A. Appendix

A.1. Proof of inevitability of reward hacking

Consider an arbitrary reward function $r(x)$ that is sufficiently smooth. Consider a non-parametric probability distribution $p(x)$ which maximizes the expected reward:

$$p^* = \arg\max \mathbb{E}_{x \sim p(x)} \left[r(x) \right] \tag{15}$$

with the constraint $\int_x p(x) \mathrm{d}x = 1$. This is written as a maximization problem with Langrange multiplier β

$$\max_p \int_x p(x) r(x) \mathrm{d}x - \beta \left(\left(\int_x p(x) \mathrm{d}x \right) - 1 \right) \tag{16}$$

$$= \int_x p(x) \left[r(x) - \beta \right] \mathrm{d}x + \beta \tag{17}$$

$$= \int_x \mathcal{L}(x, p, \dot{p}) \mathrm{d}x + \beta \tag{18}$$

This is an Euler-Langrage equation where $\mathcal{L}(x, p, \dot{p}) = p(x)(r(x) - \beta)$. The maximizer of this equation is given by:

$$\frac{\partial L}{\partial p} - \frac{\mathrm{d}}{\mathrm{d}t} \left[\frac{\partial L}{\partial \dot{p}} \right] = 0 \tag{19}$$

Since the second term in Eq. (19) is zero, we get

$$\frac{\partial L}{\partial p} = r(x) - \beta = 0 \tag{20}$$

Note that β should be a constant, for a general $r(x)$ Eq. (20) will not hold true. However, note that $p(x) \geq 0 \forall x$. Therefore, if $r(x) < \beta$, then $p(x) = 0$ and if $r(x) > \beta$, then $p(x)$ will grow indefinitely, violating the pdf constraint $\int_x p(x) \mathrm{d}x = 1$. However, β should be chosen such that $r(x) \leq \beta \forall x$. However, if $r(x) < \beta \forall x$, then $p(x) = \forall x$. Therefore, β is chosen to be $\beta = \sup r(x)$, leading to the optimal distribution

$$p^*(x) = \delta(x - x^*) \tag{21}$$

where $x^* = \arg\max r(x)$. If r has multiple maxima with the same maximum value, say $\{x_1^*, x_2^* \ldots x_n^*\}, r(x_i^*) = \sup r(x)$, then there exists a family of optimal distributions:

$$p^*(x) = \sum_i w_i \delta(x - x_i^*) \tag{22}$$

such that $\sum_i w_i = 1$. We assume that r is not 'flat' at this maximum value, therefore $p^*(x)$ lacks diversity.

For a conditional distribution $p(x|c)$ maximizing the reward $r(x, c)$, a similar derivation yields $p(x|c) = \delta(x - x^{*rc})$, where $x^{*rc} = \arg\max r(x, c)$. This completes the proof in the non-parameteric case. Note that no assumption is made about the finetuning algorithm (DPO, DRaFT, ReFL, etc.) or nature of the reward function (CLIP, JPEG compression, Aesthetics, etc.). This proves that reward hacking is an artifact of the expected reward maximization problem formulation itself.

In the parameteric case, the optimal $p^*(x)$ may not be achievable due to the parameterization. However, even with low-dimensional parameter updates like LoRA, we notice a substantial loss of image diversity when training DRaFT. Qualitative comparisons between the base, DRaFT and our regularization are shown in Figs. 13 to 16.

A.1.1 Adding dropout to reward functions does not work

Moreover, this explains why even reward functions in [13] with aggressive dropout rates (> 0.95) still led to reward collapse. Let the reward model be parameterized by φ and let $N = |\varphi|$ be the dimension of the reward model parameters. Under the dropout case with dropout parameter ξ, the expected reward maximization formulation becomes:

$$p^* = \arg\max \mathbb{E}_{x \sim p(x), u \in \mathcal{U}[0,1]^N} \left[r_{\varphi[u,\xi]}(x) \right] \tag{23}$$

where u is sampled from an i.i.d. multidimensional uniform distribution over 0 to 1, i.e. $u \sim \mathcal{U}[0,1]^N$, and $\varphi[u,\xi]$ are the parameters after applying dropout with random variable u and dropout threshold ξ. Since x and u are independent, we can simplify the expression Eq. (23) by expanding the expectation over u to obtain:

$$p^* = \arg\max \mathbb{E}_{x\sim p(x)} \left[\mathbb{E}_u \left[r_{\varphi[u,\xi]}(x)\right]\right] = \arg\max \mathbb{E}_{x\sim p(x)} \left[\tilde{r}_\xi(x)\right] \tag{24}$$

where

$$\tilde{r}_\xi(x) = \mathbb{E}_{u\in\mathcal{U}[0,1]^N} \left[r_{\varphi[u,\xi]}(x)\right] \tag{25}$$

is independent of random variable u. This is the same optimization problem as Eq. (15) with a new reward function $\tilde{r}_\xi$, therefore having the same reward hacking problem.

