

# Supplementary Material for Aligning Text-to-Image Diffusion Models to Human Preference by Classification

## 1 Theorem proof

**Theorem 1.** We say a diffusion model  $\epsilon_{\text{ali}}(\mathbf{x}_t, \mathbf{y}, t)$  is ideal alignment if it satisfies  $\|\epsilon_{\text{ali}}(\mathbf{x}_{t;\mathbf{y}}^+, \mathbf{y}, t) - \epsilon\|_2^2 = 0$  and  $\|\epsilon_{\text{ali}}(\mathbf{x}_{t;\mathbf{y}}^-, \mathbf{y}, t) - \epsilon\|_2^2 = \delta$  for any  $\mathbf{y}$ . Here,  $\mathbf{x}_{t;\mathbf{y}}^+ = \sqrt{\alpha_t}\mathbf{x}_{\mathbf{y}}^+ + \sigma_t\epsilon$  and  $\mathbf{x}_{t;\mathbf{y}}^- = \sqrt{\alpha_t}\mathbf{x}_{\mathbf{y}}^- + \sigma_t\epsilon$ . When the reference model  $\epsilon_{\text{ref}}(\mathbf{x}_t, \mathbf{y}, t) = \epsilon_{\text{ali}}(\mathbf{x}_t, \mathbf{y}, t)$  is an ideal alignment model and  $s_{\text{ali}}(\mathbf{x}, \mathbf{y}) = w_t \|\epsilon - \epsilon_{\text{ali}}(\sqrt{\alpha_t}\mathbf{x} + \sigma_t\epsilon, \mathbf{y}, t)\|_2^2$ , the AM-Softmax loss is upper bounded by the Diffusion-DPO loss. Specifically, we have

$$\begin{aligned} & \log(1 + \exp(-(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y})] - \delta))) \exp(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})]) \\ & \leq \mathbb{E}_{\epsilon, t}[\log(1 + \exp(-(s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}) - s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}))) \exp(s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y}) - s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})))] \end{aligned} \quad (1)$$

*Proof.* Under ideal alignment, we replace  $\epsilon_{\text{ref}}$  with  $\epsilon_{\text{ali}}$  in the Diffusion-DPO loss  $\mathcal{L}_{\text{DPO}}$ :

$$\mathcal{L}_{\text{DPO}} = \mathbb{E}_{\epsilon, t}[\log(1 + \exp(-(s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}) - s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}))) \exp(s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y}) - s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})))] \quad (2)$$

Since  $\|\epsilon_{\text{ali}}(\mathbf{x}_{t;\mathbf{y}}^+, \mathbf{y}, t) - \epsilon\|_2^2 = 0$ , we have  $s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y}) = 0$ , so the loss reduces to:

$$\mathcal{L}_{\text{DPO}} = \mathbb{E}_{\epsilon, t}[\log(1 + \exp(-(s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}) - s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}))) \exp(s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})))] \quad (3)$$

Now applying Jensen's inequality (due to the convexity of the function  $f(z) = \log(1 + e^z)$ ), we obtain:

$$\mathcal{L}_{\text{DPO}} \geq \log(1 + \exp(-(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y})] - s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}))) \exp(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})])) \quad (4)$$

We note that  $w_t$  is a finite constant. Without loss of generality, we assume  $w_t < 1$ . Now, under ideal alignment  $s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y}) = w_t \cdot \delta$ , we have

$$\begin{aligned} \mathcal{L}_{\text{DPO}} & \geq \log(1 + \exp(-(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y})] - \mathbb{E}_{\epsilon, t}[s_{\text{ali}}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y})])) \exp(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})])) \\ & \geq \log(1 + \exp(-(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y})] - \mathbb{E}_{\epsilon, t}[w_t] \cdot \delta)) \exp(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})])) \\ & \geq \log(1 + \exp(-(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^-, \mathbf{y})] - \delta))) \exp(\mathbb{E}_{\epsilon, t}[s_{\theta}(\mathbf{x}_{\mathbf{y}}^+, \mathbf{y})]), \end{aligned} \quad (5)$$

which matches the AM-Softmax-style upper bound, as claimed.  $\square$

**Theorem 2.** Let  $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$  denote  $N$  text prompts and  $D = \{\mathbf{x}_{\mathbf{y}_i}\}$  be corresponding aligned images. We assume that the prior prompt distribution  $p(\mathbf{y})$  and image distribution  $p(\mathbf{x})$  are uniform. To describe the discriminative ability, we define the conditional probability  $p(\mathbf{x}|\mathbf{y})$  as

$$p(\mathbf{y}|\mathbf{x}_{\mathbf{y}_i}) = \begin{cases} \frac{n}{N} & \mathbf{y} = \mathbf{y}_i, \\ \frac{N-n}{N(N-1)} & \mathbf{y} \in \mathbf{Y} - \{\mathbf{y}_i\}. \end{cases} \quad \text{where } n < N. \quad (6)$$

