TERNARY PREFERENCE OPTIMIZATION FOR TEXT-TO-IMAGE GENERATION WITH NOISY HUMAN FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent text-to-image alignment methods typically rely on two assumptions about human preference data: (1) annotators always strictly prefer one image over another, and (2) preference labels are noise-free and faithfully reflect underlying human preferences. In practice, both assumptions can be violated. For example, approximately 11% of annotations in Pick-a-Pic are ties, indicating that annotators may consider two images equally preferable. Such ties provide meaningful preference information but are typically discarded by binary preference modeling. Meanwhile, human preference labels are inherently noisy due to individual subjectivity and annotator disagreement. To address these limitations, we propose *Noisy Ternary Preference Optimization* (NTPO), a unified framework that explicitly models three preference outcomes—preferred, rejected, and tied—while accounting for noisy and inconsistent annotations. By preserving tie information and explicitly modeling annotation noise, NTPO provides a more faithful formulation of human preference alignment for text-to-image generation. Experiments across multiple benchmarks demonstrate that NTPO improves both alignment quality and robustness compared with existing binary preference optimization methods.

1 INTRODUCTION

Direct Preference Optimization (DPO) (Rafailov et al., 2023) has emerged as a simple and effective approach for aligning generative models with human preferences. Unlike reinforcement learning from human feedback, DPO directly optimizes the target model from preference data by interpreting the likelihood ratio between the target and reference policies as an implicit reward. It has been widely applied to both large language models (LLMs) (Rafailov et al., 2023) and diffusion models (DMs) (Wallace et al., 2024). However, existing text-to-image preference alignment methods (Fu et al., 2025; Lu et al., 2025) largely rely on two assumptions underlying DPO: each image pair has a strict preference ordering, and the resulting preference label is accurate and noise-free. In practice, both assumptions can be violated, limiting the reliability of preference-based alignment.

First, existing preference alignment methods typically impose a binary ordering on pairs of generated images, assuming that annotators can identify one image as preferable to the other. In practice, however, two images may be perceived as sufficiently similar in preference, resulting in a tie. To obtain strict comparisons, existing data collection pipelines often rely on over-generation and filtering, discarding tied comparisons. This is particularly wasteful given the high cost of human annotation. Moreover, several text-to-image preference datasets, such as Pick-a-Pic (Kirstain et al., 2023b) and Datapoint (Datapoint AI, 2026), retain tie cases alongside preferred and dispreferred comparisons, suggesting that ties contain meaningful preference information rather than being annotation failures. Thus, treating human preference as inherently binary may result in substantial information loss.

Second, human preference labels are inherently susceptible to noise. Human judgments in the preference dataset can vary substantially across annotators (Fu et al., 2015), while careless annotation (Kittur et al., 2008) and the lack of standardized evaluation criteria (Chen et al., 2013; Long et al., 2013) can further introduce erroneous labels. Recent studies have also identified substantial noise in text-to-image preference datasets (Yang et al., 2025; Li et al., 2026a; Liu et al., 2026b), motivating increasing attention to robust preference alignment (Wu et al., 2025; Cao et al., 2026; Li et al., 2026b). Such noise can arise from both systematic and random labeling errors, posing a challenge to reliable preference alignment.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

These observations motivate us to develop a preference optimization framework that jointly accounts for ties and noisy feedback. Specifically, we replace the binary Bradley–Terry (BT) (Bradley & Terry, 1952) preference model underlying DPO with the ternary Rao–Kupper (RK) model (Rao & Kupper, 1967), which explicitly models three outcomes: win, tie, and loss. We further extend the ternary RK model to account for noisy preference labels, yielding a noisy ternary RK model. Incorporating these models into the DPO framework leads to our Noisy Ternary Preference Optimization (NTPO) objectives for text-to-image preference alignment. Experiments on text-to-image preference benchmarks show that modeling ties allows NTPO to exploit additional preference information, while the noisy formulation improves robustness to imperfect human feedback.

Briefly, our primary contributions can be unfolded as:

❶ We first extend the conventional binary preference modeling framework to jointly account for tied judgments and noisy human feedback. Specifically, we generalize the binary BT model to a noisy ternary RK model, which explicitly models three types of preference relations: win, tie, and loss, while accounting for noise in the observed preference labels. This formulation enables a more comprehensive representation of real-world human preferences.

❷ Building on the noisy ternary RK model, we introduce Noisy Ternary Preference Optimization (NTPO), a family of preference optimization objectives designed to learn from noisy ternary preference feedback. By explicitly incorporating both tie relations and label noise, NTPO can exploit a broader range of human preference information and provides a more robust framework for preference alignment in text-to-image generation.

❸ We conduct extensive experiments on SD1.5, SDXL, Flux and AuraFlow to evaluate the effectiveness and generality of NTPO. The results demonstrate that explicitly modeling tied judgments and noisy feedback consistently improves preference alignment across various text-to-image models and preference evaluation settings. Moreover, NTPO maintains stable performance gains under different levels of preference noise, highlighting its robustness to imperfect human feedback.

2 Related Work

To the best of our knowledge, we are the first to study noisy ternary preference optimization. Existing studies on preference optimization with ties primarily focus on language-based tasks under clean preference settings, whereas recent work on noisy preference optimization for text-to-image generation mainly considers binary preference feedback. In contrast, we jointly incorporates ties and noisy preference feedback, a combination that remains largely unexplored in both text-to-image generation and large language models.

Noisy Binary Preference Alignment for Text-to-Image Models. Existing methods for handling noisy binary feedback generally follow two strategies: filtering preference data to retain reliable samples or modifying the optimization objective to reduce the influence of noisy feedback. For data filtering, FiFA (Yang et al., 2025) selects preference pairs based on preference margins, text quality, and diversity, while Semi-DPO (Liu et al., 2026b) identifies a consensus-based clean subset and iteratively pseudo-labels conflicting pairs. For objective modification, α-DPO (Li et al., 2026b) reduces the influence of noisy observations through α-divergence, while DeDPO (Pham et al., 2026) corrects systematic bias in synthetic preference annotations. However, these methods remain within the binary preference setting and therefore cannot explicitly model ties or recover clean ternary preference probabilities from noisy feedback.

Clean Ternary Preference Alignment for LLMs. Existing ternary preference alignment methods for LLMs incorporate ties through preference optimization, reward modeling, or preference annotation. At the preference optimization stage, TODO (Guo et al., 2025), DPO-RK, and DPO-D (Chen et al., 2026) extend binary preference modeling to explicitly capture win, tie, and loss outcomes using the Tie-Oriented-BT, RK, and Davidson models, respectively. For reward modeling, CROME (Srivastava et al., 2026) augments preference data with neutral pairs labeled as ties. From the annotation perspective, *When Labelers Stay Silent* (Lv et al., 2026) allows annotators to abstain on indistinguishable responses and incorporates the resulting ties into preference optimization. Despite explicitly modeling ties, these LLMs methods generally assume reliable ternary labels and do not address corrupted labels or provide robustness guarantees under noisy ternary feedback.

3 PRELIMINARIES

This section presents the preliminaries of TPO. We first introduce preference alignment and then review preference optimization methods based on pairwise preference modeling, including binary preference optimization and ternary preference optimization.

Preference Alignment aims to retrain a generative model to align its generation behavior with human preferences. Given a prompt $\boldsymbol{c}_n$, the model generates the n-th samples $\boldsymbol{x}_n^1$ and $\boldsymbol{x}_n^2$ without imposing a preference between them. For notational convenience, we denote the comparison tuple by $\boldsymbol{z}_n = (\boldsymbol{c}_n, \boldsymbol{x}_n^1, \boldsymbol{x}_n^2)$. A preference dataset is then denoted as $\mathcal{D} = \{(\boldsymbol{z}_n, y_n)\}_{n=1}^N$, where y_n represents the preference label associated with the n-th comparison sample. However, human annotators may express different preferences between the two responses. In general, there are three possible preference outcomes, denoted by y:

$$y_n = \begin{cases} 0, & \text{if } \boldsymbol{x}_n^1 \succ \boldsymbol{x}_n^2 \mid \boldsymbol{c}_n, \text{ indicating that humans prefer } \boldsymbol{x}_n^1, \\ 1, & \text{if } \boldsymbol{x}_n^1 \prec \boldsymbol{x}_n^2 \mid \boldsymbol{c}_n, \text{ indicating that humans prefer } \boldsymbol{x}_n^2, \\ 2, & \text{if } \boldsymbol{x}_n^1 \sim \boldsymbol{x}_n^2 \mid \boldsymbol{c}_n, \text{ indicating that humans have no clear preference .} \end{cases} \tag{1}$$

Accordingly, the preference probability of the n-th comparison is denoted by $p(y = y_n \mid \boldsymbol{z}_n)$. Based on how the preference label probability is modeled, existing preference alignment methods can be categorized into *binary preference optimization*, which considers $\succ$ and $\prec$, and *ternary preference optimization*, which considers both $\succ$, $\prec$ and $\sim$.

Binary Preference Optimization (BPO) models preferences using the BT model (Bradley & Terry, 1952), which assumes that a human annotator always strictly prefers one response over the other. For a response pair $(\boldsymbol{x}_n^1, \boldsymbol{x}_n^2)$, the binary preference label is defined as $y_n \in \{0, 1\}$. Under the BT model, the corresponding preference probabilities are given by

$$p^{\mathrm{BT}}(y = 0 \mid \boldsymbol{z}_n) = \sigma(\Delta(\boldsymbol{z}_n)), \qquad p^{\mathrm{BT}}(y = 1 \mid \boldsymbol{z}_n) = \sigma(-\Delta(\boldsymbol{z}_n)), \tag{2}$$

where $\Delta(\boldsymbol{z}_n) = r(\boldsymbol{c}_n, \boldsymbol{x}_n^1) - r(\boldsymbol{c}_n, \boldsymbol{x}_n^2)$, $\boldsymbol{z}_n = (\boldsymbol{c}_n, \boldsymbol{x}_n^1, \boldsymbol{x}_n^2)$, $\sigma(\cdot)$ denotes the sigmoid function, and $r(\boldsymbol{c}_n, \boldsymbol{x}_n)$ denotes the reward function. Building on the BT model, Rafailov et al. (2023) implicitly parameterize the reward function in terms of the likelihood ratio between the target policy $p_{\boldsymbol{\theta}}$ and the reference policy p_{ref}. Let β be a temperature parameter. The corresponding implicit reward margin is defined as:

$$\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) = \beta \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_n^1 \mid \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_n^1 \mid \boldsymbol{c}_n)} - \beta \log \frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_n^2 \mid \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_n^2 \mid \boldsymbol{c}_n)}, \qquad \boldsymbol{z}_n = (\boldsymbol{c}_n, \boldsymbol{x}_n^1, \boldsymbol{x}_n^2). \tag{3}$$

Substituting this margin into the BT model (2) yields the parameterized preference probabilities:

$$p_{\boldsymbol{\theta}}^{\mathrm{BT}}(y = 0 \mid \boldsymbol{z}_n) = \sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)), \qquad p_{\boldsymbol{\theta}}^{\mathrm{BT}}(y = 1 \mid \boldsymbol{z}_n) = \sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)). \tag{4}$$

Given a preference dataset consisting of N samples, the BPO objective is formulated as the negative log-likelihood:

$$\mathcal{L}_{\mathsf{BT}}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}^{\mathrm{BT}}(y = y_n \mid \boldsymbol{z}_n). \tag{5}$$

Ternary Preference Optimization (TPO) models preferences using the RK model (Rao & Kupper, 1967), which extends the BT model by allowing a human annotator to either strictly prefer one response or express no clear preference. For a response pair $(\boldsymbol{x}_n^1, \boldsymbol{x}_n^2)$, the ternary preference label y is defined as in Equation 1. Substituting the implicit reward margin $\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)$ (3) into the RK model yields the following parameterized ternary preference probabilities:

$$p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = 0 \mid \boldsymbol{z}_n) = \sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha), \quad p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = 1 \mid \boldsymbol{z}_n) = \sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha) \tag{6}$$

$$p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = 2 \mid \boldsymbol{z}_n) = \sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) + \alpha) - \sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha), \tag{7}$$

where $\alpha > 0$ is a tie parameter that controls the width of the tie region. Given the preference dataset consisting of N samples, the TPO objective is formulated as the empirical negative log-likelihood:

$$\mathcal{L}_{\mathsf{RK}}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = y_n \mid \boldsymbol{z}_n). \tag{8}$$

4 TERNARY PREFERENCE OPTIMIZATION WITH NOISY HUMAN FEEDBACK

Current alignment methods assume that preference labels are noise-free and rely on binary preference optimization to align the model with human preferences. To address the limitations of binary assumption and noisy feedback, we introduce NTPO and apply it to text-to-image generation alignment, enabling the model to jointly account for annotation noise and tied preferences.