A.2. More reward-diversity tradeoff analysis

We perform more ablations on the analysis shown in Sec. 4.3. Specifically, we train SDv1.4/SDXL on Pickscore/HPSv2 reward models and compare the reward-diversity tradeoffs for various regularizations (i.e. KL, LoRA scaling, AIG). Then, we generate images from the PartiPrompt prompt dataset containing over 1600 prompts, and all four subsets of the HPSv2 prompt dataset (containing 800 prompts each from 4 categories). Additionally, images are generated from the coverage prompts as mentioned in Sec. 4.1. Next, the training reward score is computed on the generated images on these prompt datasets, and plotted against the diversity score from the coverage dataset. These quantitative reward-diversity tradeoffs are shown in Figs. 17 to 20. AIG consistently attains better reward-diversity tradeoff than LoRA scaling and KL divergence on both SDv1.4 and SDXL architectures trained on the Pickscore dataset. On the HPSv2 trained models, we observe a smaller Pareto gap between the baseline and our method. Moreover, there is a trend reversal among the baselines. In the HPSv2 trained models, KL regularization seems to outperform LoRA scaling, but in the Pickscore trained models, LoRA scaling tends to outperform KL regularization. This highlights the versatility and reliability of AIG as an effective regularization, being less volatile to trend reversals.

CLIP-based image-text alignment We also compare CLIP [46] scores on the HPSv2 prompt data splits for all four configurations. These results are shown in Figs. 21 to 24. Note that unlike the reward-diversity analysis, maximizing Pickscore/HPSv2 rewards does not imply higher text-to-image alignment. This is evident from the DRaFT model consistently underperforming the base model in terms of CLIP score. Consequently, the theoretical optimal is sometimes very close to the base models and there is no CLIP-diversity tradeoff anymore. Therefore, we compare models only on the CLIP score for a particular value of diversity metric (FID, Recall, Spectral Distance). However, we notice that different regularizations do not deviate in CLIP score for different values of regularization. We make two interesting observations comparing AIG and LoRA scaling. First, AIG outperforms LoRA scaling overall (on PartiPrompts and 3/4 HPSv2 subsets) on all four configurations (SDv1.4/SDXL trained on Pickscore/HPSv2), achieving a notably higher CLIP score. The difference is more notable for the SDv1.4 variants, owing to the already high baseline image-text alignment of SDXL compared to SDv1.4. Second, we qualitatively observe that reward model training produces images that are stylistically more cartoony. Consequently, this leads to lower text-image alignment for HPSv2 photo prompts where the prompts mention the words ”photo” but the generated images look more cartoony and dreamish. AIG preserves more characteristics of the DRaFT alignment, that leads to a slightly lower CLIP score than LoRA scaling on this particular subset.

A.3. Qualitative Results

We present two qualitative comparisons.

Comparison with DRaFT. In this section, we show a few uncurated subsets of images generated by DRaFT and AIG, similar to those shown in user studies in Figs. 13 to 16. Note that across network architectures and reward functions, AIG consistently demonstrates high quality images compared to the base model, while much higher diversity than the DRaFT model.

Comparison with other baselines. We consider three other baselines in this work:

DOODL [65]: This method aims to use Exact Diffusion Inversion to optimize the noise latent that produces an image to maximize classifier guidance, by directly backpropagating through the pre-trained classier’s score on the generated image. The classifier can be interpreted as a reward model, and the optimization is done at inference-time. However, this method takes very long to produce results. For example, with its default configuration (50 optimization steps for 50 DDIM steps), producing the images from the PartiPrompt prompt set will take ∼775 GPU hours, as opposed to $<$ 30 minutes for our method.

ReNO [15]: This method is functionally similar to DOODL, except it works on one-step diffusion models. Given a one-step diffusion model (distilled from a DDPM/DDIM model) $G_\theta(\epsilon, \mathbf{c})$ that generates an image based on noise ϵ and (prompt) conditioning $\mathbf{c}$, and a reward function R, the ReNO objective is defined as

$$\epsilon^* = \arg\max_{\epsilon} R(G_\theta(\epsilon, \mathbf{c}), \mathbf{c}) \tag{26}$$

Since both the one-step diffusion model G and reward function R are differentiable, Eq. (26) is solved using direct optimization using gradient ascent techniques. Moreover, to prevent divergence from the initial data distribution, a regularization based on the proximity of the noise ϵ to the normal distribution $\mathcal{N}(0,1)$ is measured. In the paper, a χ^d regularization on the norm of the noise is used.