Then,  $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})$  and the optimal diffusion model  $\epsilon_{\text{opt}}(\mathbf{x}_t, \mathbf{y}, t)$ , which achieves minimal diffusion loss over both the training set and the test set, over  $D$  is given by:

$$\epsilon_{\text{opt}}(\mathbf{x}_t, \mathbf{y}, t) = \sum_{\mathbf{x}^{(i)} \in D} \frac{w_i}{\sum_{\mathbf{x}^{(j)} \in D} w_j} \cdot \epsilon_i, \quad (7)$$

$$\text{where } \epsilon_i = \frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}}{\sigma_t}, \lambda = \frac{n(N-1)}{N-n}, w_i = \begin{cases} \lambda \cdot \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}\|_2^2}{2\sigma_t^2}\right), & \mathbf{x}^{(i)} \in \{\mathbf{x}_{\mathbf{y}}\}, \\ \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}\|_2^2}{2\sigma_t^2}\right), & \mathbf{x}^{(i)} \in D - \{\mathbf{x}_{\mathbf{y}}\}. \end{cases}$$

*Proof.* As stated in the paper, we assume that each text prompt  $y$  corresponds to a single aligned image  $\mathbf{x}$ , forming a one-to-one mapping. Under the assumption that the prior distributions  $p(y)$  and  $p(\mathbf{x})$  are uniform over the same support, Bayes' rule implies:

$$p(\mathbf{x} | y) p(y) = p(y | \mathbf{x}) p(\mathbf{x}). \quad (8)$$

Since the priors are uniform and equal, they cancel out, leading to:

$$p(\mathbf{x} | y) = p(y | \mathbf{x}). \quad (9)$$

The optimal diffusion model minimizes the expected denoising error:

$$\mathbb{E}_{\mathbf{x}, t, y} [\|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon\|_2^2], \quad (10)$$

over all models  $\epsilon_\theta$  in the hypothesis space. Since this objective is additive over independent input tuples  $(\mathbf{x}_t, t, y)$ , the minimization implies pointwise optimality. That is, the optimal predictor  $\epsilon_{\text{opt}}$  satisfies:

$$\mathbb{E}_{\mathbf{x}^{(i)} \sim p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)} [\|\epsilon_{\text{opt}}(\mathbf{x}_t, t, y) - \epsilon_i\|_2^2] = \min_{\epsilon_\theta} \mathbb{E}_{\mathbf{x}^{(i)} \sim p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)} [\|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon_i\|_2^2], \quad (11)$$

where  $\epsilon_i = \frac{\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}^{(i)}}{\sigma_t}$  is the noise corresponding to candidate  $\mathbf{x}^{(i)}$ .

To find the optimal solution, we take the gradient of the expected error with respect to the prediction and set it to zero:

$$\frac{\partial}{\partial \epsilon_\theta(\mathbf{x}_t, t, y)} \mathbb{E}_{\mathbf{x}^{(i)} \sim p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)} [\|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon_i\|_2^2] = 0. \quad (12)$$

Solving this yields the optimal denoising prediction:

$$\begin{aligned} \mathbb{E}_{\mathbf{x}^{(i)} \sim p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)} [\|\epsilon_{\text{opt}}(\mathbf{x}_t, t, y) - \epsilon_i\|_2^2] &= 0, \\ \Downarrow \\ \sum_{\mathbf{x}^{(i)} \in D} p(\mathbf{x}^{(i)} | \mathbf{x}_t, y) \epsilon_{\text{opt}}(\mathbf{x}_t, t, y) &= \epsilon_{\text{opt}}(\mathbf{x}_t, t, y) = \sum_{\mathbf{x}^{(i)} \in D} p(\mathbf{x}^{(i)} | \mathbf{x}_t, y) \epsilon_i. \end{aligned} \quad (13)$$

We can write this expectation as a weighted sum:

$$\epsilon_{\text{opt}}(\mathbf{x}_t, t, y) = \underbrace{\sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} p(\mathbf{x}^{(i)} | \mathbf{x}_t, y) \epsilon_i}_{\text{A}} + \underbrace{\sum_{\mathbf{x}^{(i)} \in D - \{\mathbf{x}_y\}} p(\mathbf{x}^{(i)} | \mathbf{x}_t, y) \epsilon_i}_{\text{B}}. \quad (14)$$

Here, term A represents the contribution from positive candidates  $\mathbf{x}_y$  that aligned with  $y$ , while term B accounts for the negative influence of all other candidates in  $D$ .