4.1 NOISY TERNARY PREFERENCE OPTIMIZATION

Preference labels in the dataset are provided by human annotators and may inevitably contain errors due to factors such as carelessness and inconsistent evaluation. We refer to such annotation errors as *label flips*. Given a clean preference dataset $\mathcal{D} = \{(\boldsymbol{z}_n, y_n)\}_{n=1}^N$, let $p(y = i \mid \boldsymbol{z}_n)$ denote the clean probability of assigning class i to $\boldsymbol{z}_n$. We use $\widetilde{y}_n$ to denote the observed label provided by the human annotator, which may differ from the clean label y_n due to label flips. The corresponding noisy preference dataset is denoted by $\widetilde{\mathcal{D}} = \{(\boldsymbol{z}_n, \widetilde{y}_n)\}_{n=1}^N$, where $\widetilde{p}(\widetilde{y} = j \mid \boldsymbol{z}_n)$ denotes the probability of observing noisy label j given $\boldsymbol{z}_n$. We model the label-flip process as $p(\widetilde{y} = j \mid y = i) = q_{j,i}$, parameterized by a 3×3 transition matrix $Q = (q_{j,i}) \in \mathbb{R}^{3\times 3}$. We assume that label flips are independent of the preference sample $\boldsymbol{z}_n$. Therefore, the noisy probability $\widetilde{p}$ that an input $\boldsymbol{z}_n$ is assigned label j in the noisy preference data is given by

$$\widetilde{p}(\widetilde{y} = j \mid \boldsymbol{z}_n) = \sum_{i=0}^{2} p(\widetilde{y} = j \mid y = i) p(y = i \mid \boldsymbol{z}_n) = \sum_{i=0}^{2} q_{j,i}\, p(y = i \mid \boldsymbol{z}_n). \tag{9}$$

Exploiting the symmetry between $\prec$ and $\succ$, we have $p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = 1 \mid \boldsymbol{z}) = p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = 0 \mid \boldsymbol{z})$. Substituting the RK preference probabilities $\{p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z_n})\}_{i=0}^2$ into Equation 9 yields:

$$\widetilde{p}_{\boldsymbol{\theta},Q}^{\mathrm{RK}}(\widetilde{y} = j \mid \boldsymbol{z}) = \sum_{i=0}^{2} q_{j,i}\, p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z}). \tag{10}$$

The resulting objective $\mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta})$ consists of a RK preference model parameterized by $\boldsymbol{\theta}$ and a label-flip model parameterized by Q. Given Q, our noisy ternary preference optimization estimates $\boldsymbol{\theta}$ by maximizing the likelihood of the observed noisy labels $\widetilde{y}_n$.

$$\mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N} \log \widetilde{p}_{\boldsymbol{\theta},Q}^{\mathrm{RK}}\left(\widetilde{y} = \widetilde{y}_n \mid \boldsymbol{z}_n\right) = -\frac{1}{N}\sum_{n=1}^{N} \log\Big(\sum_{i=0}^{2} q_{\widetilde{y}_n,i}\, p_{\boldsymbol{\theta}}^{\mathrm{RK}}\left(y = i \mid \boldsymbol{z}_n\right)\Big) \tag{11}$$

4.2 UNBIASED OBJECTIVE

In this section, we investigate whether the noisy preference objective 11, optimized on the noisy dataset $\widetilde{\mathcal{D}}$ characterized by $\widetilde{p}$, can recover the clean ground-truth preference distribution p underlying the clean dataset $\mathcal{D}$. Formally, let $\boldsymbol{\theta}^*$ denote the population optimum of the preference objective on $\widetilde{\mathcal{D}}$, and let $\widetilde{p}_{\boldsymbol{\theta}^*}(\widetilde{y} = j \mid \boldsymbol{z})$ denote the preference probability induced by the resulting model, and $p(y = y_n \mid \boldsymbol{z}_n)$ denotes the ground-truth preference probability for class y_n. We define unbiasedness according to whether the induced preference distribution coincides with the ground-truth distribution. Employing this criterion, we have Definition 1.

Definition 1. *The objective is said to be* unbiased *with respect to the ground-truth preference distribution if Equation 12a holds, and* biased *if Equation 12b holds:*

$$\widetilde{p}_{\boldsymbol{\theta}^*}\left(\widetilde{y} = j \mid \boldsymbol{z}\right) = p\left(y = j \mid \boldsymbol{z}\right), \qquad \forall j \in \{0, 1, 2\}, \tag{12a}$$

$$\widetilde{p}_{\boldsymbol{\theta}^*}\left(\widetilde{y} = j \mid \boldsymbol{z}\right) \neq p\left(y = j \mid \boldsymbol{z}\right), \qquad \exists j \in \{0, 1, 2\}. \tag{12b}$$

The transition matrix Q 9 allows for arbitrary label-flip processes. In this work, we consider a practical noise model for human preference annotation that captures inherent inconsistencies in human judgments. Although annotation errors may arise from various sources, such as careless judgments and inconsistent evaluation, we focus on the latter. Careless errors can, in principle, be reduced through careful annotation procedures and quality control, whereas inconsistencies in evaluation are

intrinsic to subjective human preferences and cannot be completely eliminated. In particular, inconsistent evaluation may cause annotators to confuse two conflicting strict preferences. We therefore assume that a strict preference is flipped to its conflicting preference with probability $\eta \in [0, \frac{1}{2})$, i.e., $\boldsymbol{x}_n^1 \succ \boldsymbol{x}_n^2$ is flipped to $\boldsymbol{x}_n^1 \prec \boldsymbol{x}_n^2$, or vice versa. Under this assumption, the transition matrix Q is parameterized by the noise rate η as

$$Q = (q_{j,i}) = \begin{bmatrix} 1-\eta & \eta & 0 \\ \eta & 1-\eta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{13}$$

This transition matrix provides a simple yet practical characterization of annotation noise. Although label flips may generally occur between any pair of preference states and arise from multiple sources, our empirical observations suggest that confusion between conflicting strict preferences is the dominant source of annotation errors. We therefore ignore less prevalent error patterns and adopt the above transition matrix, yielding a more tractable model and preference optimization objective.

Although only noisy preference labels are available for training, the NTPO objective can recover the ground-truth preference distribution. Theorem 1 establishes this bias-correction property.

Theorem 1 (Proof is in the Appendix A.1)**.** *The NTPO objective 11, optimized over the noisy preference dataset $\widetilde{\mathcal{D}}$, is unbiased with respect to the preference distribution in the sense of Definition 1, satisfying Equation 12a.*

In contrast, optimizing Equation 8 over the noisy preference dataset $\widetilde{\mathcal{D}}$ generally fails to recover the ground-truth preference distribution. This bias is formally characterized in Theorem 2.

Theorem 2 (Proof is in the Appendix A.2)**.** *The TPO objective 8, optimized over the noisy preference dataset $\widetilde{\mathcal{D}}$, is biased with respect to the groundtrue preference distribution in the sense of Definition 1, satisfying Equation 12b.*

Above two theorems establish the theoretical advantage of the proposed NTPO objective over the TPO objective under noisy preference feedback: NTPO recovers the clean preference distribution, whereas TPO is generally biased by label noise.

4.3 Adaptation to Text-to-Image Models

We next adapt the NTPO objective 11 to text-to-image models. Since the optimization-dependent term only involves $p_{\boldsymbol{\theta}}^{\mathrm{RK}}$, whose preference probabilities are defined by Equations 6 and 7, the objective can be optimized through the preference difference $\Delta_{\boldsymbol{\theta}}$ 3. However, directly optimizing it in diffusion and flow-matching models is challenging because it involves a non-convex function $f(\Delta_{\boldsymbol{\theta}})$, which prevents the direct application of Jensen's inequality in Equation 14.

$$f\left(\mathbb{E}[\Delta_{\boldsymbol{\theta}}]\right) \not\leq \mathbb{E}\left[f(\Delta_{\boldsymbol{\theta}})\right]. \tag{14}$$

To obtain a tractable objective that is convex with respect to $\Delta_{\boldsymbol{\theta}}$, we introduce a variational weight matrix $R = (r_{\widetilde{y},i}) \in \mathbb{R}^{3\times 3}$, where each element $r_{\widetilde{y},i}$ represents the variational weight assigned to the clean label $y = i$ given the observed noisy label $\widetilde{y}$. These variational weights can be viewed as an approximation to the posterior distribution of the clean label. For the considered symmetric label flip, the variational weight matrix is parameterized by ρ as

$$R = [r_{\widetilde{y},i}]_{\widetilde{y},i=0}^{2} = \begin{bmatrix} 1-\rho & \rho & 0 \\ \rho & 1-\rho & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{15}$$

Using this variational weighting, we obtain the following variational surrogate objective:

$$\mathcal{L}_{\mathsf{VFC}}^{\eta}(\boldsymbol{\theta}, R) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=0}^{2} r_{\widetilde{y}_n,i} \log \frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}\left(y = i \mid \boldsymbol{z}_n\right)}{r_{\widetilde{y}_n,i}^{\rho}}. \tag{16}$$

Since Equation 16 induces a convex function $f_{\mathrm{VFC}}(\Delta_{\boldsymbol{\theta}}; \rho)$ with respect to $\Delta_{\boldsymbol{\theta}}$, Jensen's inequality can be directly applied to establish the upper bound in Equation 17.

$$f_{\mathsf{VFC}}\left(\mathbb{E}[\Delta_{\boldsymbol{\theta}}]; \rho\right) \leq \mathbb{E}\left[f_{\mathsf{VFC}}(\Delta_{\boldsymbol{\theta}}; \rho)\right]. \tag{17}$$

Furthermore, when the matrix R satisfies the optimal posterior condition, the surrogate objective 16 attains the same minimum value as the NTPO objective in 11, as confirmed in Theorem 3.

Theorem 3 (Proof is in the Appendix A.3)**.** *Under the optimal R, the surrogate and noisy preference objectives in Equation 16 and Equation 11 share the same minimum value.*

Based on Equation 16, we derive tractable training objectives for diffusion and flow matching models. Since the optimization dependent term in Equation 16 only involves the preference margin $\Delta_{\boldsymbol{\theta}}$, we instantiate this term for both model classes. Specifically, following the notation of Diffusion-DPO (Wallace et al., 2024) and Flow-GRPO (Liu et al., 2026a), we instantiate $\Delta_{\boldsymbol{\theta}}$ as $\Delta^{\mathsf{Diff}}_{\boldsymbol{\theta},t}$ (137) and $\Delta^{\mathsf{Flow}}_{\boldsymbol{\theta},t}$ (140) for diffusion and flow matching models, respectively. The corresponding objective formulations for both model classes are given below, with detailed derivations in the Appendix A.4.

$$\Delta^{\mathsf{Diff}}_{\boldsymbol{\theta},t} = -\beta T \omega_t \Big[\|\boldsymbol{\epsilon}^1 - \boldsymbol{\epsilon}^1_{\boldsymbol{\theta}}\|_2^2 - \|\boldsymbol{\epsilon}^1 - \boldsymbol{\epsilon}^1_{\mathrm{ref}}\|_2^2 - \|\boldsymbol{\epsilon}^2 - \boldsymbol{\epsilon}^2_{\boldsymbol{\theta}}\|_2^2 + \|\boldsymbol{\epsilon}^2 - \boldsymbol{\epsilon}^2_{\mathrm{ref}}\|_2^2 \Big]. \tag{18}$$

$$\Delta^{\mathsf{Flow}}_{\boldsymbol{\theta},t} = -\beta T \Big[\|\boldsymbol{v}^1 - \boldsymbol{v}^1_{\boldsymbol{\theta}}\|_2^2 - \|\boldsymbol{v}^1 - \boldsymbol{v}^1_{\mathrm{ref}}\|_2^2 - \|\boldsymbol{v}^2 - \boldsymbol{v}^2_{\boldsymbol{\theta}}\|_2^2 + \|\boldsymbol{v}^2 - \boldsymbol{v}^2_{\mathrm{ref}}\|_2^2 \Big]. \tag{19}$$

5 Experiments

To demonstrate the efficacy of our NTPO objective, we conduct a comprehensive empirical evaluation centered around the following research questions:

❶ **Performance Gains:** *Does explicitly modeling preference ties and noisy feedback improve text-to-image preference alignment?* In Section 5.2, we first compare TPO with BPO to evaluate the benefit of incorporating preference ties. We then compare TPO with NTPO to assess whether explicitly modeling noisy human feedback provides further performance gains.

❷ **Application Generality:** *Can the proposed NTPO objective generalize across different text-to-image models?* In Section 5.3, we evaluate NTPO on different text-to-image models, including SD3.5-M (Esser et al., 2024), FLUX.1-schnell (Labs et al., 2025), and AuraFlow-v0.3 (Fal, 2024), in addition to the model used in Section 5.2, to assess its generality across model architectures.

❸ **Ablation Studies:** *How do the tie and noise modeling components affect performance, and how sensitive is the proposed objective to their associated hyperparameters?* In Section 5.4, we investigate four key factors of NTPO: **(1) Effect of Transition Matrix Q Assumption.** We compare different noise transition matrix assumptions to evaluate their impact on preference optimization. **(2) Effect of Tie and Noise Modeling.** We analyze the contributions of tie modeling and noise correction by comparing binary, ternary, and noise-aware variants. **(3) Sensitivity to the Tie Parameter α.** We evaluate the influence of different α in Equations 6 and 7. **(4) Robustness to the Noise Ratio η.** We examine the robustness of NTPO when the assumed noise ratio differs from the actual annotation noise level.

5.1 Experimental Setup

Models and Datasets. We evaluate NTPO on diffusion models (SD1.5 (Rombach et al., 2022), SDXL (Podell et al., 2024)) and flow-matching models (SD3.5-M, FLUX.1-schnell, and AuraFlow-v0.3). Following previous text-to-image preference alignment works, we train on Pick-a-Pic v2 (Kirstain et al., 2023a) and evaluate on HPSv2 (Wu et al., 2023) and PartiPrompt (Yu et al., 2022).

Baseline Methods. We compare NTPO with representative binary and ternary preference optimization methods. The binary baselines include DPO, DMPO (Li et al., 2025), and α-DPO, while TPO is included as a ternary preference baseline by removing the noise correction component from NTPO.

Evaluation Metrics. We evaluate all methods on HPSV2 and PartiPrompts using human preference metrics (PickScore (PS) (Kirstain et al., 2023a), HPSv2 (Wu et al., 2023), aesthetic score (AES) (Schuhmann et al., 2022), ImageReward (IR) (Xu et al., 2023)) and the CLIP score (Radford et al., 2021) for text-image alignment. Higher scores indicate better performance.

Implementation Details. For diffusion models, we follow the Diffusion-DPO training configurations. For flow matching models, we use AdamW with a learning rate of 1×10^{-4} for 100 steps and set $\beta = 500$. Across all experiments, we use $\alpha = 0.5$ and $\eta = 0.2$. During inference, diffusion models use 50 sampling steps with CFG scales of 7.5 for SD1.5 and 5.0 for SDXL, while flow matching models use their respective default configurations. All experiments are conducted on eight NVIDIA A800 GPUs.

