 的值因任务而异，动态设置具有更好的跨场景泛化性。

当  $r = DGR$  时，正样本的梯度系数变为  $2(DGR+DGR) = 4 DGR$ ，是原来的两倍，而负样本保持  $2 DGR$ 。这意味着模型在优化时始终对正样本给予更多关注。

## 11 全文公式索引与总结

Table 6: 核心公式一览

编号	公式名称	作用
(1)	PPO 目标	奖励最大化 + KL 约束
(6)	最优策略	推导出隐式奖励
(8)	隐式奖励	连接策略与奖励
(11)	DPO 损失 (语言模型)	语言模型中的 DPO
(15)	DPO 损失 (扩散模型)	本文核心起点
(40)	DPO 梯度 (Eq.3)	揭示梯度结构
(41)	DGR 定义 (Eq.4)	控制梯度幅度
(49)	Reg-DPO 梯度 (Eq.5)	加入常数 $r$ 修补
(53)	Reg-DPO 损失 (Eq.6)	DPO + SFT 正则化
(46)	高维体积集中	解释“作弊”路径的倾向性

### 全文核心结论

1. **DPO 的结构性缺陷**: 只约束正负样本误差的差值  $A - B$ , 不约束绝对值  $A$  和  $B$ , 允许模型通过同时增大两者来“作弊”。
2. **DGR 的快速衰减**: 由于 sigmoid 和大  $\beta$  的联合作用, DGR 在训练初期迅速衰减到零, 导致梯度消失和过早收敛。
3. **Reg-DPO 的解决方案**: 在正样本梯度系数中加入不衰减的常数  $r$ , 等价于在损失函数中加入 SFT 正则化项, 直接惩罚正样本误差的膨胀, 从而间接约束整体分布偏移。
4. **与 KL 约束的本质区别**: KL 约束是隐式的、相对的、随 DGR 衰减的; SFT 正则化是显式的、绝对的、持久有效的。