Taming Preference Mode Collapse via Directional Decoupling Alignment in Diffusion Reinforcement Learning

Supplementary Material

A. D^2-Align Algorithm

The D^2-Align framework is a two-stage process that decouples reward signal correction from policy alignment. The detailed process for each stage is provided in Algorithm 1 and Algorithm 2.

Stage 1 (Algorithm 1) focuses on learning a directional correction vector $\boldsymbol{b}_v$. In this stage, the policy model G_θ is frozen, and only the vector $\boldsymbol{b}_v$ is optimized to create a guided reward signal that corrects for reward model biases.

Stage 2 (Algorithm 2) then performs the guided alignment of the policy model itself. The optimized vector $\boldsymbol{b}_v^*$ from Stage 1 is frozen, and the policy model G_θ is unfrozen. The generator's parameters θ are then updated by optimizing for the guided reward signal defined by $\boldsymbol{b}_v^*$.

Algorithm 1 D^2-Align Stage 1: Learning Directional Correction

Require: Initial policy model G_θ (as $\boldsymbol{\epsilon}_\theta$); reward model R ($\Phi_{\text{img}}, \Phi_{\text{text}}, \text{score}$); prompt dataset $\mathcal{C}$; guidance scale ω; Stage 1 total timesteps T_1 for training; diffusion coefficients α_t, σ.
Ensure: Optimized directional vector $\boldsymbol{b}_v^*$.

Initialize learnable directional vector $\boldsymbol{b}_v \in \mathbb{R}^d$
Freeze generator G_θ ($\boldsymbol{\epsilon}_\theta$)
for timestep T_1 to $t = 1$ **do**
  Sample prompt $c \sim \mathcal{C}$
  Generate clean image $\boldsymbol{x}_0 \sim G_\theta(c)$
  Get text embedding $\boldsymbol{e}_{\text{text}} \leftarrow \Phi_{\text{text}}(c)$
  Sample noise $\boldsymbol{\epsilon}_{\text{gt}} \sim \mathcal{N}(0, \mathbf{I})$
  Create noisy latent $\boldsymbol{x}_t \leftarrow \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}_{\text{gt}}$
  Predict noise $\boldsymbol{\epsilon}_{\text{pred}} \leftarrow \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c)$
  Perform one-step ODE sampling to get $\boldsymbol{x}_{t-1}$
  Reconstruct $\hat{\boldsymbol{x}}_0 \leftarrow (\boldsymbol{x}_{t-1} - \sigma_{t-1}\boldsymbol{\epsilon}_{\text{gt}})/\alpha_{t-1}$
  Get image embedding $\boldsymbol{e}_{\text{img}} \leftarrow \Phi_{\text{img}}(\hat{\boldsymbol{x}}_0)$
  Calculate $\boldsymbol{e}_+ \leftarrow \text{normalize}(\boldsymbol{e}_{\text{text}} + \boldsymbol{b}_v)$
  Calculate $\boldsymbol{e}_- \leftarrow \text{normalize}(\boldsymbol{e}_{\text{text}} - \boldsymbol{b}_v)$
  Construct guided embedding $\tilde{\boldsymbol{e}}_{\text{text}} \leftarrow \boldsymbol{e}_- + \omega \cdot (\boldsymbol{e}_+ - \boldsymbol{e}_-)$
  Compute guided reward $R_{\text{guided}} \leftarrow \text{score}(\boldsymbol{e}_{\text{img}}, \tilde{\boldsymbol{e}}_{\text{text}})$
  Update $\boldsymbol{b}_v$ by minimizing $\mathcal{L}_{\text{stage1}}(\boldsymbol{b}_v) = -R_{\text{guided}}$
return $\boldsymbol{b}_v^* \leftarrow \boldsymbol{b}_v$

Algorithm 2 D^2-Align Stage 2: Guided Alignment

Require: Policy model G_θ (as $\boldsymbol{\epsilon}_\theta$); reward model R ($\Phi_{\text{img}}, \Phi_{\text{text}}, \text{score}$); prompt dataset $\mathcal{C}$; guidance scale ω; Stage 2 total timesteps T_2 for training; diffusion coefficients α_t, σ.
Require: Frozen directional vector $\boldsymbol{b}_v^*$ (from Stage 1).
Ensure: Optimized policy model G_{θ^*}.