Table 1: Performance comparison on the HPSv2 and PartiPrompt Benchmark using SD1.5 and SDXL. Best results are shown in **bold**, and second-best results are underlined.

Model	Method	HPSv2					PartiPrompt				
		PS↑	HPS↑	AES↑	IR↑	CLIP↑	PS↑	HPS↑	AES↑	IR↑	CLIP↑
SD1.5	*Binary preference alignment*										
	SD1.5	20.93	27.02	5.437	-0.200	32.82	21.50	27.19	5.400	-0.019	31.00
	DPO	21.33	27.53	5.531	0.353	33.67	21.71	27.54	5.462	0.356	31.63
	DMPO	21.97	28.70	5.708	0.705	34.91	22.05	28.45	5.718	0.608	33.17
	α-DPO	22.09	28.71	5.709	0.697	**35.35**	22.04	28.22	5.627	0.531	33.05
	Ternary preference alignment										
	TPO	22.03	28.78	5.698	0.711	34.56	22.09	28.20	5.742	0.622	33.54
	Noisy ternary preference alignment										
	NTPO (Ours)	**22.26**	**28.93**	**5.812**	**0.838**	34.72	**22.28**	**28.78**	**5.956**	**0.741**	**33.95**
SDXL	*Binary preference alignment*										
	SDXL	23.16	29.16	6.066	0.881	37.75	22.66	28.30	5.876	0.648	34.09
	DPO	23.47	29.84	6.179	1.003	37.80	22.83	28.86	5.962	0.998	34.36
	DMPO	23.53	30.19	6.166	1.115	38.22	22.97	28.75	6.020	1.060	34.49
	α-DPO	23.43	29.82	6.262	1.112	38.43	22.84	28.91	6.080	0.898	34.41
	Ternary preference alignment										
	TPO	23.65	30.14	6.285	1.147	38.13	23.07	29.02	6.121	**1.083**	34.63
	Noisy ternary preference alignment										
	NTPO (Ours)	**23.86**	**30.34**	**6.447**	**1.188**	**38.50**	**23.26**	**29.25**	**6.237**	0.978	**34.95**



Figure 1: Qualitative comparison on SD1.5 and SDXL. NTPO produces images with improved visual quality and prompt alignment. More results are provided in the Appendix A.8.

5.2 Alignment Performance

In this section, we evaluate whether modeling ties and noisy human feedback improves text-to-image alignment. We compare NTPO with binary preference methods and its ternary variant TPO (NTPO without noise correction) through quantitative and qualitative evaluations. Due to space limitations, complete quantitative results on SD1.5 and SDXL are reported in Table 1, while a brief quantitative evaluation of flow matching models is summarized in Table 2. Qualitative comparisons are presented in Figure 1. More additional results are provided in the Appendix A.7 and A.8.

Quantitative Comparison. Table 1 summarizes the results on HPSv2 and PartiPrompts using SD1.5 and SDXL. We compare binary preference alignment (DPO, DMPO, and α-DPO), ternary preference alignment (TPO), and our NTPO. As shown in Table 1, TPO improves over binary methods by leveraging tie supervision, while NTPO further enhances alignment by correcting noisy preference bias. The complete results are provided in the Table 5 in Appendix A.7.

Qualitative Comparison. Due to space limitations, we provide qualitative comparisons on SD1.5 and SDXL in Figure 1. Additional visualizations, including results from flow-matching models, are provided in the Appendix A.8. Under the same settings, we compare the base model, DPO, DMPO, α-DPO, TPO, and NTPO. The results show that TPO improves generation quality over binary methods, while NTPO further enhances alignment by correcting noisy feedback.

Table 2: Generalization to flow matching models evaluated on HPSv2. Best results are shown in **bold**, and second-best results are underlined. Additional results are provided in the Appendix A.7.

Metric	SD3.5-M			FLUX.1-schnell			AuraFlow-v0.3		
	Base	TPO	**NTPO**	Base	TPO	**NTPO**	Base	TPO	**NTPO**
PS↑	22.95	23.17	**23.44**	23.01	23.55	**23.87**	22.45	23.03	**23.16**
HPS↑	28.88	29.32	**29.85**	29.59	30.14	**30.51**	28.15	28.77	**29.35**
AES↑	6.091	6.115	**6.197**	6.103	6.188	**6.305**	6.011	6.134	**6.247**

Table 3: Ablation study of transition matrix designs for NTPO on SD1.5. We compare the general matrix Q with the proposed simplified matrix on HPSv2 alignment and training efficiency.

Method	**Transition Matrix**	**Training Time (h)**	**PS↑**	**HPS↑**	**AES↑**	**IR↑**	**CLIP↑**
NTPO	General Q	**35**	**22.30**	**28.98**	**5.825**	0.831	**34.78**
NTPO	Simplified Q	29	22.26	28.93	5.812	**0.838**	34.72

Table 4: Ablation of tie modeling and noise correction on SD1.5. Best results in **bold**.

Model	**Method**	**Tie**	**Noise**	**HPSv2**					**PartiPrompt**				
				PS↑	**HPS↑**	**AES↑**	**IR↑**	**CLIP↑**	**PS↑**	**HPS↑**	**AES↑**	**IR↑**	**CLIP↑**
SD1.5	DPO	✗	✗	21.33	27.53	5.531	0.353	33.67	21.71	27.54	5.462	0.356	31.63
	TPO	✓	✗	22.03	28.78	5.698	0.711	34.56	22.09	28.20	5.742	0.622	33.54
	NTPO	✓	✓	**22.26**	**28.93**	**5.812**	**0.838**	**34.72**	**22.28**	**28.78**	**5.956**	**0.741**	**33.95**

5.3 Generalization Studies

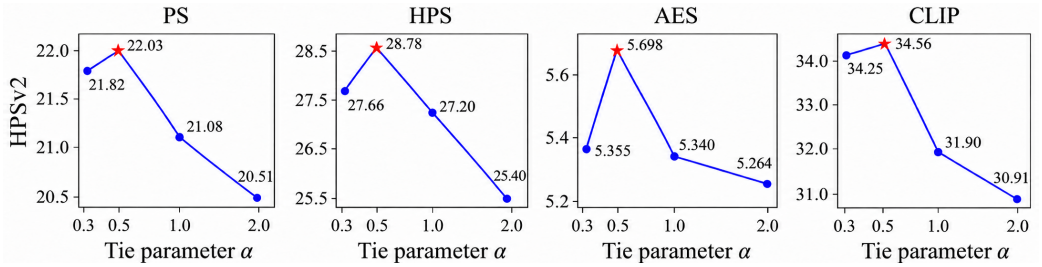
Generalization to Flow Matching Models. To verify the generality of NTPO beyond diffusion-based models, we extend our framework to recent flow-matching models, including SD3.5-M, FLUX.1-schnell, and AuraFlow-v0.3. Due to space limitations, we provide a concise evaluation in Table 2, with additional results provided in the Appendix A.7. As shown in Table 2, NTPO consistently improves over TPO across different flow-matching models, achieving higher PS, HPS, and AES scores. These results indicate that the proposed preference alignment strategy is independent of the underlying generative formulation and can effectively exploit noisy human feedback across diverse model architectures.

5.4 Ablation Studies

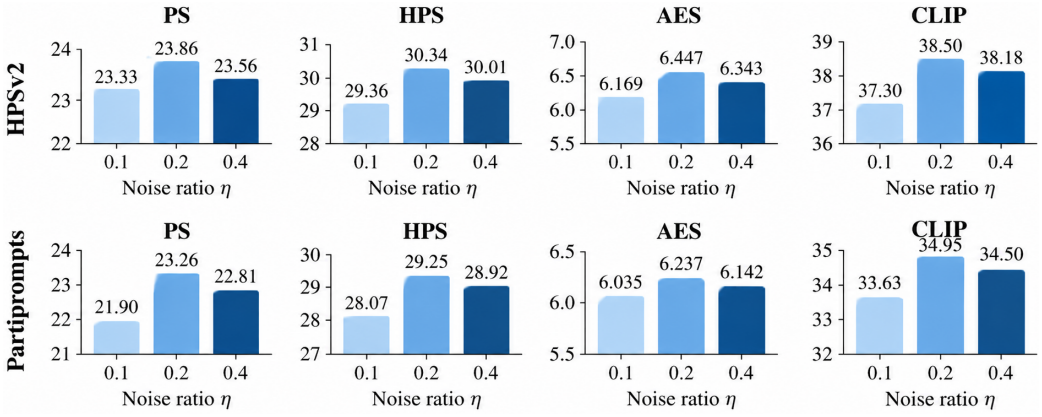
We conduct ablation studies to analyze the key design choices of NTPO. Considering the computational cost of extensive ablations, we perform these studies on SD1.5 and SDXL, with additional results provided in the Appendix A.9. Specifically, we first examine the effect of the transition matrix design, then investigate the effect of tie and noise modeling. Finally, we evaluate the sensitivity to the tie parameter α and the robustness to the noise ratio η.

Effectiveness of the Simplified Transition Matrix Q. We investigate whether the simplified transition matrix Q, introduced in Section 4.2, can achieve comparable performance to the fully parameterized transition matrix while reducing training complexity. Under the same model and training configuration, Table 3 shows that the simplified Q achieves comparable alignment performance across all metrics with reduced training time. These results demonstrate that NTPO can effectively model noisy ternary preference feedback using the simplified transition matrix while reducing the complexity of noise transition estimation. More results are provided in Appendix A.9.

Effect of Tie and Noise Modeling. We evaluate the individual contributions of tie modeling and noise correction in NTPO. We first compare the binary preference optimization method (such as DPO) with the ternary preference method (TPO) incorporating tie information. As shown in Table 4, TPO consistently outperforms DPO on both HPSv2 and PartiPrompts, demonstrating that explicitly modeling tied comparisons provides more informative preference supervision. We further compare TPO with NTPO, which additionally models noisy preference feedback. NTPO consistently improves over TPO, indicating that noise correction effectively mitigates the bias introduced by noisy preference labels.



(a) Sensitivity of TPO to the tie parameter α on SD1.5 evaluated on the HPSv2.



(b) Robustness of NTPO to the assumed noise ratio η on SDXL evaluated on the HPSv2 and PartiPrompt.

Figure 2: Sensitivity analysis of the proposed framework. (a) Sensitivity of TPO to the tie parameter α. (b) Robustness of NTPO to the assumed noise ratio η on HPSv2 and PartiPrompt.

Sensitivity to the Tie Parameter α. We first investigate the effect of the tie parameter α in TPO, which controls the strength of the tie region in the RK model. As shown in Figure 2a, we evaluate TPO with different values of α on SD1.5 using the HPSv2 benchmark. The results show that an excessively small α leads to insufficient modeling of tied comparisons, while an excessively large α weakens the distinction between tied and strict preference cases, resulting in degraded alignment performance. Among the evaluated settings, $\alpha = 0.5$ achieves the best overall performance and is therefore adopted as the default setting.

Robustness to the Noise Ratio η. We further investigate the robustness of NTPO to the assumed noise ratio η. Specifically, we apply Cleanlab (Northcutt et al., 2021) to filter Pick-a-Pic v2 and obtain a clean preference dataset, from which we construct a noisy dataset by randomly flipping preference labels with a probability of 20%. Detailed implementation is provided in Appendix A.5. We set $\eta = 0.2$ according to the injected noise level and further evaluate NTPO under different noise ratio assumptions. As shown in Figure 2b, the matched noise ratio ($\eta = 0.2$) achieves the best overall performance. Underestimating the noise ratio ($\eta = 0.1$) causes a substantial performance drop, whereas overestimating it ($\eta = 0.4$) results in only a minor degradation. This indicates that NTPO is robust to noise-ratio misspecification when the assumed noise ratio overestimates the actual noise level.

6 Conclusion

In this work, we investigate ternary preference optimization for text-to-image alignment under noisy human feedback. Existing methods typically assume strict pairwise preferences and noise-free annotations, overlooking tied comparisons and noisy feedback, which can lead to biased preference learning. To address this issue, we propose NTPO, a unified framework that preserves tie information through ternary preference modeling and corrects noisy annotations through label noise modeling. NTPO recovers clean preference probabilities from noisy data while remaining robust to preference noise. Experiments on diffusion and flow-matching models demonstrate that NTPO improves alignment quality and robustness over existing preference optimization methods.

7 AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript and for retrieval and discovery of relevant literature. All AI-assisted text was carefully reviewed and revised by the authors, and the retrieved literature was manually verified against the original sources for relevance and correctness. Generative AI tools were not used to generate experimental results or draw scientific conclusions. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

8 ETHICS STATEMENT

This work studies preference optimization for text-to-image generative models using existing preference datasets and does not involve the collection of new human-subject data. Our method is intended to improve the robustness of preference learning under noisy human feedback. We acknowledge that text-to-image generative models may inherit biases and potential risks from their training data and preference annotations. We follow the ICLR Code of Ethics and take responsibility for the ethical considerations associated with this work.

9 REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide full methodological details in Sections 4, with additional proofs in Appendix A.1–A.3. We also provide additional derivations in Appendix A.4.

REFERENCES

Ralph Allan Bradley and Milton E Terry. Rank Analysis of Incomplete Block Designs: I. the Method of Paired Comparisons. *Biometrika*, 1952.

Xiaoyang Cao, Zelai Xu, Mo Guang, Kaiwen Long, Michiel Bakker, Yu Wang, and Chao Yu. REPO: Robust Enhanced Policy Optimization as A General Framework for LLM Alignment. In *International Conference on Learning Representations*, 2026.

Jinghong Chen, Guangyu Yang, Weizhe Lin, Jingbiao Mei, Chenxu Lyu, and Bill Byrne. On Extending Direct Preference Optimization to Accommodate Ties. *Neural Information Processing Systems*, 2026.

Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise Ranking Aggregation in a Crowdsourced Setting. In *Web Search and Data Mining*, 2013.