The conditional distribution  $p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)$  captures both aligned (positive) and unaligned (negative) samples. It can be computed via Bayes' rule as:

$$p(\mathbf{x}^{(i)} | \mathbf{x}_t, y) = \frac{p(\mathbf{x}^{(i)} | y) p(\mathbf{x}_t | \mathbf{x}^{(i)}, y)}{p(\mathbf{x}_t | y)} = \frac{p(\mathbf{x}^{(i)} | y) q(\mathbf{x}_t | \mathbf{x}^{(i)})}{p(\mathbf{x}_t | y)}, \quad (15)$$

where the transition distribution  $q(\mathbf{x}_t | \mathbf{x}^{(i)})$  is Gaussian:  $q(\mathbf{x}_t | \mathbf{x}^{(i)}) = \mathcal{N}(\mathbf{x}_t | \sqrt{\alpha_t} \mathbf{x}^{(i)}, \sigma_t^2 I)$ .

Substituting  $p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)$  and  $\epsilon_i$  into the equation for A, we obtain:

$$\begin{aligned} \text{A} &= \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} p(\mathbf{x}^{(i)} | \mathbf{x}_t, y) \frac{\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}^{(i)}}{\sigma_t} \\ &= \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} \frac{p(\mathbf{x}^{(i)} | y) q(\mathbf{x}_t | \mathbf{x}^{(i)})}{p(\mathbf{x}_t | y)} \frac{\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}^{(i)}}{\sigma_t} \\ &= \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} \frac{p(\mathbf{x}^{(i)} | y)}{p(\mathbf{x}_t | y)} \frac{1}{(2\pi\sigma_t)^{\frac{n}{2}}} \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}^{(i)}\|_2^2}{2\sigma_t^2}\right) \frac{\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}^{(i)}}{\sigma_t}. \end{aligned} \quad (16)$$



To avoid numerical problem caused by  $\frac{1}{(2\pi\sigma_t)^{\frac{n}{2}}}$  and intractable  $\frac{p(\mathbf{x}^{(i)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})}$ , we reorganize this equation using softmax function:

$$A = \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} \frac{\frac{p(\mathbf{x}^{(i)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})} \frac{1}{(2\pi\sigma_t)^{\frac{n}{2}}} \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}\|_2^2}{2\sigma_t^2}\right)}{\sum_{j=1}^{|D|} p(\mathbf{x}^{(j)} | \mathbf{x}_t, \mathbf{y})} \frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}}{\sigma_t}. \quad (17)$$

To simplify the expression, we introduce

$$\epsilon_i = \frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}}{\sigma_t}, \quad \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)}) := \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}\|_2^2}{2\sigma_t^2}\right). \quad (18)$$

We rewrite the expression as follows

$$\begin{aligned} A &= \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} \frac{\frac{p(\mathbf{x}^{(i)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{j=1}^{|\{\mathbf{x}_y\}|} \frac{p(\mathbf{x}^{(j)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{j=1}^{|D-\{\mathbf{x}_y\}|} \frac{p(\mathbf{x}^{(j)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i \\ &= \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} \frac{\frac{n}{N} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{j=1}^{|\{\mathbf{x}_y\}|} \frac{n}{N} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{j=1}^{|D-\{\mathbf{x}_y\}|} \frac{N-n}{N(N-1)} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i \\ &= \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} \frac{\frac{n(N-1)}{N-n} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{j=1}^{|\{\mathbf{x}_y\}|} \frac{n(N-1)}{N-n} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{j=1}^{|D-\{\mathbf{x}_y\}|} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i \\ &= \sum_{\mathbf{x}^{(i)} \in \{\mathbf{x}_y\}} \frac{\lambda \cdot \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{\mathbf{x}^{(j)} \in \{\mathbf{x}_y\}} \lambda \cdot \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{\mathbf{x}^{(j)} \in D-\{\mathbf{x}_y\}} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i. \end{aligned} \quad (19)$$