Unfreeze generator G_θ ($\boldsymbol{\epsilon}_\theta$)
for timestep T_2 to $t = 1$ **do**
  Sample prompt $c \sim \mathcal{C}$
  Generate clean image $\boldsymbol{x}_0 \sim G_\theta(c)$
  Get text embedding $\boldsymbol{e}_{\text{text}} \leftarrow \Phi_{\text{text}}(c)$
  Sample noise $\boldsymbol{\epsilon}_{\text{gt}} \sim \mathcal{N}(0, \mathbf{I})$
  Create noisy latent $\boldsymbol{x}_t \leftarrow \alpha_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\epsilon}_{\text{gt}}$
  Predict noise $\boldsymbol{\epsilon}_{\text{pred}} \leftarrow \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, c)$
  Perform one-step ODE sampling to get $\boldsymbol{x}_{t-1}$
  Reconstruct $\hat{\boldsymbol{x}}_0 \leftarrow (\boldsymbol{x}_{t-1} - \sigma_{t-1}\boldsymbol{\epsilon}_{\text{gt}})/\alpha_{t-1}$
  Get image embedding $\boldsymbol{e}_{\text{img}} \leftarrow \Phi_{\text{img}}(\hat{\boldsymbol{x}}_0)$
  Calculate $\boldsymbol{e}_+ \leftarrow \text{normalize}(\boldsymbol{e}_{\text{text}} + \boldsymbol{b}_v^*)$
  Calculate $\boldsymbol{e}_- \leftarrow \text{normalize}(\boldsymbol{e}_{\text{text}} - \boldsymbol{b}_v^*)$
  Construct guided embedding $\tilde{\boldsymbol{e}}_{\text{text}} \leftarrow \boldsymbol{e}_- + \omega \cdot (\boldsymbol{e}_+ - \boldsymbol{e}_-)$
  Compute guided reward $R_{\text{guided}} \leftarrow \text{score}(\boldsymbol{e}_{\text{img}}, \tilde{\boldsymbol{e}}_{\text{text}})$
  Update generator θ by minimizing $\mathcal{L}_{\text{stage2}}(\theta) = -R_{\text{guided}}$
return $G_{\theta^*} \leftarrow G_\theta$

B. Experimental Setting Details

We conduct all reinforcement learning experiments using FLUX.1.Dev [30] as the base T2I model, following the state-of-the-art methodology. For fairness and comprehensive comparison, all alignment baselines are retrained on the Human Preference Dataset (HPD) v2 [66] using two standardized reward combinations: HPS-v2.1 [66] and HPS-v2.1 [66] + CLIP [23] Score. All experiments are conducted on NVIDIA H20 GPUs with 96GB memory.

B.1. Baseline Implementation Details

To ensure a fair comparison, we standardize the training steps for major baselines and align all reward functions. For the baseline comparisons, we follow the original implementations provided in their official repositories. Details on the training steps for each competing method are provided below.

Table 3. **Comprehensive Hyperparameters of D^2-Align.**

Category	Parameter	Value
GENERAL & MODEL	Random Seed	42
	Resolution (H × W)	720 × 720
	Mixed Precision	bf16
	Gradient Checkpointing	Enabled
	Dataloader Workers	4
	Use EMA	Disabled
OPTIMIZATION	Optimizer	AdamW
	Learning Rate	5×10^{-6}
	Weight Decay	1×10^{-4}
	Gradient Clip Norm	0.1
	LR Warmup Steps	0
	Gradient Accumulation Steps	2
RL ALIGNMENT	ω	1.5
	Stage 1 (b_v) Max Steps	3000
	Stage 2 (G_θ) Max Steps	20
INFERENCE	Sampling Steps	25
	Inference Steps	50
	Train Guidance	1.0
	Shift	3
	Train Batch Size	1
	SP Size / Train SP Batch Size	1 / 1

- **DanceGRPO** [69]: We follow the default open-source hyperparameters and train the model for 300 steps.
- **Flow-GRPO** [40]: Hyperparameters were adopted from a stable PickScore training configuration. The training length is set to 300 steps to maintain alignment with DanceGRPO.
- **SRPO** [55]: The model is trained using its default configuration for the official recommended length of 20 steps.

B.2. D^2-Align Hyperparameters

Our proposed D^2-Align framework employs the two-stage training approach. Stage 1 (Directional Correction) trains the correction vector $\boldsymbol{b}_v$ for 3000 steps, after which the policy model G_θ is aligned in Stage 2 with training 20 steps. Key hyperparameters specific to the D^2-Align are listed in Tab. 3.

C. DivGenBench Construction and Metrics

C.1. Prompt Construction and Examples

Our **3,200** "keyword-driven" prompts are systematically generated using distinct templates for each of the four dimensions, as summarized in Tab. 4. Each dimension uses a specific templating strategy to augment base content with explicit attribute keywords.

- **ID**: Motivated by [28], prompts are built using the template: "A high-quality portrait photo of a/an [age] [ethnicity] [gender] [features]". We combine 3 ages, 6 ethnicities, 2 genders, and 40 physical features (e.g., "with arched eyebrows") derived from CelebA [43]. A conflict-resolution mechanism ensures logical coherence.
- **Style**: We pair 27 classic art styles from WikiArt [58] with base content prompts from Parti [72].
- **Layout**: We combine 80 COCO [37] object classes with 4 counts (two, three, four, five) using designed template to test numerical control and spatial diversity.
- **Tonal**: We augment Parti [72] base prompts with 18 fine-grained keywords across 3 sub-dimensions: Saturation, Contrast, and Brightness.

C.2. Metric Calculation Details

To quantify the extent of PMC, we employ four dimension-customized metrics that measure the model's generative breadth.