ReFL [69]: Directly optimizing LDMs with a reward models is expensive due to its many sampling steps. However, [69] observe that the the rewards of the images in the middle of the sampling chain are indicative of the final scores. Therefore, the LDM chooses a model and a randomly chosen timestep in the middle of the sample chain, and computes gradient *only* with respect to that step. This prevents an expensive gradient computation step for finetuning the LDM.

[13] already show that DRaFT performs quantitatively better than these baselines; we focus on qualitative differences. For all methods, we use their default recommended configurations. We qualitatively evaluate all models by generating images from the PartiPrompt dataset. Qualitative comparison is shown in Figs. 11 and 12.

A.4. User Study

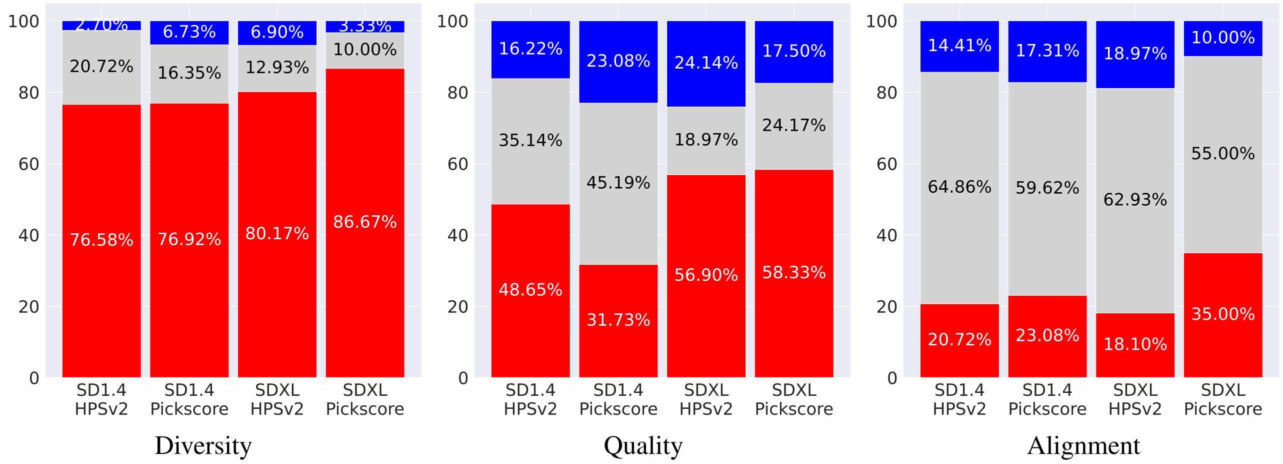
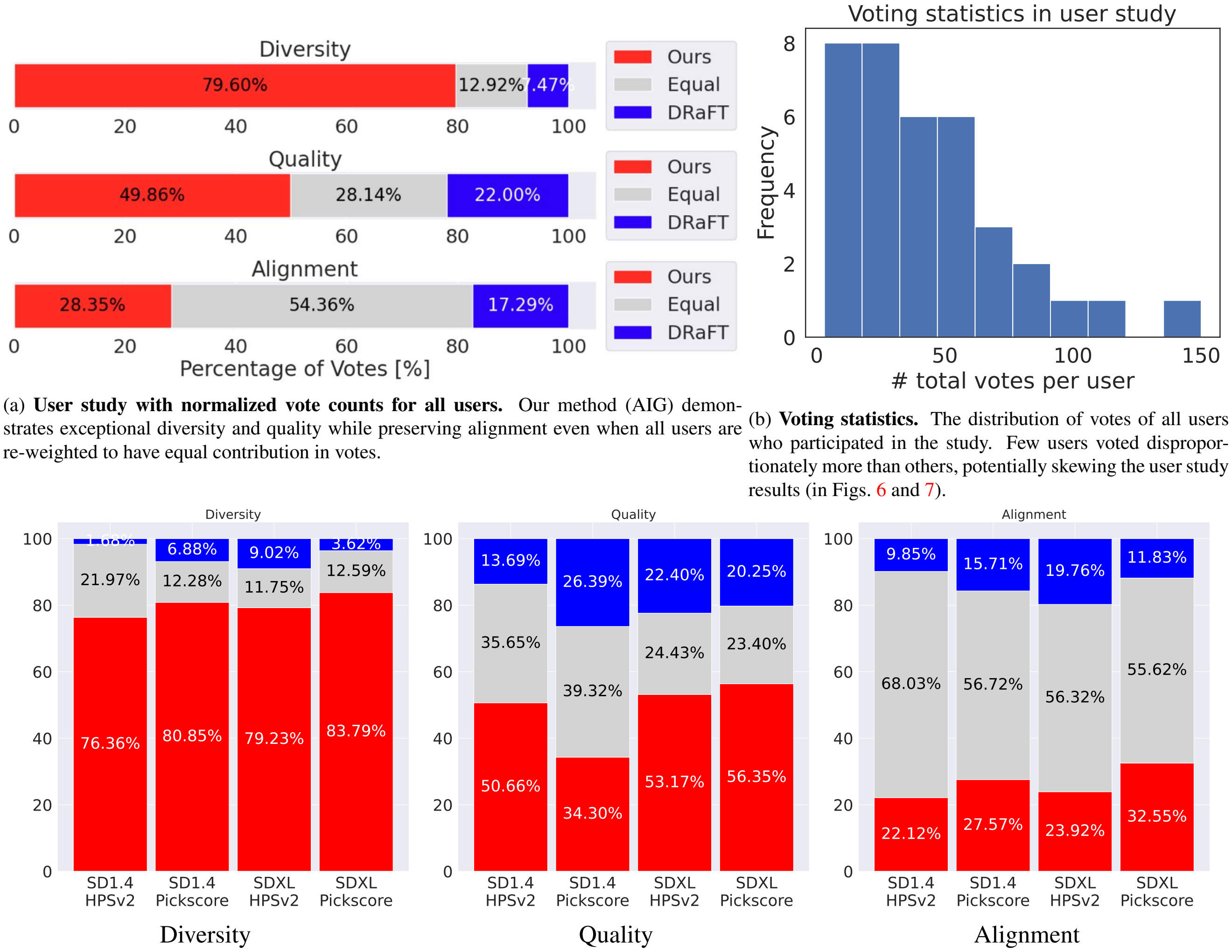