where  $\lambda = \frac{n(N-1)}{N-n}$ . Similarly,

$$\begin{aligned} B &= \sum_{\mathbf{x}^{(i)} \in D-\{\mathbf{x}_y\}} \frac{\frac{p(\mathbf{x}^{(i)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{j=1}^{|\{\mathbf{x}_y\}|} \frac{p(\mathbf{x}^{(j)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{j=1}^{|D-\{\mathbf{x}_y\}|} \frac{p(\mathbf{x}^{(j)}|\mathbf{y})}{p(\mathbf{x}_t|\mathbf{y})} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i \\ &= \sum_{\mathbf{x}^{(i)} \in D-\{\mathbf{x}_y\}} \frac{\frac{N-n}{N(N-1)} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{j=1}^{|\{\mathbf{x}_y\}|} \frac{n}{N} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{j=1}^{|D-\{\mathbf{x}_y\}|} \frac{N-n}{N(N-1)} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i \\ &= \sum_{\mathbf{x}^{(i)} \in D-\{\mathbf{x}_y\}} \frac{\text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{j=1}^{|\{\mathbf{x}_y\}|} \frac{n(N-1)}{N-n} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{j=1}^{|D-\{\mathbf{x}_y\}|} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i \\ &= \sum_{\mathbf{x}^{(i)} \in D-\{\mathbf{x}_y\}} \frac{\text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)})}{\sum_{\mathbf{x}^{(j)} \in \{\mathbf{x}_y\}} \lambda \cdot \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)}) + \sum_{\mathbf{x}^{(j)} \in D-\{\mathbf{x}_y\}} \text{sim}(\mathbf{x}_t, \mathbf{x}^{(j)})} \epsilon_i. \end{aligned} \quad (20)$$

With 14 and 18, the final expression becomes :

$$\epsilon_{\text{opt}}(\mathbf{x}_t, \mathbf{y}, t) = \sum_{\mathbf{x}^{(i)} \in D} \frac{w_i}{\sum_{\mathbf{x}^{(j)} \in D} w_j} \cdot \epsilon_i, \quad (21)$$

where

$$\begin{aligned} \epsilon_i &= \frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}}{\sigma_t}, \quad \lambda = \frac{n(N-1)}{N-n}, \\ w_i &= \begin{cases} \lambda \cdot \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)}) = \lambda \cdot \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}\|_2^2}{2\sigma_t^2}\right), & \mathbf{x}^{(i)} \in \{\mathbf{x}_y\}, \\ \text{sim}(\mathbf{x}_t, \mathbf{x}^{(i)}) = \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}^{(i)}\|_2^2}{2\sigma_t^2}\right), & \mathbf{x}^{(i)} \in D - \{\mathbf{x}_y\}, \end{cases} \end{aligned}$$

When  $n \rightarrow N$ , we have  $\lambda \rightarrow \infty$ , and the weights on unaligned (negative) samples vanish. Hence, the optimal predictor becomes fully dominated by aligned (positive) samples, recovering a perfect class-conditional denoiser.

□

**Supplementary Explanation: From  $\epsilon_\theta$  to  $\epsilon_{\text{opt}}$  in Equation (13)** In the initial step of the proof, we consider the expected squared error between the model’s predicted noise  $\epsilon_\theta(\mathbf{x}_t, t, y)$  and the ground-truth noise  $\epsilon_i$ , averaged over a conditional distribution of data samples:

$$\mathbb{E}_{\mathbf{x}^{(i)} \sim p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)} \left[ \|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon_i\|_2^2 \right].$$

To find the optimal prediction under this objective, we differentiate the expected error with respect to the model’s output  $\epsilon_\theta$ , and set the gradient to zero:

$$\frac{\partial}{\partial \epsilon_\theta(\mathbf{x}_t, t, y)} \mathbb{E}_{\mathbf{x}^{(i)} \sim p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)} \left[ \|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon_i\|_2^2 \right] = 0.$$

At this stage,  $\epsilon_\theta$  is treated as a free variable (i.e., an optimization target) rather than a fixed model prediction. The goal is to compute the value of  $\epsilon_\theta$  that minimizes the expected squared error.

Solving this equation yields the value of  $\epsilon_\theta$  that achieves the minimum. We denote this solution as:

$$\epsilon_{\text{opt}}(\mathbf{x}_t, t, y) := \arg \min_{\epsilon_\theta} \mathbb{E}_{\mathbf{x}^{(i)} \sim p(\mathbf{x}^{(i)} | \mathbf{x}_t, y)} \left[ \|\epsilon_\theta(\mathbf{x}_t, t, y) - \epsilon_i\|_2^2 \right].$$

This optimal solution corresponds to the expected value of the ground-truth noise under the given conditional distribution. Thus, while the derivation begins by optimizing with respect to  $\epsilon_\theta$ , the result is a closed-form expression for the optimal value, which we denote  $\epsilon_{\text{opt}}$  to distinguish it from arbitrary predictions made by a parametric model.