Identity Divergence Score (IDS): We use ArcFace [11] to extract a 512-D identity embedding v_i for each of N generated faces. The score is the average pairwise cosine similarity between all unique identity vectors. A **lower** score signifies greater identity diversity.

$$\text{IDS} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \tag{11}$$

Artistic Style Coverage (ASC): Inspired by the Image Retrieval Score (IRS) [13], this metric quantifies the retrievable Style diversity of a generative model relative to the ground-truth data's diversity. We use the CSD [56] feature extractor ($\mathcal{F}$) and define three datasets: a ground-truth (GT) training set $\mathcal{X}_{train}$ (with N_{train} images from WikiArt [58]), a GT reference set $\mathcal{X}_{test}$ (with $N_{sample} = N_{test}$ from WikiArt [58]), and the generated synthetic set $\mathcal{X}_{synth}$ (with $N_{sample} = N_{synth}$ from DivGenBench).The score is computed in three steps:

- **Retrieval:** For a query set $\mathcal{X}_q$ (either $\mathcal{X}_{test}$ or $\mathcal{X}_{synth}$), we first find $\mathcal{X}_{learned}$, the set of unique training images that are the nearest neighbor (in CSD feature space $\mathcal{F}$) to at least one image in $\mathcal{X}_q$, as defined in Eq. (12). We then get the count $N_{learned} = |\mathcal{X}_{learned}(\mathcal{X}_q)|$.

$$\begin{aligned} \mathcal{X}_{learned}(\mathcal{X}_q) = \{x \in \mathcal{X}_{train} \mid \exists g \in \mathcal{X}_q \text{ s.t.} \\ x = \arg\min_{x' \in \mathcal{X}_{train}} d(\mathcal{F}(g), \mathcal{F}(x'))\} \end{aligned} \tag{12}$$

- **Estimation:** Using $N_{learned}$, N_{sample}, and N_{train}, we compute the maximum likelihood estimate of the total "learnable" images s^* (Eq. (13)), and the corresponding "infinite" retrieval score, IRS_∞ (Eq. (14)). This estimates the fraction of $\mathcal{X}_{train}$ that would be retrieved given infinite query samples.

$$s^*(\mathcal{X}_q) = \arg\max_s P(N_{learned}, N_{sample}, s) \tag{13}$$

$$IRS_\infty(\mathcal{X}_q) = s^*(\mathcal{X}_q)/N_{train} \tag{14}$$

- **Adjustment:** To correct for the feature extractor's inherent "measurement gap", the final ASC score is the **Adjusted IRS** ($IRS_{\infty,a}$), shown in Eq. (15). This is the

Table 4. **Prompt Construction Templates and Examples for Each Dimension of DivGenBench.** Brackets indicate keywords sampled from curated attribute lists.

Dimension	One of Templates	Example
ID	A high-quality portrait photo of an [age] [ethnicities] [gender] with [feature].	A high-quality portrait photo of an elderly South Asian woman with arched eyebrows wearing a necklace.
Style	A painting in the style of [style], depicting [base prompt].	A painting in the style of Rococo, depicting a silver fire hydrant next to a sidewalk.
Layout	A studio shot of [number] [object] on a clean white background.	A studio shot of three boats on a clean white background.
Tonal	An image of [base prompt], rendered with [tonal] properties.	An image of ten children on a couch, rendered with dimly lit properties.

ratio of the synthetic set's estimated diversity to the real reference set's estimated diversity. A **higher** score is better.

$$\text{ASC} = \frac{IRS_{\infty}(\mathcal{X}_{synth})}{IRS_{\infty}(\mathcal{X}_{test})} \tag{15}$$

Spatial Dispersion Index (SDI): This metric evaluates the diversity of object layouts across multiple images generated from the *same* text prompt, effectively measuring the model's ability to produce spatially varied results. For M images per prompt, we first use Grounding DINO [41] to detect the bounding boxes $L_i = \{b_j\}$ of the target objects in each image. The Similarity $\text{Sim}_{\text{Layout}}$ (Eq. 16) is calculated by finding the optimal bipartite matching of the detected bounding boxes via the Hungarian algorithm on the IoU matrix, normalized by the maximum number of objects. We then compute the average pairwise **Layout Similarity ($\overline{\text{Sim}}_{\text{Layout}}$)** between all pairs of images, as defined in Eq. 17. Finally, the **SDI** is defined as one minus the average Layout Similarity across all M images, averaged over all prompts P (Eq. 18). A **higher** score signifies greater Layout diversity.

$$\begin{aligned}\text{Sim}_{\text{Layout}}(L_i, L_p) = &\frac{1}{\max(|L_i|, |L_p|)} \\ &\times \sum_{(j,l)\in\mathcal{P}} \text{IoU}(b_j \in L_i, b_l \in L_p)\end{aligned} \tag{16}$$

where $\mathcal{P}$ is the set of optimal matching pairs found via the Hungarian algorithm.