Datapoint AI. Text-to-Image Human Preferences 2M: the Datapoint Image Bench voting record, 2026. URL `https://huggingface.co/datasets/datapointai/text-2-image-human-preferences-2m`. 2,161,160 leaderboard-grade pairwise votes over 30 models, 500 prompts, full round-robin.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In *International Conference on Machine Learning*, 2024.

Fal. Introducing AuraFlow v0.1, an Open Exploration of Large Rectified Flow Models, 2024. Available at: `https://blog.fal.ai/auraflow/` [Accessed: 2024-11-07].

Minghao Fu, Guo-Hua Wang, Liangfu Cao, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. CHATS: Combining Human-Aligned Optimization and Test-Time Sampling for Text-to-Image Generation. *International Conference on Machine Learning*, 2025.

Yanwei Fu, Timothy M Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. Robust Subjective Visual Property Prediction from Crowdsourced Pairwise Labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

Yuxiang Guo, Lu Yin, Bo Jiang, and Jiaqi Zhang. TODO: Enhancing LLM Alignment with Ternary Preferences. In *International Conference on Learning Representations*, 2025.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation. In *Neural Information Processing Systems*, 2023a.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An Open Dataset of User Preferences for Text-to-Image Generation. *Neural Information Processing Systems*, 2023b.

Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing User Studies with Mechanical Turk. In *SIGCHI Conference on Human Factors in Computing Systems*, 2008.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space, 2025.

Binxu Li, Minkai Xu, Jiaqi Han, Meihua Dang, and Stefano Ermon. Divergence Minimization Preference Optimization for Diffusion Model Alignment. *ArXiv*, 2025.

Ming Li, Jie Wu, Jiaxing Cui, Xiaojie Li, Rui Wang, and Chen Chen. ViPO: Visual Preference Optimization at Scale. In *International Conference on Learning Representations*, 2026a.

Yang Li, Songlin Yang, Wei Wang, Xiaoxuan Han, and Jing Dong. $alpha$-DPO: Robust Preference Alignment for Diffusion Models via $alpha$ Divergence. In *International Conference on Learning Representations*, 2026b.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-GRPO: Training Flow Matching Models via Online RL. *Neural Information Processing Systems*, 2026a.

Xinxin Liu, Ming Li, Zonglin Lyu, Yuzhang Shang, and Chen Chen. Learning from Noisy Preferences: A Semi-Supervised Learning Approach to Direct Preference Optimization. *International Conference on Learning Representations*, 2026b.

Chengjiang Long, Gang Hua, and Ashish Kapoor. Active Visual Recognition with Expertise Estimation in Crowdsourcing. In *International Conference on Computer Vision*, 2013.

Yunhong Lu, Qichao Wang, Hengyuan Cao, Xiaoyin Xu, and Min Zhang. Smoothed Preference Optimization via Renoise Inversion for Aligning Diffusion Models with Varied Human Preferences. *International Conference on Machine Learning*, 2025.

Jiaqi Lv, Zihan Zhang, Chengjing Yang, Shiyu Xia, Ning Xu, and Xin Geng. When Labelers Stay Silent: The Power of Ties in Cost-Effective Preference Learning. In *International Conference on Machine Learning*, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards Wide-Spectrum Human Preference Score. In *International Conference on Computer Vision*, 2025.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels. *Journal of Artificial Intelligence Research*, 2021.

Khiem Pham, Quang Nguyen, Tung Nguyen, Jingsen Zhu, Michele Santacatterina, Dimitris Metaxas, and Ramin Zabih. DeDPO: Debiased Direct Preference Optimization for Diffusion Models. *ArXiv*, 2026.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *International Conference on Learning Representations*, 2024.

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 *International Conference on Machine Learning*, 2021.

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. *Neural Information Processing Systems*, 2023.

Pejaver V Rao and Lawrence L Kupper. Ties in Paired-Comparison Experiments: A Generalization of the Bradley-Terry Model. *Journal of the American Statistical Association*, 1967.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Computer Vision and Pattern Recognition*, 2022.

Christoph Schuhmann et al. Laion-aesthetics v2+ dataset. 2022.

Pragya Srivastava, Harman Singh, Rahul Madhavan, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuveer, et al. Robust Reward Modeling via Causal Rubrics. In *International Conference on Learning Representations*, 2026.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion Model Alignment using Direct Preference Optimization. In *Computer Vision and Pattern Recognition*, 2024.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization. In *International Conference on Learning Representations*, 2025.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score: Better Aligning Text-to-Image Models with Human Preference. In *International Conference on Computer Vision*, 2023.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and Evaluating Human Preferences for Text-to-Image Generation. *Neural Information Processing Systems*, 2023.

Yongjin Yang, Sihyeon Kim, Hojung Jung, Sangmin Bae, SangMook Kim, Se-Young Yun, and Kimin Lee. Automated Filtering of Human Feedback Data for Aligning Text-to-Image Diffusion Models. In *International Conference on Learning Representations*, 2025.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Transactions on Machine Learning Research*, 2022.

A Appendix

A.1 Proof of Theorem 1

Theorem 1. *The NTPO objective 11, optimized over the noisy preference dataset $\widetilde{\mathcal{D}}$, is unbiased with respect to the preference distribution in the sense of Definition 1, satisfying Equation 12a.*

Proof. According to the noisy preference transition process, the observed noisy preference label $\widetilde{y}$ is generated from the clean preference label y through the transition matrix Q defined in Equation 13. Therefore, for each preference outcome $j \in \{0, 1, 2\}$, we obtain:

$$\widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) = \sum_{i=0}^{2} q_{j,i} p(y = i|\boldsymbol{z}). \tag{20}$$

The NTPO objective in Equation 11, optimized over the noisy preference distribution $\mathcal{P}^{\eta}$, is

$$\mathcal{L}_{\mathrm{NTPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}^{\eta}} \left[\sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log p^{\mathrm{RK}}_{\boldsymbol{\theta},Q}(\widetilde{y} = j|\boldsymbol{z}) \right]. \tag{21}$$

Using the entropy–KL decomposition, we rewrite the above objective as

$$\begin{aligned} \mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}) =& -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}^{\eta}} \left[\sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) + \sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log \frac{\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta},Q}(\widetilde{y} = j|\boldsymbol{z})}{\widetilde{p}(\widetilde{y} = j|\boldsymbol{z})} \right] \\ =& \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}^{\eta}} \left[-\sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) + \sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log \frac{\widetilde{p}(\widetilde{y} = j|\boldsymbol{z})}{\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta},Q}(\widetilde{y} = j|\boldsymbol{z})} \right] \end{aligned} \tag{22}$$

The first term in Equation 22 is independent of $\boldsymbol{\theta}$. The second term corresponds to the KL divergence between the noisy preference distributions, whose expanded form is

$$\sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log \frac{\widetilde{p}(\widetilde{y} = j|\boldsymbol{z})}{\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta},Q}(\widetilde{y} = j|\boldsymbol{z})}. \tag{23}$$

Therefore, minimizing the NTPO objective is equivalent to minimizing this KL divergence, whose minimum is achieved when

$$\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta}^*,Q}(\widetilde{y} = j|\boldsymbol{z}) = \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}), \quad j \in \{0, 1, 2\}. \tag{24}$$

According to the forward correction formulation in Equation 10, the model prediction satisfies

$$\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta},Q}(\widetilde{y} = j|\boldsymbol{z}) = \sum_{i=0}^{2} q_{j,i} \widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta}}(\widetilde{y} = i|\boldsymbol{z}). \tag{25}$$

Combining Equation 20, Equation 24, and Equation 25, we obtain

$$\sum_{i=0}^{2} q_{j,i} \widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta}^*}(\widetilde{y} = i|\boldsymbol{z}) = \sum_{i=0}^{2} q_{j,i} p(y = i|\boldsymbol{z}), \quad j \in \{0, 1, 2\}. \tag{26}$$

The above equality corresponds to the linear system induced by the transition matrix Q defined in Equation 13. According to the transition matrix defined in Equation 13, its determinant is

$$\begin{aligned} \det(Q) &= \begin{vmatrix} 1-\eta & \eta & 0 \\ \eta & 1-\eta & 0 \\ 0 & 0 & 1 \end{vmatrix} \\ &= (1-\eta)^2 - \eta^2 \\ &= 1 - 2\eta. \end{aligned} \tag{27}$$

Under the assumed noise condition $\eta < 1/2$, we have

$$\det(Q) = 1 - 2\eta > 0, \tag{28}$$

which guarantees that Q is invertible. Hence, multiplying both sides of Eq. equation 26 by Q^{-1} gives

$$\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta}^*}(\widetilde{y} = i|\boldsymbol{z}) = p(y = i|\boldsymbol{z}), \quad i \in \{0, 1, 2\}, \tag{29}$$

which satisfies Equation 12a. Therefore, the NTPO objective optimized over the noisy preference distribution $\mathcal{P}^\eta$ recovers the clean preference distribution. According to Definition 1, the NTPO objective is unbiased. □

A.2 Proof of Theorem 2

Theorem 2. *The TPO objective 8, optimized over the noisy preference dataset $\widetilde{\mathcal{D}}$, is biased with respect to the groundtrue preference distribution in the sense of Definition 1, satisfying Equation 12b.*

Proof. The standard TPO objective in Equation 8 directly treats the observed noisy preference label $\widetilde{y}$ as the groundtruth preference label. Therefore, optimized over the noisy preference distribution $\mathcal{P}^\eta$, its population objective can be written as

$$\mathcal{L}_{\mathrm{TPO}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}^\eta}\left[\sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y = j|\boldsymbol{z})\right]. \tag{30}$$

Similar to the decomposition used in the proof of Theorem 1 in Section A.1, we rewrite the objective as

$$\mathcal{L}_{\mathrm{TPO}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z}\sim\mathcal{P}^\eta}\left[-\sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) + \sum_{j=0}^{2} \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \log \frac{\widetilde{p}(\widetilde{y} = j|\boldsymbol{z})}{p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y = j|\boldsymbol{z})}\right] \tag{31}$$

The first term is independent of $\boldsymbol{\theta}$. Therefore, minimizing the TPO objective is equivalent to minimizing the second term, which achieves its minimum when

$$p^{\mathrm{RK}}_{\boldsymbol{\theta}^*}(y = j|\boldsymbol{z}) = \widetilde{p}(\widetilde{y} = j|\boldsymbol{z}), \quad j \in \{0, 1, 2\}. \tag{32}$$

However, according to the noisy preference transition process defined in Equation 13, the noisy preference distribution is

$$\widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) = \sum_{i=0}^{2} q_{j,i} p(y = i|\boldsymbol{z}). \tag{33}$$

For the transition matrix Q defined in Equation 13, when $\eta > 0$, the noisy preference distribution differs from the clean preference distribution. Specifically,

$$\widetilde{p}(\widetilde{y} = j|\boldsymbol{z}) \neq p(y = j|\boldsymbol{z}), \quad \exists j \in \{0, 1, 2\}. \tag{34}$$

Combining the above inequality with Equation 32, we obtain

$$p^{\mathrm{RK}}_{\boldsymbol{\theta}^*}(y = j|\boldsymbol{z}) \neq p(y = j|\boldsymbol{z}), \quad \exists j \in \{0, 1, 2\}. \tag{35}$$

Since the model optimized by TPO on the noisy preference dataset $\widetilde{\mathcal{D}}$ induces the distribution $\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta}^*}$, we have

$$\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta}^*}(\widetilde{y} = j|\boldsymbol{z}) \neq p(y = j|\boldsymbol{z}), \quad \exists j \in \{0, 1, 2\}. \tag{36}$$

Therefore, the TPO objective optimized over the noisy preference distribution $\mathcal{P}^\eta$ cannot recover the groundtruth preference distribution. According to Definition 1, the TPO objective is biased and satisfies Equation 12b.

□

A.3 Proof of Theorem 3

Theorem 3. *Under the optimal R, the surrogate and noisy preference objectives in Equation 16 and Equation 11 share the same minimum value.*

Proof. According to the forward correction formulation in Equation 10, the noisy preference likelihood is obtained by applying the transition matrix Q to the RK preference probabilities (as formulated in Equation 6 and Equation 7) .