## 2 Notations

Table 1: Notations used in our method.

Symbol	Description
$\mathbf{x}_0$	Clean image sampled from real data distribution $q(\mathbf{x}_0)$ .
$\mathbf{x}_y^\pm$	Clean image $\mathbf{x}_0$ aligned $(+)$ or misaligned $(-)$ with prompt $y$ in preference pairs.
$\mathbf{x}_{t;y}^\pm$	Noisy version of $\mathbf{x}_y^\pm$ at timestep $t$ .
$\mathbf{x}_{y^\pm}$	Clean image $\mathbf{x}_0$ conditioned on aligned $(+)$ and misaligned $(-)$ prompts $y^\pm$ .
$\epsilon$	Ground-truth noise sampled from $\mathcal{N}(0, I)$ , added during forward diffusion.
$\epsilon_\theta(\mathbf{x}, t)$	Noise predicted by the model $\epsilon_\theta$ at timestep $t$ , given noisy input $\mathbf{x}$ .
$\epsilon_{\text{ali}}(\mathbf{x}, y, t)$	Noise predicted by the ideal alignment model for input image $\mathbf{x}$ , prompt $y$ and timestep $t$ .
$\epsilon_{\text{ref}}(\mathbf{x}, y, t)$	Noise predicted by the reference diffusion model for input image $\mathbf{x}$ , prompt $y$ and timestep $t$ .
$\epsilon_{\text{opt}}(\mathbf{x}, y, t)$	Noise predicted by the optimal diffusion model for input image $\mathbf{x}$ , prompt $y$ and timestep $t$ .
$s_\theta(\mathbf{x}, t)$	Score predicted by the training model for image sample $\mathbf{x}$ and prompt $y$ .
$s_{\text{ref}}(\mathbf{x}, y)$	Score predicted by the reference model for image sample $\mathbf{x}$ and prompt $y$ .
$\delta$	Margin used to enforce separation between aligned $(+)$ and misaligned $(-)$ scores.
$\Delta_y^\pm$	Margin offsets applied to the expected scores of aligned $(+)$ and misaligned $(-)$ images with prompt $y$ .
$O_y^\pm$	Ideal score targets for aligned $(+)$ and misaligned $(-)$ images with prompt $y$ .
$\eta_y^\pm$	Scaling factors that modulate the contributions of aligned $(+)$ and misaligned $(-)$ images with prompt $y$ .
$\iota(x_y^-, x_y^+, y)$	Score difference in ABC loss between aligned $(+)$ and misaligned $(-)$ images with prompt $y$ .

### 3 More Evaluation Details

Table 2: Win rate on PartiPrompts for SD1.5 and SDXL-based models.

	PickScore	HPS	Aesthetics	CLIP
vs. SD1.5-Base	<b>60.02 <math>\pm</math> 2.40%</b>	<b>81.51 <math>\pm</math> 2.21%</b>	<b>74.27 <math>\pm</math> 2.14%</b>	<b>59.72 <math>\pm</math> 2.45%</b>
vs. SD1.5-DPO	<b>55.85 <math>\pm</math> 2.44%</b>	<b>73.02 <math>\pm</math> 2.38%</b>	<b>64.90 <math>\pm</math> 2.34%</b>	44.97 $\pm$ 2.43%
vs. SD1.5-SPO	<b>51.16 <math>\pm</math> 2.41%</b>	<b>61.59 <math>\pm</math> 2.41%</b>	47.60 $\pm$ 2.45%	<b>60.02 <math>\pm</math> 2.36%</b>
vs. SD1.5-KTO	<b>57.77 <math>\pm</math> 2.41%</b>	44.72 $\pm$ 2.38%	<b>53.90 <math>\pm</math> 2.44%</b>	47.22 $\pm$ 2.45%
vs. SDXL-Base	<b>74.38 <math>\pm</math> 1.84%</b>	<b>79.26 <math>\pm</math> 1.97%</b>	<b>80.20 <math>\pm</math> 1.56%</b>	<b>52.46 <math>\pm</math> 2.04%</b>
vs. SDXL-DPO	<b>73.22 <math>\pm</math> 2.18%</b>	<b>72.50 <math>\pm</math> 2.33%</b>	<b>68.25 <math>\pm</math> 1.39%</b>	<b>50.51 <math>\pm</math> 1.88%</b>
vs. SDXL-SPO	<b>52.49 <math>\pm</math> 2.34%</b>	40.31 $\pm$ 2.44%	<b>59.93 <math>\pm</math> 2.41%</b>	<b>55.53 <math>\pm</math> 2.44%</b>
vs. SDXL-MAPO	<b>65.35 <math>\pm</math> 1.81%</b>	<b>81.17 <math>\pm</math> 2.07%</b>	<b>72.10 <math>\pm</math> 1.82%</b>	46.97 $\pm$ 2.06%

Table 3: Win rate on HPS benchmarks for SD1.5 and SDXL-based models.