$$\overline{\text{Sim}}_{\text{Layout}} = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{p=i+1}^{M} \text{Sim}_{\text{Layout}}(L_i, L_p) \tag{17}$$

$$\text{SDI} = \frac{1}{P} \sum_{r=1}^{P} \left(1 - \overline{\text{Sim}}_{\text{Layout}}^{(r)}\right) \tag{18}$$

Photographic Variance Score (PVS): This metric, inspired by APG [53], quantifies the spread of generated tonal values. For a set of N images $G = \{g_i\}_{i=1}^{N}$, we first extract a scalar value for each perceptual dimension. For each image g_i, **Saturation** (s_i) is the mean of the S-channel (from an RGB-to-HSV conversion), **Brightness** (v_i) is the mean of the V-channel, and **Contrast** (c_i) is the standard deviation of the grayscale-converted image. We then form three value sets $\mathbf{s} = \{s_i\}_{i=1}^{N}$, $\mathbf{v} = \{v_i\}_{i=1}^{N}$, and $\mathbf{c} = \{c_i\}_{i=1}^{N}$. PVS is the sum of the standard deviations of these three sets. A **higher** score indicates greater tonal control.

$$\text{PVS} = \text{std}(\mathbf{s}) + \text{std}(\mathbf{v}) + \text{std}(\mathbf{c}) \tag{19}$$

D. Extended Experiments

D.1. User Study on HPDv2

To validate the effectiveness of D^2-Align in alignment with human preference, we conducted a comprehensive user study following [4, 69]. We compared our method against the base model (FLUX) and three competitive RL-based baselines: DanceGRPO, Flow-GRPO, and SRPO.

D.1.1. Experimental Setup

We randomly selected `100` prompts from *HPDv2*. For each prompt, we generated images using all five methods with the same random seed. We recruited `20` evaluators who were presented with the generated images in a randomized, blind manner. The evaluators were asked to select the best image based on the following criteria:

- **Detail Preservation:** The clarity, sharpness, and richness of details in the generated image.
- **Color Consistency:** The naturalness, harmony, and realism of the colors.
- **Image-Text Alignment:** How well the generated image accurately reflects the content and intent of the text prompt.
- **Overall:** Considering all the above factors, which image do you prefer?

Table 5. **Prompts Used for The Qualitative Examples in Figure 1 of The Main Paper Grouped by Their Corresponding Dimension.** The numbering (1-16) corresponds to the images in Figure 1, read from left-to-right, top-to-bottom within each dimensional category.

No.	Dimension	Prompt
1	Face	A high-quality portrait photo of a young Middle Eastern woman who is attractive with arched eyebrows
2		A high-quality portrait photo of a middle-aged Middle Eastern woman with a receding hairline who is attractive
3		A high-quality portrait photo of a middle-aged White woman with a big nose
4		A high-quality portrait photo of a young Middle Eastern woman with an oval face
5	Style	An artwork of corgi pizza, in the Baroque style.
6		Imagine a panda bear playing ping pong using a blue paddle against an ostrich using a red paddle. Now, picture it in the style of Fauvism.
7		An image of a giant cobra snake made from salad, with strong Action painting influences.
8		A masterpiece of Pointillism, showing a hot air balloon
9-12	Layout	A clear, top-down view of two tennis rackets arranged on a large white table.
13	Tonal	a woman with sunglasses and red hair, monochrome, black and white
14		a tiger in a forest, desaturated, muted colors
15		An image of an ornate treasure chest with a broad sword propped up against it, glowing in a dark cave, rendered with natural colors properties
16		Photograph of three red lego blocks, captured with neon colors, fluorescent

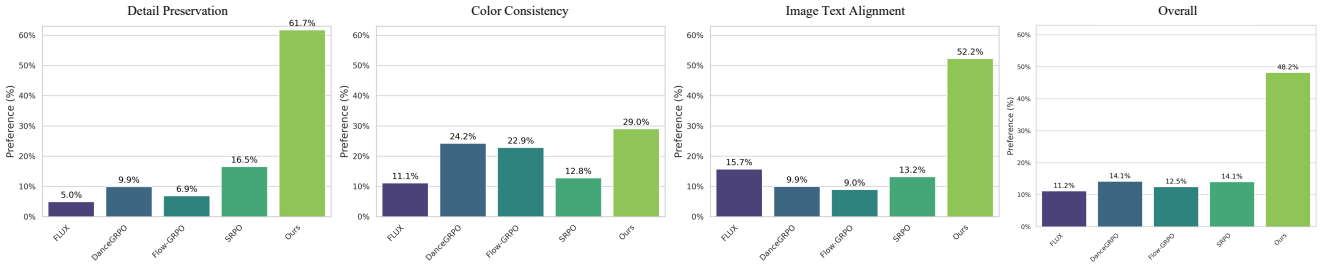

Figure 8. **Human Preference Evaluation on HPDv2.** We conducted a user study comparing D^2-Align against the base model FLUX and state-of-the-art RL alignment methods (DanceGRPO, Flow-GRPO, SRPO). The evaluation spans four distinct dimensions: Detail Preservation, Color Consistency, Image-Text Alignment, and Overall Preference. D^2-Align achieves a dominant lead in Detail Preservation (61.7%) and Image-Text Alignment (52.2%), significantly outperforming baselines that suffer from mode collapse artifacts. Ultimately, our method secures the highest Overall Preference rate of 48.2%.