Figure 7. **User study results aggregated by DM+reward configuration** User preferences are consistent across models and reward functions, confirming that the user preference towards our method is due to the AIG itself and is not due to any specific architecture or reward model.

In this section, we provide more details on the user preference study. The objective of the user study is to quantify if the proposed method: AIG leads to increased diversity at the cost of any loss of quality or alignment. To this end, we use the 'coverage prompt' dataset, which is a subset of 40 prompts from the PartiPrompt prompt dataset. For each prompt, 50 images are generated for all methods with the same noise latents for consistency. These images are generated for all four configurations - SDv1.4 and SDXL models that are trained on Pickscore and HPSv2 rewards. The web UI assigns a unique user ID to a browser session, and randomly chooses a prompt, DM+reward configuration, and selects 9 random indices from the 50 generated images without replacement, randomly shuffles the order, and displays the images side by side (Fig. 10). We recruit 36 participants and provide them basic app usage instructions prior to conducting the study, and we collect more than 1500 total votes from all users. Users are referenced by browser cookie information, allowing us to preserve user anonymity while collecting user-specific voting statistics. Both overall voting results, and results aggregated by configurations are summarized and discussed in Sec. 4.4.

User-normalized preference However, our user study allows the users to cast a different number of votes as per their convenience. This leads to a slight non-uniformity in the distribution of votes (Fig. 8b), which can skew the preference scores towards users who cast more votes. To highlight this potential discrepancy between overall vote distribution and user-normalized vote distribution, we compute re-normalized preference scores as follows. Instead of counting votes towards a particular baseline (DRaFT, AIG or Equal) followed by normalization, we normalize the number of votes of each individual user to sum to a 100. Next, we aggregate these normalized votes for each user. Essentially, we calculate the average preference of users irrespective of the total number of votes cast by each user. These results are shown in Fig. 8. Interestingly, the trends do not drastically shift from that in Figs. 6 and 7, showing that user agreeability on preferences is high. If users disagreed on preferences, then the unweighted and reweighted preference statistics may have been different.



(a) **User study with normalized vote counts for all users.** Our method (AIG) demonstrates exceptional diversity and quality while preserving alignment even when all users are re-weighted to have equal contribution in votes.

(b) **Voting statistics.** The distribution of votes of all users who participated in the study. Few users voted disproportionately more than others, potentially skewing the user study results (in Figs. 6 and 7).

(c) **User study results aggregated by DM+reward configuration with normalized votes** User preferences are consistent across models and reward functions, indicating the improvement is due to the AIG itself. Images best viewed zoomed in.