For the strict preference cases, the forward-corrected likelihood can be written as

$$\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta},Q}(\widetilde{y}=0|\boldsymbol{z}) = (1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) + \eta\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha), \tag{37}$$

where $\Delta_{\boldsymbol{\theta}}$ in Equation 3 denotes the preference margin. Therefore, the corresponding log-likelihood term in the NTPO objective is

$$\log\left((1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) + \eta\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)\right). \tag{38}$$

To apply Jensen's inequality with respect to the preference margin $\Delta_{\boldsymbol{\theta}}$, the negative log-likelihood function should be convex. We define

$$f(\Delta_{\boldsymbol{\theta}}) = -\log\left((1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) + \eta\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)\right). \tag{39}$$

For simplicity, we define

$$g(\Delta_{\boldsymbol{\theta}}) = (1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) + \eta\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha). \tag{40}$$

Then,

$$f(\Delta_{\boldsymbol{\theta}}) = -\log g(\Delta_{\boldsymbol{\theta}}). \tag{41}$$

The first derivative of $g(\Delta_{\boldsymbol{\theta}})$ is

$$\begin{aligned} g'(\Delta_{\boldsymbol{\theta}}) =& \frac{\partial g(\Delta_{\boldsymbol{\theta}})}{\partial \Delta_{\boldsymbol{\theta}}} \\ =& (1-\eta)\frac{\partial}{\partial \Delta_{\boldsymbol{\theta}}}\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) + \eta\frac{\partial}{\partial \Delta_{\boldsymbol{\theta}}}\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha) \\ =& (1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)(1-\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)) \\ & - \eta\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)(1-\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)). \end{aligned} \tag{42}$$

Since

$$f(\Delta_{\boldsymbol{\theta}}) = -\log g(\Delta_{\boldsymbol{\theta}}), \tag{43}$$

its first derivative is

$$f'(\Delta_{\boldsymbol{\theta}}) = -\frac{g'(\Delta_{\boldsymbol{\theta}})}{g(\Delta_{\boldsymbol{\theta}})}. \tag{44}$$

The second derivative of $f(\Delta_{\boldsymbol{\theta}})$ is

$$\begin{aligned} f''(\Delta_{\boldsymbol{\theta}}) &= -\frac{g''(\Delta_{\boldsymbol{\theta}})g(\Delta_{\boldsymbol{\theta}}) - (g'(\Delta_{\boldsymbol{\theta}}))^2}{g^2(\Delta_{\boldsymbol{\theta}})} \\ &= \frac{(g'(\Delta_{\boldsymbol{\theta}}))^2 - g''(\Delta_{\boldsymbol{\theta}})g(\Delta_{\boldsymbol{\theta}})}{g^2(\Delta_{\boldsymbol{\theta}})}. \end{aligned} \tag{45}$$

Next, we compute $g''(\Delta_{\boldsymbol{\theta}})$. Since

$$\sigma'(x) = \sigma(x)(1-\sigma(x)), \tag{46}$$

we have

$$\begin{aligned} \sigma''(x) &= \frac{\partial}{\partial x}\left(\sigma(x)(1-\sigma(x))\right) \\ &= \sigma'(x)(1-\sigma(x)) - \sigma(x)\sigma'(x) \\ &= \sigma'(x)(1-2\sigma(x)) \\ &= \sigma(x)(1-\sigma(x))(1-2\sigma(x)). \end{aligned} \tag{47}$$

Therefore,

$$\frac{\partial^2}{\partial \Delta_{\boldsymbol{\theta}}^2}\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) =\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)(1-\sigma(\Delta_{\boldsymbol{\theta}}-\alpha))\cdot(1-2\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)), \tag{48}$$

and

$$\frac{\partial^2}{\partial \Delta_{\boldsymbol{\theta}}^2}\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha) =\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)(1-\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha))\cdot(1-2\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)). \tag{49}$$

Thus,

$$\begin{aligned} g''(\Delta_{\boldsymbol{\theta}}) =&(1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)(1-\sigma(\Delta_{\boldsymbol{\theta}}-\alpha))\cdot(1-2\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)) \\ &+\eta\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)(1-\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha))\cdot(1-2\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)). \end{aligned} \tag{50}$$

Substituting $g'(\Delta_{\boldsymbol{\theta}})$ and $g''(\Delta_{\boldsymbol{\theta}})$ into Equation 45, we obtain

$$f''(\Delta_{\boldsymbol{\theta}}) = \frac{(g'(\Delta_{\boldsymbol{\theta}}))^2 - g''(\Delta_{\boldsymbol{\theta}})g(\Delta_{\boldsymbol{\theta}})}{g^2(\Delta_{\boldsymbol{\theta}})}. \tag{51}$$

Because $g(\Delta_{\boldsymbol{\theta}}) > 0$, the sign of $f''(\Delta_{\boldsymbol{\theta}})$ depends on the numerator. The term $g''(\Delta_{\boldsymbol{\theta}})$ contains

$$1-2\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) \tag{52}$$

and

$$1-2\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha), \tag{53}$$

whose signs vary with $\Delta_{\boldsymbol{\theta}}$. Therefore, the second-order derivative $f''(\Delta_{\boldsymbol{\theta}})$ is not always non-negative, i.e.,

$$f''(\Delta_{\boldsymbol{\theta}}) \ngeq 0. \tag{54}$$

Hence, $f(\Delta_{\boldsymbol{\theta}})$ is not globally convex with respect to $\Delta_{\boldsymbol{\theta}}$, and the direct application of Jensen's inequality is not valid.

For the tie preference case, the forward correction does not introduce a mixture term because the tie label remains unchanged under the transition matrix in Equation 13. Therefore,

$$\widetilde{p}^{\mathrm{RK}}_{\boldsymbol{\theta},Q}(\widetilde{y}=2|\boldsymbol{z}) = p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=2|\boldsymbol{z}). \tag{55}$$

According to the Rao–Kupper preference model, the probability of a tie (as formulated in 7) is

$$p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=2|\boldsymbol{z}) = \sigma\left(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z})+\alpha\right) - \sigma\left(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z})-\alpha\right). \tag{56}$$

For simplicity, we denote $\Delta_{\boldsymbol{\theta}}(\boldsymbol{z})$ as $\Delta_{\boldsymbol{\theta}}$. Then,

$$\begin{aligned} \sigma(\Delta_{\boldsymbol{\theta}}+\alpha)-\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) =&\frac{1}{1+e^{-\Delta_{\boldsymbol{\theta}}-\alpha}} - \frac{1}{1+e^{-\Delta_{\boldsymbol{\theta}}+\alpha}} \\ =&\frac{(1+e^{-\Delta_{\boldsymbol{\theta}}+\alpha})-(1+e^{-\Delta_{\boldsymbol{\theta}}-\alpha})}{(1+e^{-\Delta_{\boldsymbol{\theta}}-\alpha})(1+e^{-\Delta_{\boldsymbol{\theta}}+\alpha})} \\ =&\frac{e^{-\Delta_{\boldsymbol{\theta}}}(e^{\alpha}-e^{-\alpha})}{(1+e^{-\Delta_{\boldsymbol{\theta}}-\alpha})(1+e^{-\Delta_{\boldsymbol{\theta}}+\alpha})} \\ =&(e^{2\alpha}-1)\frac{e^{-\Delta_{\boldsymbol{\theta}}-\alpha}}{(1+e^{-\Delta_{\boldsymbol{\theta}}-\alpha})(1+e^{-\Delta_{\boldsymbol{\theta}}+\alpha})} \\ =&(e^{2\alpha}-1)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha). \end{aligned} \tag{57}$$

Therefore, the tie probability can be rewritten as

$$p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=2|\boldsymbol{z}) = (e^{2\alpha}-1)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha). \tag{58}$$

The corresponding negative log-likelihood term is

$$\begin{aligned} f_{\mathrm{tie}}(\Delta_{\boldsymbol{\theta}}) =& -\log\left(p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=2|\boldsymbol{z})\right) \\ =& -\log\left((e^{2\alpha}-1)\sigma(\Delta_{\boldsymbol{\theta}}-\alpha)\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha)\right) \\ =& -\log(e^{2\alpha}-1) - \log\sigma(\Delta_{\boldsymbol{\theta}}-\alpha) - \log\sigma(-\Delta_{\boldsymbol{\theta}}-\alpha). \end{aligned} \tag{59}$$

Since α is a constant, the first term in Equation 59 is independent of $\Delta_{\boldsymbol{\theta}}$. Therefore, it does not affect the convexity of $f_{\text{tie}}(\Delta_{\boldsymbol{\theta}})$. For the sigmoid function, we have

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)). \tag{60}$$

The first derivative of the negative log-sigmoid function is

$$\begin{aligned}\frac{\partial}{\partial x}[-\log \sigma(x)] &= -\frac{\sigma'(x)}{\sigma(x)} \\ &= -\frac{\sigma(x)(1-\sigma(x))}{\sigma(x)} \\ &= -(1 - \sigma(x)).\end{aligned} \tag{61}$$

Taking the second derivative gives

$$\frac{\partial^2}{\partial x^2}[-\log \sigma(x)] = \sigma'(x) = \sigma(x)(1 - \sigma(x)). \tag{62}$$

Therefore,

$$\frac{\partial^2}{\partial \Delta_{\boldsymbol{\theta}}^2}[-\log \sigma(\Delta_{\boldsymbol{\theta}} - \alpha)] = \sigma(\Delta_{\boldsymbol{\theta}} - \alpha)(1 - \sigma(\Delta_{\boldsymbol{\theta}} - \alpha)) \geq 0, \tag{63}$$

and

$$\frac{\partial^2}{\partial \Delta_{\boldsymbol{\theta}}^2}[-\log \sigma(-\Delta_{\boldsymbol{\theta}} - \alpha)] = \sigma(-\Delta_{\boldsymbol{\theta}} - \alpha)(1 - \sigma(-\Delta_{\boldsymbol{\theta}} - \alpha)) \geq 0. \tag{64}$$

Combining the above results, the second-order derivative of the tie loss is

$$\frac{\partial^2 f_{\text{tie}}(\Delta_{\boldsymbol{\theta}})}{\partial \Delta_{\boldsymbol{\theta}}^2} = \sigma(\Delta_{\boldsymbol{\theta}} - \alpha)(1 - \sigma(\Delta_{\boldsymbol{\theta}} - \alpha)) + \sigma(-\Delta_{\boldsymbol{\theta}} - \alpha)(1 - \sigma(-\Delta_{\boldsymbol{\theta}} - \alpha)) \geq 0. \tag{65}$$

Therefore, $f_{\text{tie}}(\Delta_{\boldsymbol{\theta}})$ is convex with respect to the preference margin $\Delta_{\boldsymbol{\theta}}$. Hence, the tie preference case satisfies the convexity condition required by Jensen's inequality.

Thus, the difficulty arises from the noisy strict-preference log-mixture, rather than from the tie term.

We now construct the variational surrogate using the variational weight matrix R defined in Equation 15. We first explicitly examine the surrogate loss for the strict preference and tie cases.

For a strict preference, we consider $\widetilde{y}_n = 0$ without loss of generality, since the case $\widetilde{y}_n = 1$ is symmetric. According to the transition matrix Q in Equation 13 and the variational weight matrix R in Equation 15, we have

$$q_{0,0} = 1 - \eta, \qquad q_{0,1} = \eta, \qquad r_{0,0} = 1 - \rho, \qquad r_{0,1} = \rho. \tag{66}$$

Moreover, according to the RK preference probability in Equation 6,

$$p_{\boldsymbol{\theta}}^{\text{RK}}(y = 0 \mid \boldsymbol{z}_n) = \sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha)\,, \tag{67}$$

$$p_{\boldsymbol{\theta}}^{\text{RK}}(y = 1 \mid \boldsymbol{z}_n) = \sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha)\,. \tag{68}$$

Substituting these expressions into Equation 16, the strict-preference surrogate term can be written as

$$\begin{aligned}f_{\mathsf{VFC}}^{\text{strict}}(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n); \rho) = &-(1-\rho)\log \sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha) \\ &- \rho \log \sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha) \\ &+ (1-\rho)\log\frac{1-\rho}{1-\eta} + \rho \log \frac{\rho}{\eta}.\end{aligned} \tag{69}$$

The last two terms in Equation 69 are independent of $\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)$. To examine convexity, we use

$$\frac{\partial}{\partial x}\left[-\log \sigma(x)\right] = -\sigma(-x), \tag{70}$$

and

$$\frac{\partial^2}{\partial x^2}\left[-\log\sigma(x)\right] = \sigma(x)\sigma(-x) \geq 0. \tag{71}$$

Therefore,

$$\begin{aligned}&\frac{\partial^2}{\partial \Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)^2}\left[-(1-\rho)\log\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\right]\\ =&(1-\rho)\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)+\alpha) \geq 0,\end{aligned} \tag{72}$$

and

$$\begin{aligned}&\frac{\partial^2}{\partial \Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)^2}\left[-\rho\log\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\right]\\ =&\rho\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)+\alpha) \geq 0.\end{aligned} \tag{73}$$

Hence,

$$\begin{aligned}&\frac{\partial^2 f_{\mathsf{VFC}}^{\text{strict}}\left(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n);\rho\right)}{\partial \Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)^2}\\ =&(1-\rho)\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)+\alpha)\\ &+\rho\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)+\alpha)\\ \geq&0,\end{aligned} \tag{74}$$

where $0 \leq \rho \leq 1$. Therefore, the VFC surrogate for a strict preference is globally convex with respect to $\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)$.

For the tie case $\widetilde{y}_n = 2$, the transition matrix Equation 13 and the variational matrix Equation 15 satisfy

$$q_{2,2} = r_{2,2} = 1, \qquad q_{2,0} = q_{2,1} = r_{2,0} = r_{2,1} = 0. \tag{75}$$

Thus, the variational surrogate reduces exactly to the negative log-likelihood of the RK tie probability. According to Equation 7,

$$\begin{aligned}p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=2 \mid \boldsymbol{z}_n) =&\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)+\alpha)-\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\\ =&\left(e^{2\alpha}-1\right)\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,.\end{aligned} \tag{76}$$

Therefore, the tie surrogate term is

$$\begin{aligned}f_{\mathsf{VFC}}^{\text{tie}}\left(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)\right) =& -\log p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=2 \mid \boldsymbol{z}_n)\\ =& -\log\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\\ &-\log\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\\ &-\log\left(e^{2\alpha}-1\right).\end{aligned} \tag{77}$$

Since the last term in Equation 77 is independent of $\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)$, applying Equation 71 gives

$$\begin{aligned}&\frac{\partial^2 f_{\mathsf{VFC}}^{\text{tie}}\left(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)\right)}{\partial \Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)^2}\\ =&\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)+\alpha)\\ &+\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha)\,\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)+\alpha)\\ \geq&0.\end{aligned} \tag{78}$$

Therefore, both the strict-preference surrogate in Equation 69 and the tie surrogate in Equation 77 are globally convex with respect to the preference margin $\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)$. Consequently, the complete VFC surrogate induced by Equation 16 is convex with respect to $\Delta_{\boldsymbol{\theta}}$, and Jensen's inequality can be validly applied as in Equation 17.