	PickScore	HPS	Aesthetics	CLIP
vs. SD1.5-Base	<b>74.83 <math>\pm</math> 3.10%</b>	<b>85.75 <math>\pm</math> 2.87%</b>	<b>68.84 <math>\pm</math> 3.31%</b>	<b>59.65 <math>\pm</math> 3.53%</b>
vs. SD1.5-DPO	<b>53.46 <math>\pm</math> 3.51%</b>	<b>71.50 <math>\pm</math> 3.46%</b>	<b>64.19 <math>\pm</math> 3.56%</b>	<b>52.06 <math>\pm</math> 3.57%</b>
vs. SD1.5-SPO	45.35 $\pm$ 3.57%	<b>54.99 <math>\pm</math> 3.53%</b>	38.08 $\pm$ 3.47%	<b>64.83 <math>\pm</math> 3.13%</b>
vs. SD1.5-KTO	<b>52.28 <math>\pm</math> 3.52%</b>	42.88 $\pm$ 3.48%	<b>52.86 <math>\pm</math> 3.57%</b>	<b>53.93 <math>\pm</math> 3.55%</b>
vs. SDXL-Base	<b>79.35 <math>\pm</math> 2.23%</b>	<b>70.17 <math>\pm</math> 2.58%</b>	<b>72.28 <math>\pm</math> 2.86%</b>	<b>60.38 <math>\pm</math> 3.26%</b>
vs. SDXL-DPO	<b>77.26 <math>\pm</math> 3.06%</b>	<b>69.54 <math>\pm</math> 3.42%</b>	<b>70.19 <math>\pm</math> 2.39%</b>	<b>57.06 <math>\pm</math> 3.07%</b>
vs. SDXL-SPO	<b>51.16 <math>\pm</math> 3.34%</b>	<b>52.41 <math>\pm</math> 3.57%</b>	46.78 $\pm$ 3.53%	<b>59.87 <math>\pm</math> 3.58%</b>
vs. SDXL-MAPO	<b>68.55 <math>\pm</math> 2.37%</b>	<b>64.89 <math>\pm</math> 3.11%</b>	<b>68.18 <math>\pm</math> 2.96%</b>	<b>51.14 <math>\pm</math> 3.19%</b>

Table 4: Scores on PartiPrompts for SD1.5 and SDXL-based models.

	PickScore	HPS	Aesthetics	CLIP
SD1.5-Base	21.25 $\pm$ 2.02	26.98 $\pm$ 2.85	5.29 $\pm$ 1.17	29.57 $\pm$ 9.71
SD1.5-DPO	21.49 $\pm$ 2.19	27.16 $\pm$ 3.13	5.36 $\pm$ 1.08	29.80 $\pm$ 9.46
SD1.5-SPO	21.53 $\pm$ 2.33	27.33 $\pm$ 3.74	5.89 $\pm$ 1.04	28.13 $\pm$ 9.23
SD1.5-KTO	21.46 $\pm$ 1.96	27.70 $\pm$ 3.21	5.62 $\pm$ 0.99	30.78 $\pm$ 8.09
SD1.5-ABC	21.85 $\pm$ 2.04	27.97 $\pm$ 2.85	5.93 $\pm$ 1.01	31.07 $\pm$ 8.14
SDXL-Base	22.76 $\pm$ 2.45	28.49 $\pm$ 3.59	5.86 $\pm$ 1.05	35.76 $\pm$ 9.66
SDXL-DPO	22.94 $\pm$ 2.37	28.93 $\pm$ 3.52	6.01 $\pm$ 0.98	36.01 $\pm$ 8.73
SDXL-SPO	23.56 $\pm$ 2.64	29.12 $\pm$ 3.52	6.26 $\pm$ 0.92	33.82 $\pm$ 9.84
SDXL-MAPO	22.82 $\pm$ 2.40	28.62 $\pm$ 3.62	5.98 $\pm$ 1.04	36.58 $\pm$ 9.39
SDXL-ABC	23.79 $\pm$ 2.27	29.42 $\pm$ 3.29	6.35 $\pm$ 0.89	36.81 $\pm$ 8.51

Table 5: Scores on HPS benchmarks for SD1.5 and SDXL-based models.