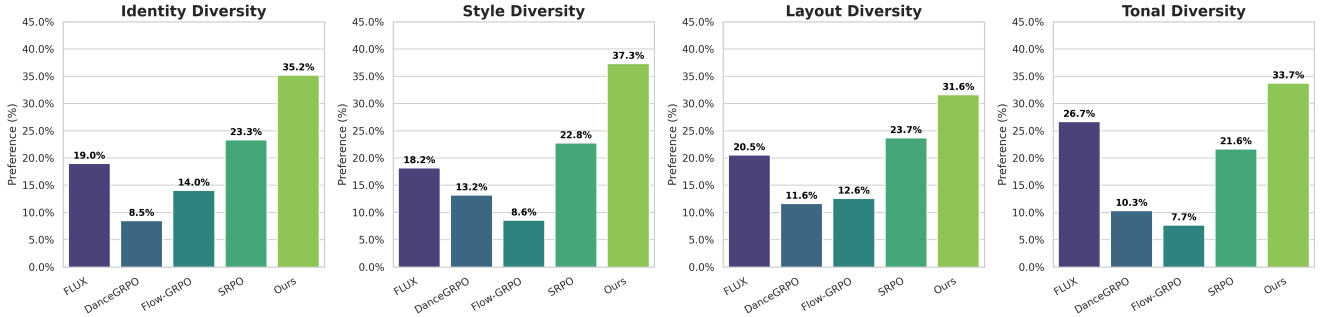

Figure 9. **Human Preference on Diversity (DivGenBench).** We evaluated user preferences across four key diversity dimensions: Identity, Style, Layout, and Tonal. The results reveal a severe PMC in existing RL baselines (DanceGRPO, Flow-GRPO), which often score lower than the Base Model (FLUX), particularly in Tonal and Style diversity. In contrast, D^2-Align consistently achieves the highest preference rates (e.g., 37.3% in Style and 35.2% in Identity), demonstrating its ability to break the trade-off between human preference and generative diversity.

D.1.2. Analysis of Results

The results of the user study are presented in Figure 8. The findings demonstrate a clear and consistent preference for our proposed method, D^2-Align, across all evaluated metrics. Specifically, in the *Detail Preservation* category, D^2-Align was preferred in 61.7% of cases, significantly out-

performing the runner-up, SRPO (16.5%). A similar dominant trend is observed for *Image-Text Alignment*, where D^2-Align achieved a 52.2% preference rate. Furthermore, for *Color Consistency*, our method was chosen 29.0% of the time, again marking a lead over all baselines.

Aggregating the votes, the *Overall* preference for D^2-Align stands at 48.2%, confirming its comprehensive superiority. This strong performance in human evaluations validates that D^2-Align not only improves alignment from a theoretical standpoint but also translates to tangible and perceptually superior generation quality that is easily recognized by human users.

D.2. User Study on DivGenBench

While the HPDv2 study confirms our method's alignment quality, it does not explicitly measure the severity of PMC. To quantitatively assess whether models sacrifice diversity for higher scores, we conducted a second user study using our proposed *DivGenBench*.

D.2.1. Experimental Setup

We sampled `20` distinct templates from each of the four dimensions in DivGenBench: *Identity*, *Style*, *Layout*, and *Tonal*, totaling `80` evaluation sets. For each set, we generated images using varied prompts designed to probe the model's generative boundaries (e.g., requesting specific "Low-key" lighting or distinct "Cubism" styles). `20` evaluators were asked to identify which model best reflected the requested diversity and avoided generating repetitive or homogeneous outputs.

D.2.2. Analysis of Results

The diversity preference results are presented in Figure 9. Two critical observations emerge from the data:

Evidence of Preference Mode Collapse. The results provide strong empirical evidence of PMC in existing RL methods. In several dimensions, the baseline RL models (DanceGRPO and Flow-GRPO) perform significantly worse than the unaligned Base Model (FLUX). For instance, in *Tonal Diversity*, Flow-GRPO drops to a mere 7.7% preference rate, and DanceGRPO to 10.3%, compared to FLUX's 26.7%. Similarly, in *Style Diversity*, Flow-GRPO (8.6%) lags behind FLUX (18.2%). This confirms that naively optimizing for reward models drives the generator into a narrow mode (e.g., always generating overexposed or realistic styles), actively destroying the inherent diversity of the base model.

Superior Diversity Preservation. In contrast, D^2-Align effectively mitigates this collapse. Our method achieves the highest preference scores across all four dimensions, surpassing both the collapsed baselines and the Base Model.