We now show that the NTPO objective is upper-bounded by the variational surrogate objective. For the n-th training sample, the corresponding term of the NTPO objective in Equation 11 is

$$\ell_{\mathsf{NTPO},n}(\boldsymbol{\theta}) = -\log\left(\sum_{i=0}^{2} q_{\widetilde{y}_n,i}\, p_{\boldsymbol{\theta}}^{\mathrm{RK}}\left(y=i \mid \boldsymbol{z}_n\right)\right). \tag{79}$$

According to the variational weight matrix R defined in Equation 15, for each observed label $\widetilde{y}_n$, the corresponding row satisfies

$$r_{\widetilde{y}_n,i} \geq 0, \qquad \sum_{i=0}^{2} r_{\widetilde{y}_n,i} = 1. \tag{80}$$

For the indices with positive variational weights, we can rewrite the mixture probability in Equation 79 as

$$\begin{aligned} &\sum_{i=0}^{2} q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z}_n) \\ &= \sum_{i=0}^{2} r_{\widetilde{y}_n,i} \frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z}_n)}{r_{\widetilde{y}_n,i}}. \end{aligned} \tag{81}$$

For the zero entries of R, the corresponding entries of Q are also zero under Equation 13 and Equation 15, so these terms make no contribution to either side of Equation 81. Since $-\log(\cdot)$ is convex on the positive real domain, i.e.,

$$\frac{\partial^2}{\partial x^2}[-\log x] = \frac{1}{x^2} > 0, \qquad x > 0, \tag{82}$$

we can apply Jensen's inequality to Equation 81. Using Equation 80, we obtain

$$\begin{aligned} &-\log\left(\sum_{i=0}^{2} r_{\widetilde{y}_n,i} \frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z}_n)}{r_{\widetilde{y}_n,i}}\right) \\ &\leq -\sum_{i=0}^{2} r_{\widetilde{y}_n,i} \log \frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z}_n)}{r_{\widetilde{y}_n,i}}. \end{aligned} \tag{83}$$

Substituting Equation 81 into the left-hand side gives

$$\begin{aligned} &-\log\left(\sum_{i=0}^{2} q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z}_n)\right) \\ &\leq -\sum_{i=0}^{2} r_{\widetilde{y}_n,i} \log \frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y = i \mid \boldsymbol{z}_n)}{r_{\widetilde{y}_n,i}}. \end{aligned} \tag{84}$$

The left-hand side of Equation 84 is exactly the sample-wise NTPO loss in Equation 79, whereas the right-hand side is the sample-wise variational surrogate in Equation 16. Therefore,

$$\ell_{\mathsf{NTPO},n}(\boldsymbol{\theta}) \leq \ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R). \tag{85}$$

Averaging the above inequality over all N training samples yields

$$\begin{aligned} \mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{n=1}^{N} \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta}) \\ &\leq \frac{1}{N}\sum_{n=1}^{N} \ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R) \\ &= \mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R). \end{aligned} \tag{86}$$

Hence,

$$\boxed{\mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}) \leq \mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R)}. \tag{87}$$

Therefore, Equation 16 provides a variational upper bound on the original noisy preference objective in Equation 11.

We next derive the equality condition of the variational upper bound in Equation 84. Recall that Jensen's inequality becomes an equality if all arguments of the convex function with positive weights are identical. Therefore, equality in Equation 83 holds when

$$\frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=i \mid \boldsymbol{z}_n)}{r_{\widetilde{y}_n,i}} = \frac{q_{\widetilde{y}_n,k} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=k \mid \boldsymbol{z}_n)}{r_{\widetilde{y}_n,k}}, \tag{88}$$

for all i and k such that $r_{\widetilde{y}_n,i} > 0$ and $r_{\widetilde{y}_n,k} > 0$.

Hence, there exists a common positive value such that

$$r_{\widetilde{y}_n,i} \propto q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=i \mid \boldsymbol{z}_n). \tag{89}$$

Using the normalization condition

$$\sum_{i=0}^{2} r_{\widetilde{y}_n,i} = 1, \tag{90}$$

the optimal variational weights are

$$r^*_{\widetilde{y}_n,i} = \frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=i \mid \boldsymbol{z}_n)}{\sum_{k=0}^{2} q_{\widetilde{y}_n,k} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=k \mid \boldsymbol{z}_n)}, \qquad i \in \{0,1,2\}. \tag{91}$$

We now explicitly verify the optimal variational weights for the strict preference case. For $\widetilde{y}_n = 0$, according to Equation 13,

$$q_{0,0} = 1-\eta, \qquad q_{0,1} = \eta, \qquad q_{0,2} = 0. \tag{92}$$

Combining this with the RK probabilities in Equation 6, we obtain

$$r^*_{0,0} = \frac{(1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha)}{(1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha) + \eta\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha)}, \tag{93}$$

$$r^*_{0,1} = \frac{\eta\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha)}{(1-\eta)\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha) + \eta\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n) - \alpha)}, \tag{94}$$

$$r^*_{0,2} = 0. \tag{95}$$

According to the parameterization of R in Equation 15, this gives

$$1-\rho^* = r^*_{0,0}, \qquad \rho^* = r^*_{0,1}. \tag{96}$$

The case $\widetilde{y}_n = 1$ follows symmetrically. For the tie case $\widetilde{y}_n = 2$, since

$$q_{2,0} = q_{2,1} = 0, \qquad q_{2,2} = 1, \tag{97}$$

the optimal variational weights are directly given by

$$r^*_{2,0} = r^*_{2,1} = 0, \qquad r^*_{2,2} = 1, \tag{98}$$

which is identical to the third row of R in Equation 15.

We next show that the gap between the variational surrogate and the original NTPO loss vanishes under the optimal R. For the n-th sample, the gap is

$$\begin{aligned} &\ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R) - \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta}) \\ &= -\sum_{i=0}^{2} r_{\widetilde{y}_n,i} \log \frac{q_{\widetilde{y}_n,i} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=i \mid \boldsymbol{z}_n)}{r_{\widetilde{y}_n,i}} \\ &\quad + \log\left(\sum_{k=0}^{2} q_{\widetilde{y}_n,k} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=k \mid \boldsymbol{z}_n)\right). \end{aligned} \tag{99}$$

Using $\sum_{i=0}^{2} r_{\widetilde{y}_n,i} = 1$, the second term can be written as

$$\begin{aligned} &\log\left(\sum_{k=0}^{2} q_{\widetilde{y}_n,k} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=k \mid \boldsymbol{z}_n)\right) \\ &= \sum_{i=0}^{2} r_{\widetilde{y}_n,i} \log\left(\sum_{k=0}^{2} q_{\widetilde{y}_n,k} p_{\boldsymbol{\theta}}^{\mathrm{RK}}(y=k \mid \boldsymbol{z}_n)\right). \end{aligned} \tag{100}$$

Therefore, Equation 99 becomes

$$\begin{aligned}&\ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R) - \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta})\\ =&\sum_{i=0}^{2} r_{\widetilde{y}_n,i} \log \frac{r_{\widetilde{y}_n,i} \sum_{k=0}^{2} q_{\widetilde{y}_n,k} p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=k \mid \boldsymbol{z}_n)}{q_{\widetilde{y}_n,i} p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=i \mid \boldsymbol{z}_n)}.\end{aligned} \tag{101}$$

Substituting the optimal variational weights in Equation 91, we have

$$\frac{r^{*}_{\widetilde{y}_n,i} \sum_{k=0}^{2} q_{\widetilde{y}_n,k} p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=k \mid \boldsymbol{z}_n)}{q_{\widetilde{y}_n,i} p^{\mathrm{RK}}_{\boldsymbol{\theta}}(y=i \mid \boldsymbol{z}_n)} = 1. \tag{102}$$

Hence,

$$\begin{aligned}&\ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R^*) - \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta})\\ =&\sum_{i=0}^{2} r^{*}_{\widetilde{y}_n,i} \log 1\\ =&0.\end{aligned} \tag{103}$$

Therefore,

$$\ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R^*) = \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta}). \tag{104}$$

Equivalently, the expression in Equation 101 is the expanded form of the KL divergence between the variational weights and the posterior weights in Equation 91. Since the KL divergence is non-negative, we have

$$\ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R) - \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta}) \geq 0, \tag{105}$$

with equality if and only if

$$R = R^*. \tag{106}$$

Thus, the variational upper bound is tight under the optimal R.

Since the variational upper bound is tight for every training sample under the optimal R, as shown in Equation 104, we have

$$\ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R^*) = \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta}), \qquad n = 1, \ldots, N. \tag{107}$$

Averaging the above equality over all N samples yields

$$\begin{aligned}\mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R^*) &= \frac{1}{N} \sum_{n=1}^{N} \ell^{\eta}_{\mathsf{VFC},n}(\boldsymbol{\theta}, R^*)\\ &= \frac{1}{N} \sum_{n=1}^{N} \ell_{\mathsf{NTPO},n}(\boldsymbol{\theta})\\ &= \mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}).\end{aligned} \tag{108}$$

On the other hand, from the variational upper bound in Equation 87, for any admissible R,

$$\mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}) \leq \mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R). \tag{109}$$

Combining this inequality with Equation 108, we obtain

$$\mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R^*) = \mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}) \leq \mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R), \tag{110}$$

for any admissible R. Therefore, R^* minimizes the variational surrogate for a fixed $\boldsymbol{\theta}$, and

$$\min_{R} \mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R) = \mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}). \tag{111}$$

Taking the minimum with respect to $\boldsymbol{\theta}$ on both sides of Equation 111 gives

$$\begin{aligned}\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathsf{NTPO}}(\boldsymbol{\theta}) &= \min_{\boldsymbol{\theta}} \left[\min_{R} \mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R)\right] \\ &= \min_{\boldsymbol{\theta}, R} \mathcal{L}^{\eta}_{\mathsf{VFC}}(\boldsymbol{\theta}, R). \end{aligned} \tag{112}$$

Hence, under the optimal variational matrix R^*, the upper bound is tight not only for each individual sample but also for the entire noisy preference dataset. Consequently, the variational surrogate objective in Equation 16 and the original noisy preference objective in Equation 11 share the same minimum value, which proves Theorem 3. □

A.4 Derivation of the VFC Objectives for Diffusion and Flow Matching Models

Since preference datasets are typically constructed with a fixed label order, where the preferred generation is placed first and the dispreferred generation second, our training objective only considers the strict preference case $\boldsymbol{x}_n^1 \succ \boldsymbol{x}_n^2$ and the tie case $\boldsymbol{x}_n^1 \sim \boldsymbol{x}_n^2$. We use the indicator $\widetilde{\mathbb{I}}_n$, where $\widetilde{\mathbb{I}}_n = 0$ denotes a strict preference and $\widetilde{\mathbb{I}}_n = 1$ denotes a tie.

For completeness, we briefly introduce the diffusion and flow matching formulations used in the following derivation. For each candidate generation $\boldsymbol{x}_0^j$, where $j \in \{1, 2\}$, we denote its intermediate state at timestep t by $\boldsymbol{x}_t^j$.

Diffusion Models. A diffusion model gradually perturbs a clean sample $\boldsymbol{x}_0^j$ with Gaussian noise according to the forward process

$$q(\boldsymbol{x}_t^j \mid \boldsymbol{x}_0^j) = \mathcal{N}\left(\alpha_t \boldsymbol{x}_0^j, \sigma_t^2 \boldsymbol{I}\right), \tag{113}$$

which admits the reparameterization

$$\boldsymbol{x}_t^j = \alpha_t \boldsymbol{x}_0^j + \sigma_t \boldsymbol{\epsilon}^j, \qquad \boldsymbol{\epsilon}^j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{114}$$

The reverse process is parameterized by $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^j \mid \boldsymbol{x}_t^j, \boldsymbol{c}_n)$. Under the noise-prediction parameterization,

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^j = \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^j, t, \boldsymbol{c}_n) \tag{115}$$

denotes the noise predicted by the trainable model. The standard diffusion training objective is

$$\mathcal{L}_{\mathrm{Diff}}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{c}, \boldsymbol{x}_0) \sim \mathcal{D},\, t,\, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[\omega_t \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c})\right\|_2^2\right], \tag{116}$$

where ω_t denotes the timestep-dependent weighting coefficient determined by the diffusion process.

We further denote the noise prediction of the reference model by

$$\boldsymbol{\epsilon}_{\mathrm{ref}}^j = \boldsymbol{\epsilon}_{\mathrm{ref}}(\boldsymbol{x}_t^j, t, \boldsymbol{c}_n). \tag{117}$$

Flow Matching Models. Flow matching constructs a time-dependent probability path connecting a simple source distribution to the data distribution. A conditional path can be written as

$$\boldsymbol{x}_t^j = \alpha_t \boldsymbol{x}_0^j + \sigma_t \boldsymbol{\epsilon}^j, \qquad \boldsymbol{\epsilon}^j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{118}$$

with the corresponding conditional target velocity

$$\boldsymbol{v}^j = \dot{\alpha}_t \boldsymbol{x}_0^j + \dot{\sigma}_t \boldsymbol{\epsilon}^j. \tag{119}$$

The trainable velocity field is denoted by

$$\boldsymbol{v}_{\boldsymbol{\theta}}^j = \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t^j, t, \boldsymbol{c}_n). \tag{120}$$

The standard flow matching objective minimizes the squared error between the predicted and target velocities:

$$\mathcal{L}_{\mathrm{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{c}, \boldsymbol{x}_0) \sim \mathcal{D},\, t,\, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left[\left\|\boldsymbol{v} - \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c})\right\|_2^2\right]. \tag{121}$$

Similarly, we denote the velocity predicted by the reference model by

$$\boldsymbol{v}_{\mathrm{ref}}^{j} = \boldsymbol{v}_{\mathrm{ref}}(\boldsymbol{x}_t^j, t, \boldsymbol{c}_n). \tag{122}$$

In the following, T denotes the total number of timesteps used in the trajectory decomposition, and $\beta > 0$ denotes the preference-strength coefficient. These formulations allow the KL-based timestep preference margin to be instantiated using the noise-prediction error for diffusion models and the velocity-prediction error for flow matching models.