	PickScore	HPS	Aesthetics	CLIP
SD1.5-Base	21.17 $\pm$ 2.45	27.61 $\pm$ 3.10	5.45 $\pm$ 1.01	35.73 $\pm$ 7.38
SD1.5-DPO	21.71 $\pm$ 2.41	28.23 $\pm$ 3.35	5.59 $\pm$ 0.99	36.66 $\pm$ 7.95
SD1.5-SPO	21.99 $\pm$ 2.76	28.53 $\pm$ 3.69	5.96 $\pm$ 1.10	33.14 $\pm$ 9.15
SD1.5-KTO	21.79 $\pm$ 2.53	28.95 $\pm$ 3.12	5.62 $\pm$ 0.96	37.01 $\pm$ 8.01
SD1.5-ABC	21.97 $\pm$ 2.45	28.86 $\pm$ 3.02	5.72 $\pm$ 0.87	37.18 $\pm$ 7.72
SDXL-Base	23.26 $\pm$ 2.50	29.38 $\pm$ 3.59	6.08 $\pm$ 1.03	37.24 $\pm$ 6.74
SDXL-DPO	23.59 $\pm$ 2.65	29.86 $\pm$ 3.40	6.14 $\pm$ 1.00	38.34 $\pm$ 5.98
SDXL-SPO	23.76 $\pm$ 2.70	30.30 $\pm$ 3.18	6.48 $\pm$ 0.86	37.62 $\pm$ 7.14
SDXL-MAPO	23.60 $\pm$ 2.53	29.92 $\pm$ 3.52	6.19 $\pm$ 0.92	38.61 $\pm$ 7.17
SDXL-ABC	24.39 $\pm$ 2.38	30.67 $\pm$ 3.08	6.54 $\pm$ 0.86	38.97 $\pm$ 6.02

We report extended evaluation tables with empirical confidence intervals, computed by discarding the top 5% of deviations from the mean to cover 95% of the scores. We also provide the absolute scores for each setting.

## 4 More Visualization Results

Table 6: Detailed prompts used for generated images in Figure 3.

Image	Prompt
Row 1, Col1	A view of the Orion constellation in the night sky.
Row 1, Col2	Snow mountain and tree reflection in the lake.
Row 1, Col3	An old oil painting of Dubrovnik on canvas.
Row 1, Col4	A white country home with a wrap-around porch.
Row 1, Col5	Sunset over the sea.
Row 2, Col1	Two pianos next to each other.
Row 2, Col2	Teacups surrounding a kettle.
Row 2, Col3	A pumpkin with a candle in it.
Row 2, Col4	An impressionistic painting of tree and a building.
Row 2, Col5	A plant with small flowers with purple petals.
Row 3, Col1	A cinematic shot of Avatar Azula.
Row 3, Col2	An anime-style depiction of a boy that showcases impressive artistic skill.
Row 3, Col3	One child on a couch.
Row 3, Col4	A portrait of young girl.
Row 3, Col5	A smiling beautiful sorceress wearing a modest high necked blue suit surrounded by swirling rainbow aurora, hyper-realistic, cinematic, post-production.
Row 4, Col1	A rabbit sitting on a turtle's back.
Row 4, Col2	A young badger delicately sniffing yellow roses, richly textured oil painting.
Row 4, Col3	A pile of chicken eggs with a confused chicken nearby.
Row 4, Col4	An eagle.
Row 4, Col5	A cute digital art of a unicorn.
Row 5, Col1	Anime illustration of Gundam mech suit on Pixiv.
Row 5, Col2	A room with two chairs and paintings.
Row 5, Col3	A picture of some food in the plate.
Row 5, Col4	A shiny VW van with a cityscape painted on it and parked on grass.
Row 5, Col5	Splashing milk and berries in a porcelain bowl, vintage wooden tabletop, black background, soft light.

Table 7: Detailed prompts used for generated images in Figure 6.