- **Identity & Style:** We achieve dominant preference rates of 35.2% for *Identity* and 37.3% for *Style*, significantly outperforming the runner-up SRPO (about 23%). This indicates our method can generate diverse faces and artistic styles without reverting to a mean template.
- **Layout & Tonal:** Crucially, in the dimensions most susceptible to collapse, our method maintains robustness. In *Tonal Diversity*, where baselines fail, D^2-Align leads with 33.7%, demonstrating successful disentanglement of "quality" from "lighting bias."

These results, combined with the HPDv2 findings, validate that D^2-Align can simultaneously improve human preference alignment while preserving and even enhancing generative diversity.

D.3. Generalizability Study with DanceGRPO

To further evaluate the intrinsic value and generalizability of our learned corrective signal $\mathbf{b}_v$, we conducted an extension experiment by applying it as a plug-and-play component to an external Reinforcement Learning framework. Specifically, we selected DanceGRPO [69], a representative method that, while effective in improving alignment, is susceptible to PMC. The objective of this experiment is to verify whether the directionality captured by $\mathbf{b}_v$ acts as a universal corrective signal for the reward model and can effectively mitigate mode collapse in other algorithms.

D.3.1. Experimental Setup

We maintain the original training logic and hyperparameters of DanceGRPO. The only modification lies in the reward calculation mechanism. Let $\mathbf{b}_v^*$ denote the optimal parameter learned and frozen from Stage 1 of our method. We substitute the naive reward of DanceGRPO with our proposed guided reward R_{guided}. Formally, during the DanceGRPO training process, for every generated sample $(\boldsymbol{x}_0, c)$, the reward is computed as:

$$R_{\text{guided}}(\boldsymbol{x}_0, c; \mathbf{b}_v^*) = \text{score}(\Phi_{\text{img}}(\boldsymbol{x}_0), \tilde{\boldsymbol{e}}_{\text{text}}) \tag{20}$$

where $\tilde{\boldsymbol{e}}_{\text{text}}$ is the rectified text embedding constructed using $\mathbf{b}_v^*$ via Eq. (8) (as defined in the main paper), with the guidance scale ω kept consistent.

D.3.2. Analysis of Results

The quantitative comparisons in Tab. 6 and Tab. 7 demonstrate that integrating our corrective signal significantly enhances the robustness of DanceGRPO. In terms of alignment, while vanilla DanceGRPO achieves the highest HPS-v2.1 score, it suffers from regression in generalized metrics. In contrast, applying our learned $\mathbf{b}_v$ effectively mitigates reward overfitting: it achieves a **4.7%** improvement in Aesthetic Score and restores semantic consistency with a **6.1%** gain in CLIP score, suggesting a shift towards true human preference. Crucially, for diversity, our method effectively

Table 6. **Comprehensive Quantitative Evaluation of Metrics for Human Preference Alignment and Semantic Consistency**. We compare FLUX, DanceGRPO, and DanceGRPO incorporated with our learned $\boldsymbol{b}_v$. All RL-based methods utilize HPS-v2.1 as the reward model. **Ranking is performed between the RL-based methods.** The best score is shown in **bold**.

Method	Human Preference Alignment					Semantic Consistency & Accuracy		
	Aesthetic ↑	ImageReward ↑	Pick Score ↑	Q-Align ↑	HPS-v2.1 ↑	CLIP ↑	DeQA ↑	GenEval ↑
FLUX	6.417	1.670	0.240	4.922	0.310	0.315	4.456	0.663
DanceGRPO	6.068	1.664	0.241	4.930	**0.361**	0.293	4.400	0.522
DanceGRPO + Our learned $\boldsymbol{b}_v$	**6.353**	**1.677**	**0.242**	**4.947**	0.319	**0.311**	**4.496**	**0.641**

resolves the mode collapse observed in the baseline. By filtering out the low-diversity manifold, our corrective signal reduces the IDS score by **20.1%** and expands the ASC score by a remarkable **57.7%**, surpassing both the baseline and the pre-trained FLUX. These results confirm that $\mathbf{b}_v$ forces the external optimizer to explore a broader solution space without requiring complex re-tuning of the training configuration.

Table 7. **Quantitative Evaluation of Generative Diversity on DivGenBench**. We compare FLUX, DanceGRPO, and DanceGRPO enhanced with our learned $\boldsymbol{b}_v$. All RL-based methods utilize HPS-v2.1 as the reward model. We report Identity Divergence Score (IDS), Artistic Style Coverage (ASC), Spatial Dispersion Index (SDI), and Photographic Variance Score (PVS). **Ranking is performed between the RL-based methods.** The best score is shown in **bold**.

Method	IDS ↓	ASC ↑	SDI ↑	PVS ↑
FLUX	0.280	0.179	0.563	0.408
DanceGRPO	0.348	0.130	0.488	0.259
DanceGRPO + Our learned $\boldsymbol{b}_v$	**0.278**	**0.205**	**0.604**	**0.437**

E. Visualization

E.1. Prompts for Figure 1

We present the example prompts used to generate the qualitative comparisons in Figure 1 of the main paper. The prompts, grouped by their corresponding dimension, are detailed in Tab. 5.