According to the variational objective in Equation 16 and the variational matrix R in Equation 15, the two cases can be combined into the following unified objective:

$$\begin{aligned}
&\mathcal{L}_{\mathsf{VFC}}^{\eta}(\boldsymbol{\theta}, \rho) \\
=&\mathbb{E}_{(\boldsymbol{z}_n, \widetilde{\mathbb{I}}_n)\sim\mathcal{P}^{\eta}}\Bigg[(1-\widetilde{\mathbb{I}}_n)\Bigg(-(1-\rho)\log\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha) \\
&\qquad -\rho\log\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha) + (1-\rho)\log\frac{1-\rho}{1-\eta} + \rho\log\frac{\rho}{\eta}\Bigg) \\
&+\widetilde{\mathbb{I}}_n\Bigg(-\log\sigma(\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha) - \log\sigma(-\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)-\alpha) - \log\big(e^{2\alpha}-1\big)\Bigg)\Bigg].
\end{aligned} \tag{123}$$

We next instantiate the preference margin $\Delta_{\boldsymbol{\theta}}(\boldsymbol{z}_n)$ using the complete generative trajectories. For each candidate $\boldsymbol{x}_0^j$, $j \in \{1,2\}$, let $\boldsymbol{x}_{0:T}^j = (\boldsymbol{x}_0^j, \ldots, \boldsymbol{x}_T^j)$ denote its complete trajectory. Following Equation 3, we define the corresponding trajectory-level preference margin as

$$\Delta_{\boldsymbol{\theta}}^{\mathrm{traj}}(\boldsymbol{z}_n) = \beta\mathbb{E}_{\substack{\boldsymbol{x}_{1:T}^1\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0^1)\\ \boldsymbol{x}_{1:T}^2\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0^2)}}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T}^1 \mid \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{0:T}^1 \mid \boldsymbol{c}_n)} - \log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T}^2 \mid \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{0:T}^2 \mid \boldsymbol{c}_n)}\right]. \tag{124}$$

The reverse processes of the target and reference models factorize as

$$\begin{aligned}
p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}) &= p(\boldsymbol{x}_T)\prod_{s=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{s-1} \mid \boldsymbol{x}_s, \boldsymbol{c}), \\
p_{\mathrm{ref}}(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}) &= p(\boldsymbol{x}_T)\prod_{s=1}^{T} p_{\mathrm{ref}}(\boldsymbol{x}_{s-1} \mid \boldsymbol{x}_s, \boldsymbol{c}).
\end{aligned} \tag{125}$$

Since the terminal distribution $p(\boldsymbol{x}_T)$ is shared by the target and reference models, it cancels in the likelihood ratio. Therefore,

$$\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{0:T}^j \mid \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{0:T}^j \mid \boldsymbol{c}_n)} = \sum_{s=1}^{T}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{s-1}^j \mid \boldsymbol{x}_s^j, \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{s-1}^j \mid \boldsymbol{x}_s^j, \boldsymbol{c}_n)}. \tag{126}$$

Substituting Equation 126 into Equation 124 gives

$$\Delta_{\boldsymbol{\theta}}^{\mathrm{traj}}(\boldsymbol{z}_n) = \beta\mathbb{E}_{\substack{\boldsymbol{x}_{1:T}^1\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0^1)\\ \boldsymbol{x}_{1:T}^2\sim q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0^2)}}\left[\sum_{s=1}^{T}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{s-1}^1 \mid \boldsymbol{x}_s^1, \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{s-1}^1 \mid \boldsymbol{x}_s^1, \boldsymbol{c}_n)} - \sum_{s=1}^{T}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{s-1}^2 \mid \boldsymbol{x}_s^2, \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{s-1}^2 \mid \boldsymbol{x}_s^2, \boldsymbol{c}_n)}\right]. \tag{127}$$

The expectation over the complete trajectory can further be decomposed using the conditional posterior $q(\boldsymbol{x}_{s-1} \mid \boldsymbol{x}_s, \boldsymbol{x}_0)$. Hence,

$$\begin{aligned}
\Delta_{\boldsymbol{\theta}}^{\mathrm{traj}}(\boldsymbol{z}_n) =&\beta\sum_{s=1}^{T}\mathbb{E}_{\substack{\boldsymbol{x}_s^1\sim q(\boldsymbol{x}_s|\boldsymbol{x}_0^1)\\ \boldsymbol{x}_s^2\sim q(\boldsymbol{x}_s|\boldsymbol{x}_0^2)}}\Bigg[\mathbb{E}_{q(\boldsymbol{x}_{s-1}^1|\boldsymbol{x}_s^1,\boldsymbol{x}_0^1)}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{s-1}^1 \mid \boldsymbol{x}_s^1, \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{s-1}^1 \mid \boldsymbol{x}_s^1, \boldsymbol{c}_n)} \\
&\qquad -\mathbb{E}_{q(\boldsymbol{x}_{s-1}^2|\boldsymbol{x}_s^2,\boldsymbol{x}_0^2)}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{s-1}^2 \mid \boldsymbol{x}_s^2, \boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{s-1}^2 \mid \boldsymbol{x}_s^2, \boldsymbol{c}_n)}\Bigg].
\end{aligned} \tag{128}$$

Sampling $t \sim \mathrm{Unif}\{1,\dots,T\}$, the summation over the complete trajectory can equivalently be expressed as a single-timestep expectation:

$$\Delta^{\mathrm{traj}}_{\boldsymbol{\theta}}(\boldsymbol{z}_n) = \beta T \mathbb{E}_{\substack{t\sim\mathrm{Unif}\{1,\dots,T\}\\ \boldsymbol{x}_t^1\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^1)\\ \boldsymbol{x}_t^2\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^2)}}\Bigg[\mathbb{E}_{q(\boldsymbol{x}_{t-1}^1|\boldsymbol{x}_t^1,\boldsymbol{x}_0^1)}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^1\mid\boldsymbol{x}_t^1,\boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{t-1}^1\mid\boldsymbol{x}_t^1,\boldsymbol{c}_n)} - \mathbb{E}_{q(\boldsymbol{x}_{t-1}^2|\boldsymbol{x}_t^2,\boldsymbol{x}_0^2)}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^2\mid\boldsymbol{x}_t^2,\boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{t-1}^2\mid\boldsymbol{x}_t^2,\boldsymbol{c}_n)}\Bigg]. \tag{129}$$

Each conditional transition log-ratio in Equation 129 can be expressed as a difference of KL divergences. Specifically, for $j \in \{1,2\}$,

$$\begin{aligned}&\mathbb{E}_{q(\boldsymbol{x}_{t-1}^j|\boldsymbol{x}_t^j,\boldsymbol{x}_0^j)}\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{c}_n)}{p_{\mathrm{ref}}(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{c}_n)} \qquad (130)\\ =&-D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{x}_0^j)\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{c}_n)\right)+D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{x}_0^j)\|p_{\mathrm{ref}}(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{c}_n)\right).\end{aligned}$$

For compactness, let

$$\begin{aligned}q_t^j &= q(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{x}_0^j),\\ p_{\boldsymbol{\theta},t}^j &= p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{c}_n),\\ p_{\mathrm{ref},t}^j &= p_{\mathrm{ref}}(\boldsymbol{x}_{t-1}^j\mid\boldsymbol{x}_t^j,\boldsymbol{c}_n).\end{aligned} \tag{131}$$

Then the single-timestep preference margin becomes

$$\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t} = -\beta T\left[D_{\mathrm{KL}}\left(q_t^1\|p_{\boldsymbol{\theta},t}^1\right)-D_{\mathrm{KL}}\left(q_t^1\|p_{\mathrm{ref},t}^1\right)-D_{\mathrm{KL}}\left(q_t^2\|p_{\boldsymbol{\theta},t}^2\right)+D_{\mathrm{KL}}\left(q_t^2\|p_{\mathrm{ref},t}^2\right)\right], \tag{132}$$

such that

$$\Delta^{\mathrm{traj}}_{\boldsymbol{\theta}}(\boldsymbol{z}_n) = \mathbb{E}_{\substack{t\sim\mathrm{Unif}\{1,\dots,T\}\\ \boldsymbol{x}_t^1\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^1)\\ \boldsymbol{x}_t^2\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^2)}}\left[\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t}\right]. \tag{133}$$

Substituting $\Delta^{\mathrm{traj}}_{\boldsymbol{\theta}}$ into Equation 123 gives the trajectory-level VFC objective. Since both the strict-preference and tie terms of the VFC surrogate are convex with respect to the preference margin, as established in Equation 74 and Equation 78, Jensen's inequality in Equation 17 gives

$$\begin{aligned}&\mathcal{L}^{\eta}_{\mathsf{VFC\text{-}RK},\mathsf{traj}}(\boldsymbol{\theta},\rho) \le \mathbb{E}_{(\boldsymbol{c}_n,\boldsymbol{x}_0^1,\boldsymbol{x}_0^2,\widetilde{\mathbb{I}}_n)\sim\mathcal{P}^{\eta},\ t\sim\mathcal{U}\{1,\dots,T\},\ \boldsymbol{x}_t^1,\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^1),\ \boldsymbol{x}_t^2\sim q(\boldsymbol{x}_t|\boldsymbol{x}_0^2)}\\ &\Bigg[(1-\widetilde{\mathbb{I}}_n)\left(-(1-\rho)\log\sigma\big(\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t}-\alpha\big)-\rho\log\sigma\big(-\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t}-\alpha\big)+(1-\rho)\log\frac{1-\rho}{1-\eta}+\rho\log\frac{\rho}{\eta}\right)\\ &\quad+\widetilde{\mathbb{I}}_n\left(-\log\sigma\big(\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t}-\alpha\big)-\log\sigma\big(-\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t}-\alpha\big)-\log\big(e^{2\alpha}-1\big)\right)\Bigg].\end{aligned} \tag{134}$$

Equation equation 134 provides a common single-timestep VFC objective for both diffusion and flow matching models. The two model classes differ only in how the KL-based timestep margin $\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t}$ is instantiated by their respective training parameterizations.

For diffusion models, minimizing

$$D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{c})\right) \tag{135}$$

is equivalent to minimizing the noise-prediction error

$$\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t,\boldsymbol{c})\|_2^2. \tag{136}$$

Therefore, the KL-based timestep preference margin can be written as

$$\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t} = \Delta^{\mathsf{Diff}}_{\boldsymbol{\theta},t} = -\beta T \omega_t \left[\left\|\boldsymbol{\epsilon}^1 - \boldsymbol{\epsilon}^1_{\boldsymbol{\theta}}\right\|_2^2 - \left\|\boldsymbol{\epsilon}^1 - \boldsymbol{\epsilon}^1_{\mathrm{ref}}\right\|_2^2 - \left\|\boldsymbol{\epsilon}^2 - \boldsymbol{\epsilon}^2_{\boldsymbol{\theta}}\right\|_2^2 + \left\|\boldsymbol{\epsilon}^2 - \boldsymbol{\epsilon}^2_{\mathrm{ref}}\right\|_2^2 \right], \quad (137)$$

where ω_t denotes the timestep-dependent weighting coefficient.

For flow matching models, minimizing

$$D_{\mathrm{KL}}\left(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \| p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c})\right) \quad (138)$$

is equivalent to minimizing the velocity-prediction error

$$\left\|\boldsymbol{v} - \boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c})\right\|_2^2 . \quad (139)$$

Accordingly, the KL-based timestep preference margin becomes

$$\Delta^{\mathrm{KL}}_{\boldsymbol{\theta},t} = \Delta^{\mathsf{Flow}}_{\boldsymbol{\theta},t} = -\beta T \left[\left\|\boldsymbol{v}^1 - \boldsymbol{v}^1_{\boldsymbol{\theta}}\right\|_2^2 - \left\|\boldsymbol{v}^1 - \boldsymbol{v}^1_{\mathrm{ref}}\right\|_2^2 - \left\|\boldsymbol{v}^2 - \boldsymbol{v}^2_{\boldsymbol{\theta}}\right\|_2^2 + \left\|\boldsymbol{v}^2 - \boldsymbol{v}^2_{\mathrm{ref}}\right\|_2^2 \right]. \quad (140)$$

A.5 Preference Label Auditing with HPSv3 and Cleanlab

A.5.1 Setup

We consider pairwise human preference annotations from Pick-a-Pic v2. Each example consists of a text prompt c, two generated images (I_0, I_1), and a human preference encoded by `label_0`. We retain only *decisive* comparisons with `label_0` $\in \{0, 1\}$ and exclude ties (`label_0` $= 0.5$). We map annotations to a binary label $y \in \{0, 1\}$ used throughout Cleanlab:

$$y = 0 \;\Leftrightarrow\; \text{human prefers } I_0 \quad (\texttt{label_0} = 1), \tag{141}$$

$$y = 1 \;\Leftrightarrow\; \text{human prefers } I_1 \quad (\texttt{label_0} = 0). \tag{142}$$

A.6 Pairwise Probabilities from HPSv3

We use **HPSv3** Ma et al., 2025 as an external human-preference model. For each pair (c, I_0, I_1), HPSv3 assigns scalar scores s_0 and s_1 by running its reward inferencer on (c, I_0) and (c, I_1) separately. Following the official API, each score is the mean μ of the model's preference logit for that image–prompt pair.

Cleanlab requires a predicted class-probability matrix $\hat{\mathbf{p}}_i \in \Delta^1$ for each example i, where $\hat{p}_{i,0} = P(\text{prefer } I_0 \mid c_i)$ and $\hat{p}_{i,1} = P(\text{prefer } I_1 \mid c_i)$. We convert $(s_{i,0}, s_{i,1})$ into $\hat{\mathbf{p}}_i$ via a two-way softmax:

$$\ell_{i,c} = \frac{\alpha}{T}\, s_{i,c}, \quad c \in \{0, 1\}, \tag{143}$$

$$\hat{p}_{i,c} = \frac{\exp\big(\ell_{i,c} - \max_{c' \in \{0,1\}} \ell_{i,c'}\big)}{\sum_{c' \in \{0,1\}} \exp\big(\ell_{i,c'} - \max_{c'' \in \{0,1\}} \ell_{i,c''}\big)}, \tag{144}$$

where $T > 0$ is a temperature and $\alpha > 0$ is a logit scale. In our full-scale audit we set $T = 1$ and $\alpha = 1$, reflecting that HPSv3's μ is already on a usable preference scale. The vector $\hat{\mathbf{p}}_i$ is *not* produced by training a classifier on Pick-a-Pic; it is the preference distribution implied by a fixed HPSv3 prior.