Image	Prompt
Row 1, Col1	A realistic and elegant depiction of a young Garfield in the style of Stefan Kostic by Stanley Lau, also known as Artgerm.
Row 1, Col2	A 3D model of Huggy Wuggy from Poppy Playtime video game, depicted as an oil painting in a hyperrealistic style by Yanjun Cheng.
Row 1, Col3	An oil painting of astronauts in space.
Row 1, Col4	A close-up portrait of a girl with an autumn leaves headdress.
Row 1, Col5	Album cover featuring a golden swimming princess, created by Conrad Roset.
Row 2, Col1	A skyline with several buildings with old architecture.
Row 2, Col2	A futuristic building towers above a river, reflecting on the water with beautiful lighting.
Row 2, Col3	A castle with towers and catapults is heading towards an iceberg at dawn in a highly detailed painting.
Row 2, Col4	The image features a cathedral-like structure in a canyon grotto, created by Sparth and Greg Rutkowski with micro details and 3D sculpture techniques.
Row 2, Col5	A Chinese painting of Shanghai featuring blooming sakura trees.
Row 3, Col1	Fruit and vegetable displayed in glass container on table.
Row 3, Col2	One slice of simple cheese pizza on a paper plate.
Row 3, Col3	A collage of roses painted in oil by Anders Zorn, with intricate detail and elegant composition.
Row 3, Col4	A throw rug on the floor.
Row 3, Col5	Graffiti of a rocket ship on a brick wall.
Row 4, Col1	Three-quarters front view of a blue 1977 Ford F-150 coming around a curve in a mountain road and looking over a green valley on a cloudy day.
Row 4, Col2	An armchair in the shape of an avocado.
Row 4, Col3	A detailed portrait of a Border Collie with ominous atmosphere.
Row 4, Col4	A small house in the wilderness.
Row 4, Col5	The image depicts the character Totoro from Studio Ghibli, illustrated by Joe Fenton.
Row 5, Col1	An office break room with a table.
Row 5, Col2	A living room filled with books furniture and a flat screen TV.
Row 5, Col3	An advertising painting of a tennis court with bright colors, featuring art by Claude Monet done in impressionistic oil painting style.
Row 5, Col4	A wooden deck overlooking a mountain valley.
Row 5, Col5	A kitchen has a stainless steel refrigerator and other appliances.






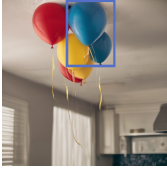




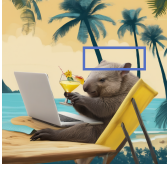




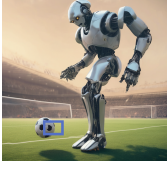


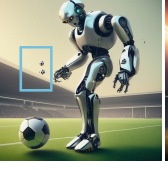

Prompts	SDXL-Base	SDXL-DPO	SDXL-SPO	SDXL-MAPO	SDXL-ABC
<i>"KEEP OFF THE GRASS" on a sign next to a lawn.</i>					
<i>Red and yellow balloons hanging from a ceiling fan.</i>					
<i>A wombat wearing a white panama hat and a floral Hawaiian shirt.</i>					
<i>A robot kicking a soccer ball.</i>					

Figure 1: Qualitative Comparison for SDXL based Text-to-Image Alignment.

Prompts	SDXL-Base	SDXL-DPO	SDXL-SPO	SDXL-MAPO	SDXL-ABC
<i>A close-up of a long-island ice tea cocktail.</i>					
<i>A kitchen in natural wood style.</i>					
<i>A beat-up truck at the base of the Great Pyramid.</i>					
<i>A palm tree forest in front of the Kremlin.</i>					

Figure 2: Qualitative Comparison for SDXL based Aesthetic Alignment.





Figure 3: Sample Images Generated by SDXL-ABC. Prompts are provided in Table 6.



Prompts	SD1.5-Base	SD1.5-DPO	SD1.5-SPO	SD1.5-KTO	SD1.5-ABC
<i>Ceramic bowl with <b>two</b> ice cream scoops, <b>blueberries</b>, and <b>banana</b> slices.</i>					
<i>A small round wooden <b>table</b> beside a sofa, with <b>a potted plant</b> on top.</i>					
<i>Gray tote bag on a hanger with <b>a black cat</b> print.</i>					
<i>A strawberry shortcake with <b>one</b> burning candle standing behind it.</i>					

Figure 4: Qualitative Comparison for SD1.5 based Text-to-Image Alignment.

Prompts	SD1.5-Base	SD1.5-DPO	SD1.5-SPO	SD1.5-KTO	SD1.5-ABC
<i>A digital painting of a motorcycle in sharp focus, created as concept art.</i>					
<i>The benches along the sidewalk are covered with snow.</i>					
<i>Two men at a cafe talking about something.</i>					
<i>A wizard with fox-like features in detailed, furry artwork.</i>					

Figure 5: Qualitative Comparison for SD1.5 based Aesthetic Alignment.





Figure 6: Sample Images Generated by SD1.5-ABC. Prompts are provided in Table 7.