E.2. Results on HPDv2

In this section, we present qualitative comparisons on the HPDv2 [66] benchmark to visually evaluate the performance of our method against the baseline approach and advanced RL-based methods. Fig. 10 illustrates the visual outputs where all competing RL-based methods were trained using only HPS-v2.1 [66] as the reward model. Fig. 11 shows the results for the same set of methods, but where the models were trained using a combined reward signal from both HPS-v2.1 [66] and CLIP [23], inspired by DanceGRPO [69].

E.3. Results on DivGenBench

We provide a comprehensive visual evaluation of our method on DivGenBench. Observing a critical trade-off in existing work, we strategically focus our comparative visualization on the two state-of-the-art methods that achieved the highest performance on the HPDv2 benchmark yet simultaneously recorded the lowest diversity scores on DivGenBench.

As shown in Fig. 12, our method generates distinct identities, avoiding the mode collapse seen in baselines. It further demonstrates a broad range of artistic styles without defaulting to generic aesthetics (Fig. 13), produces diverse spatial layouts (Fig. 14), and maintains a wide tonal spectrum in brightness and contrast (Fig. 15), contrasting with the monotonic distributions of competing methods.

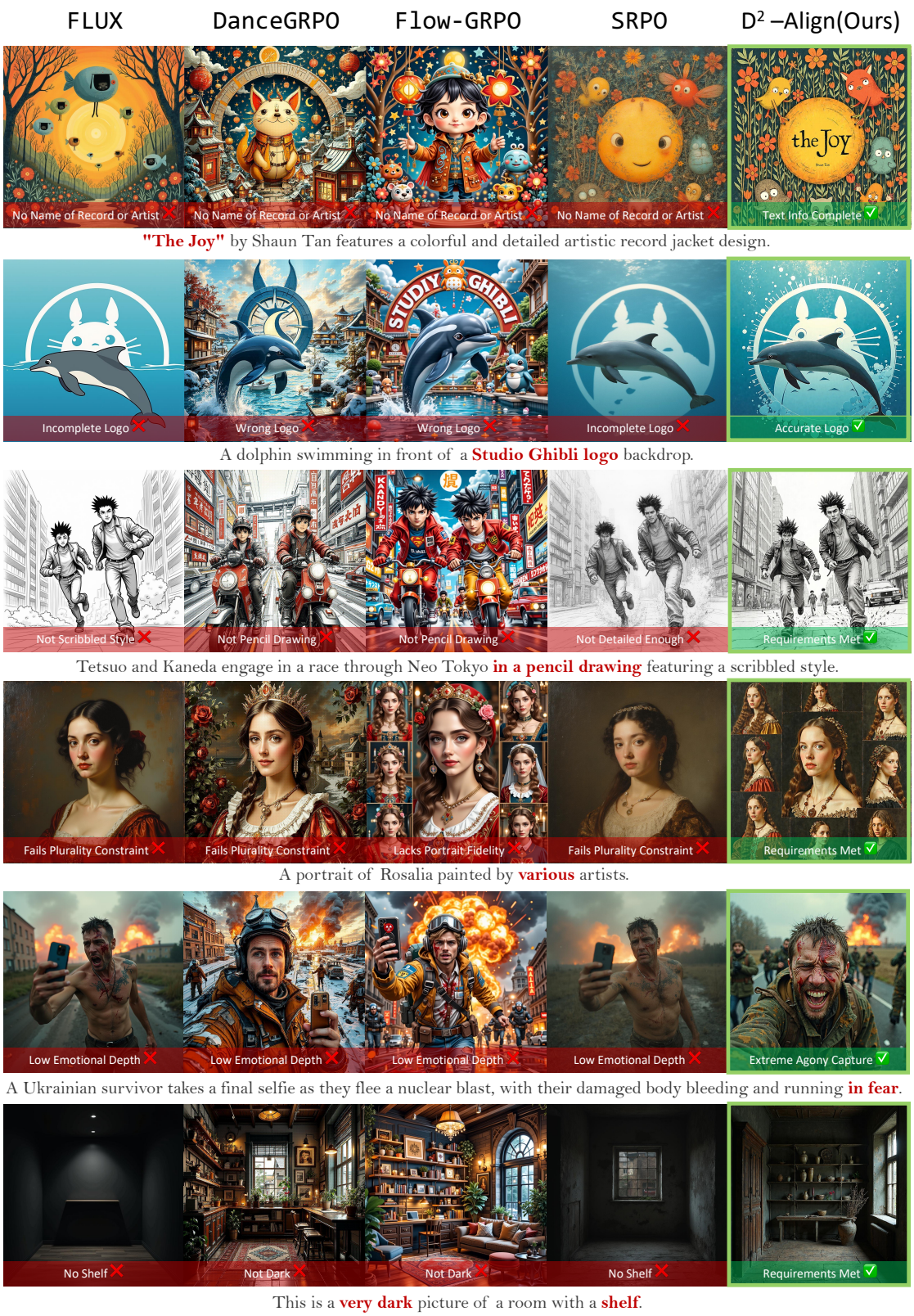

Figure 10. Qualitative comparison of D^2-Align against SOTAs on the HPDv2 benchmark. All RL-based methods are trained using HPS-v2.1 as the reward model.