A.6.1 Cleanlab Label-Issue Detection

Given aligned pairs $(y_i, \hat{\mathbf{p}}_i)_{i=1}^N$, we apply **Cleanlab** Northcutt et al. (2021) to identify likely label errors and rank examples by label quality.

Flagging issues. We call `find_label_issues` with human labels $\mathbf{y}$ and predicted probabilities $\hat{\mathbf{P}} = [\hat{\mathbf{p}}_1; \ldots; \hat{\mathbf{p}}_N]^\top$, using `filter_by=both`, i.e., Cleanlab's combined pruning criterion over noise-rate and class-conditional views of suspicious labels. This yields a boolean mask `is_label_issue`$_i$ for each example.

Quality scores. We compute `label_quality_score`$_i$ via `get_label_quality_scores` with `method=self_confidence`: higher scores indicate stronger agreement between the given label y_i and $\hat{\mathbf{p}}_i$; lower scores indicate more suspicious annotations. We sort examples by ascending quality score for manual review or downstream filtering.

Model disagreement (diagnostic). We also record the HPSv3-preferred class $\hat{y}_i = \arg\max_{c \in \{0,1\}} \hat{p}_{i,c}$ and the confidence on the human label, $\hat{p}_{i,y_i}$, as diagnostics alongside Cleanlab outputs. Disagreement ($\hat{y}_i \neq y_i$) correlates with but is not identical to `is_label_issue`$_i$.

A.6.2 Pipeline Summary

The audit pipeline is:

1. filter decisive Pick-a-Pic pairs and map `label_0` to $y \in \{0, 1\}$;
2. score both images with HPSv3 to obtain (s_0, s_1);
3. form $\hat{\mathbf{p}}$ via Eq. equation 144;
4. run Cleanlab issue detection and self-confidence quality scoring on $(y, \hat{\mathbf{p}})$.

HPSv3 supplies a preference prior largely independent of the crowdsourcing pipeline; Cleanlab, given $\hat{\mathbf{p}}$, systematically highlights human labels that are least consistent with that prior, supporting dataset cleaning and noise analysis.

Using HPSv3 and Cleanlab as above, we filter Pick-a-Pic v2 training pairs to remove potentially mislabeled or low-quality preference annotations and obtain a clean preference dataset. We then inject synthetic label noise by randomly flipping preference labels with a probability of $\eta = 0.2$, resulting in the noisy preference dataset used for training and evaluation.

Table 5: Performance comparison on the HPSv2 and PartiPrompt Benchmarks across diffusion and flow matching models. Best results are shown in **bold**, and second-best results are underlined.

Model	Method	HPSv2					PartiPrompt				
		PS↑	HPS↑	AES↑	IR↑	CLIP↑	PS↑	HPS↑	AES↑	IR↑	CLIP↑
SD1.5	*Binary preference alignment*										
	SD1.5	20.93	27.02	5.437	-0.200	32.82	21.50	27.19	5.400	-0.019	31.00
	DPO	21.33	27.53	5.531	0.353	33.67	21.71	27.54	5.462	0.356	31.63
	DMPO	21.97	28.70	5.708	0.705	34.91	22.05	28.45	5.718	0.608	33.17
	α-DPO	22.09	28.71	5.709	0.697	**35.35**	22.04	28.22	5.627	0.531	33.05
	Ternary preference alignment										
	TPO	22.03	28.78	5.698	0.711	34.56	22.09	28.20	5.742	0.622	33.54
	Noisy ternary preference alignment										
	NTPO (Ours)	**22.26**	**28.93**	**5.812**	**0.838**	34.72	**22.28**	**28.78**	**5.956**	**0.741**	**33.95**
SDXL	*Binary preference alignment*										
	SDXL	23.16	29.16	6.066	0.881	37.75	22.66	28.30	5.876	0.648	34.09
	DPO	23.47	29.84	6.179	1.003	37.80	22.83	28.86	5.962	0.998	34.36
	DMPO	23.53	30.19	6.166	1.115	38.22	22.97	28.75	6.020	1.060	34.49
	α-DPO	23.43	29.82	6.262	1.112	38.43	22.84	28.91	6.080	0.898	34.41
	Ternary preference alignment										
	TPO	23.65	30.14	6.285	1.147	38.13	23.07	29.02	6.121	**1.083**	34.63
	Noisy ternary preference alignment										
	NTPO (Ours)	**23.86**	**30.34**	**6.447**	**1.188**	**38.50**	**23.26**	**29.25**	**6.237**	0.978	**34.95**
SD3.5-M	Base	22.95	28.88	6.091	1.165	37.04	22.58	29.08	6.053	0.887	34.01
	TPO	23.17	29.32	6.115	1.183	37.20	22.94	29.15	6.081	0.912	34.32
	NTPO (Ours)	**23.44**	**29.85**	**6.197**	**1.249**	**37.32**	**23.30**	**29.33**	**6.112**	**0.938**	**34.75**
FLUX.1-schnell	Base	23.01	29.59	6.103	1.218	**37.78**	23.15	29.42	6.135	1.051	34.68
	TPO	23.55	30.14	6.188	1.198	37.53	**23.34**	**29.60**	6.158	**1.102**	35.01
	NTPO (Ours)	**23.87**	**30.51**	**6.305**	**1.233**	37.60	23.30	29.51	**6.176**	1.083	**35.24**
AuraFlow-v0.3	Base	22.45	28.15	6.011	**1.111**	36.88	21.98	28.69	**6.002**	0.859	**33.85**
	TPO	23.03	28.77	6.134	1.085	37.01	22.15	29.03	5.890	0.836	33.43
	NTPO (Ours)	**23.16**	**29.35**	**6.247**	1.102	**37.23**	**22.33**	**29.32**	5.935	**0.943**	33.58

A.7 More Quantitative Results

We provide the overall performance comparison in Table 5. The results show that NTPO can be effectively applied to both diffusion and flow matching models, consistently improving their preference alignment performance.



Figure 3: More qualitative comparisons on SD1.5. NTPO produces images with improved visual quality and prompt alignment.

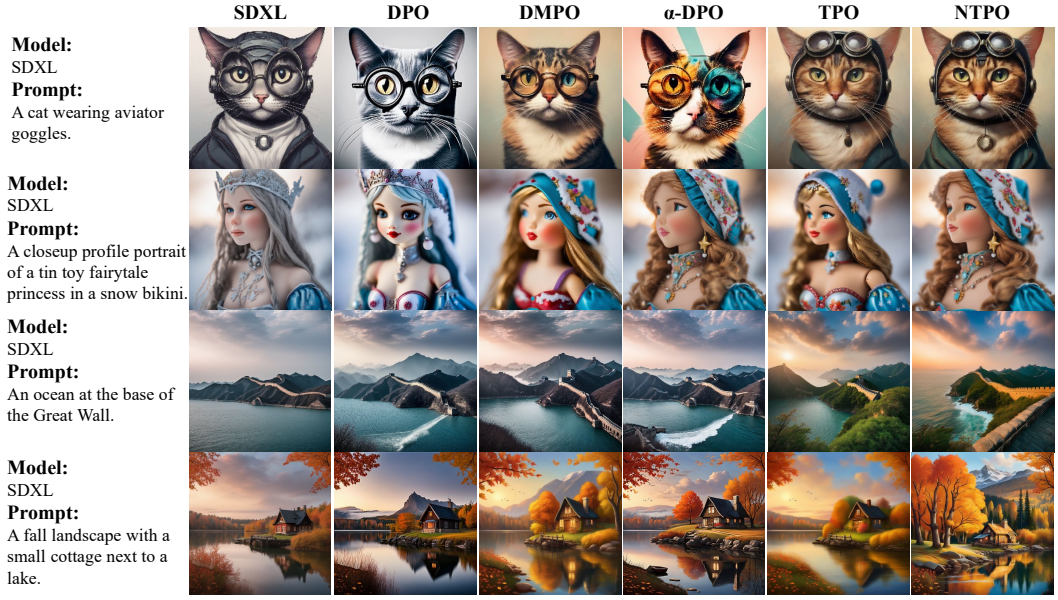


Figure 4: More qualitative comparisons SDXL. NTPO produces images with improved visual quality and prompt alignment.

A.8 More Qualitative Comparisons

We provide additional qualitative comparisons on SD1.5 and SDXL in Figures 3 and 4. Additional results on flow-matching models, including SD3.5-M, FLUX.1-schnell, and AuraFlow-v0.3, are provided in Figures 5a, 5b, and 6, respectively.



(a) More qualitative comparisons on SD3.5-M.



(b) More qualitative comparisons on FLUX.1-schnell.

Figure 5: More qualitative comparisons on flow-matching models. NTPO produces images with improved visual quality and prompt alignment.

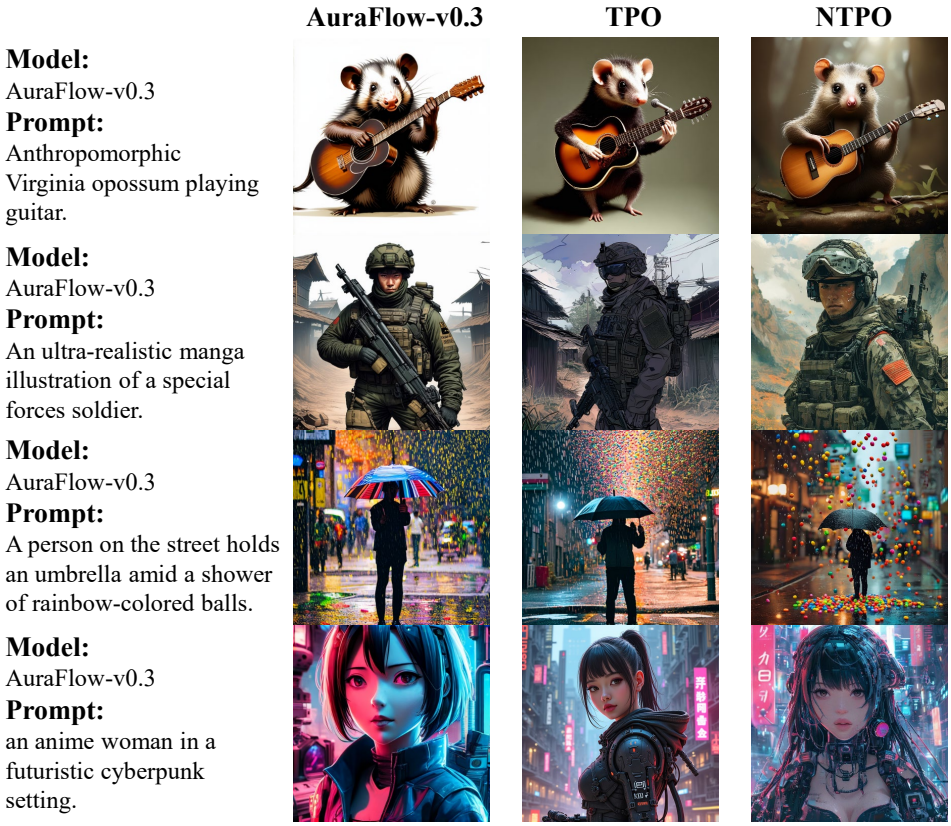


Figure 6: More qualitative comparisons AuraFlow-v0.3. NTPO produces images with improved visual quality and prompt alignment.

Table 6: Ablation study of transition matrix designs for NTPO on SD1.5 and SDXL. We compare the general matrix Q with the proposed simplified matrix under full-parameter training for 2000 steps, evaluating both HPSv2 alignment performance and training efficiency.

Model	Transition Matrix	Training Time (h)	PS↑	HPS↑	AES↑	IR↑	CLIP↑
SD1.5	General Q	**35**	**22.30**	**28.98**	**5.825**	0.831	**34.78**
	Simplified Q	29	22.26	28.93	5.812	**0.838**	34.72
SDXL	General Q	**92**	**23.92**	**30.37**	6.421	1.173	**38.54**
	Simplified Q	85	23.86	30.34	**6.447**	**1.188**	38.50

Table 7: Ablation study of transition matrix designs for NTPO on flow matching models. We compare the general matrix Q with the proposed simplified matrix under LoRA fine-tuning for 100 steps, evaluating both HPSv2 alignment performance and training efficiency.

Model	Transition Matrix	Training Time (h)	PS↑	HPS↑	AES↑	IR↑	CLIP↑
SD3.5-M	General Q	**15**	**23.49**	**29.92**	**6.201**	**1.252**	**37.40**
	Simplified Q	10	23.44	29.85	6.197	1.249	37.32
FLUX.1-schnell	General Q	**52**	**23.91**	**30.56**	**6.311**	1.245	**37.71**
	Simplified Q	44	23.87	30.51	6.305	**1.233**	37.60
AuraFlow-v0.3	General Q	**27**	**23.20**	**29.39**	6.252	1.083	**37.30**
	Simplified Q	22	23.16	29.35	**6.247**	**1.102**	37.23

A.9 Additional Ablation Studies of Transition Matrix Designs

We further investigate the effect of different transition matrix designs in NTPO. Specifically, we compare the general transition matrix with the proposed simplified matrix Q to evaluate whether the simplified noise model can maintain preference alignment performance while reducing training cost.

We report the ablation results for diffusion and flow matching models separately due to their different training configurations. For SD1.5 and SDXL, we perform full-parameter fine-tuning for 2000 training steps, whereas the flow matching models are fine-tuned with LoRA for only 100 steps. Therefore, their training time is not directly comparable across the two settings.

Table 6 reports the results on SD1.5 and SDXL. Overall, the simplified transition matrix achieves alignment performance comparable to the general transition matrix across the evaluation metrics, while requiring lower training cost. This indicates that the simplified matrix provides an effective and efficient approximation for modeling preference noise.

We further evaluate the two transition matrix designs on flow matching models in Table 7. Similar trends are observed under LoRA fine-tuning, suggesting that the proposed simplified transition matrix remains effective across different generative model architectures and training paradigms.