Figure 11. Qualitative comparison of D^2-Align against SOTAs on the HPDv2 benchmark. All RL-based methods are trained using HPS-v2.1 and CLIP as the reward models.

Figure 12. Qualitative comparison on the ID dimension of DivGenBench. Our method generates diverse identities adhering to required demographic features.

Figure 13. Qualitative comparison on the Style dimension of DivGenBench. Our method faithfully renders diverse artistic styles specified in the prompts.

Figure 14. Qualitative comparison on the Layout dimension of DivGenBench. Our method not only achieves precise adherence to object counts but also generates diverse and novel spatial arrangements.

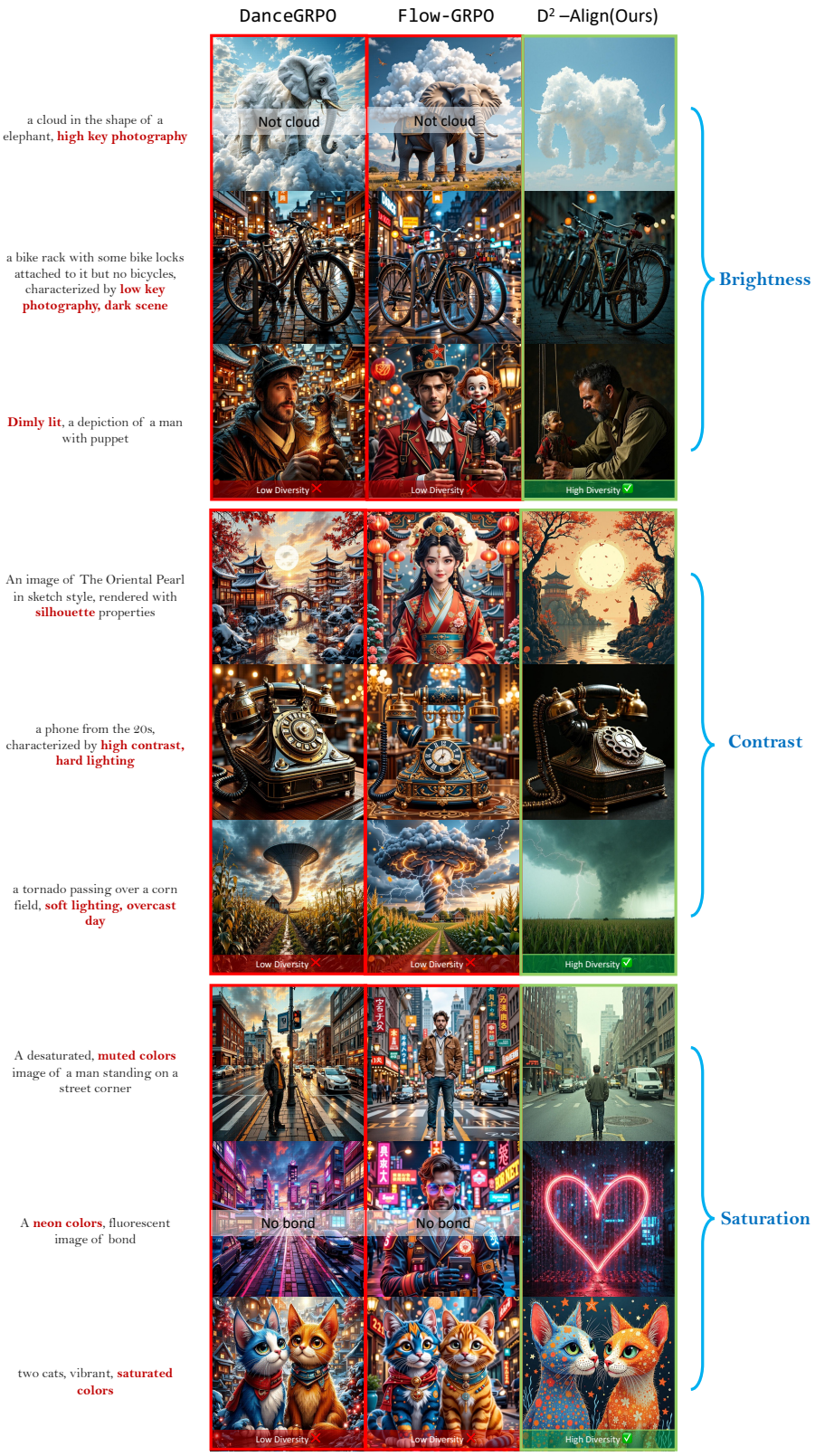

Figure 15. Qualitative comparison on the Tonal dimension of DivGenBench. Our method generates a diverse spectrum of brightness, contrast, and saturation levels while maintaining high image fidelity.