Figure 8. **User Study comparing AIG with DRaFT with user-based reweighing of votes**. We recruited 36 participants who compared the quality, diversity and alignment of AIG and DRaFT, resulting in more than 1500 votes. In contrast to Fig. 6, this study reweighs each user contribution to have the same number of votes, to avoid skewing the user study in favor of users who voted more than others (Fig. 8b).

A.5. Limitations and Future Work

Although our work aims to study regularization techniques that are inference-time, and mitigate the 'reference mismatch' problem, it does not completely eliminate it. For example, in AIG, the earlier sampling steps are dominated by the score function of the base model, with the underlying assumption that each mode of the original data distribution has a 'high-reward region' close to it. This assumption is usually true for stylistic changes (e.g. Pickscore or HPSv2 rewards that primarily alter the style of the images) but may not necessarily hold for text-to-image alignment (e.g. changing spatial relationships, counts or attributes of objects). This assumption is also similar to the motivation used in works like SDEdit [37], which recover realistic images from a 'partially noisy label'. Consequently, we observe a huge improvement in quality, but do not improve text-to-image alignment significantly. The lack of improvement of text-to-image alignment also raises questions about the efficacy of reward models trained on human preference data themselves. Most reward models trained on large human preference datasets achieve maximum validation accuracies of 65-70% [30, 68, 69], questioning the presence of any discernable, objective learnable signal present in these images w.r.t. alignment. This forms the basis for future work by using finegrained alignment using Large Multimodal Models (LMMs) to identify and correct mistakes in the image that do not align with the prompt. Another line of future work pertains to the choice of γ. Since our choice of γ allows immense flexibility of the interplay of the base and DRaFT sampling dynamics, a user interface can be built where users can choose from a predetermined set of γ curves, followed by finetuning these curves using spline interpolation from points clicked on by the user.

A.6. More intuition for Annealed Importance Guidance

A.6.1 Asymmetric Mixing Dynamics

Consider a data distribution composed of only a finite set of images $\boldsymbol{\mu}_i \in \mathbb{R}^d$, i.e. $p_{\text{data}}(\mathbf{x}) = \sum_{i=1}^N w_i \delta(\mathbf{x} - \boldsymbol{\mu}_i)$, $\sum_i w_i = 1$. The average distance between any two 'modes' of the distribution is $m_{\text{data}} = \mathbb{E}_{i \neq j}\left[\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2\right]$. Now, consider the forward diffusion $\mathbf{x}_t = \sqrt{\tilde{\alpha}_t}\mathbf{x} + \sigma_t \epsilon$, $\epsilon \in \mathcal{N}(0, 1)$. The distribution is given by

$$q_t(\mathbf{x}_t) = \int_{\mathbf{x}_0} q(\mathbf{x}_t|\mathbf{x}_0) p_{\text{data}}(\mathbf{x}_0) \mathrm{d}\mathbf{x}_0 \tag{27}$$

$$= \int_{\mathbf{x}_0} \frac{1}{\sqrt{2\pi\sigma_t^d}} \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\tilde{\alpha}_t}\mathbf{x}_0\|_2^2}{2\sigma_t^2}\right)\left(\sum_i w_i \delta(\mathbf{x}_0 - \boldsymbol{\mu}_i)\right) \mathrm{d}\mathbf{x}_0 \tag{28}$$

$$q_t(\mathbf{x}_t) = \frac{1}{\sqrt{2\pi\sigma_t^d}} \sum_i w_i \exp\left(-\frac{\|\mathbf{x}_t - \sqrt{\tilde{\alpha}_t}\boldsymbol{\mu}_i\|_2^2}{2\sigma_t^2}\right) \tag{29}$$

Therefore, $q_t(\mathbf{x}_t)$ is a Gaussian Mixture Model (GMM) with means $\boldsymbol{\mu}_i^{(t)} = \sqrt{\tilde{\alpha}_t}\boldsymbol{\mu}_i$. The average distance between the means is now $m_{\text{data}}^{(t)} = \mathbb{E}_{i \neq j}\left[\|\boldsymbol{\mu}_i^{(t)} - \boldsymbol{\mu}_j^{(t)}\|_2\right] = \sqrt{\tilde{\alpha}_t}\mathbb{E}_{i \neq j}\left[\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2\right] = \sqrt{\tilde{\alpha}_t} m_{\text{data}}$. The average distance between any two modes decreases, as they collapse onto each other to form a unimodal Gaussian distribution. Two data modes are therefore easy to tell apart from each other for small t instead of larger t. Therefore, the score matching objective must be most discriminative in terms of mode recovery from the later stages, where samples from $q_t(\mathbf{x}_t)$ cannot be reliably distinguished from each other in terms of the mode of the data distribution that they originated from. We demonstrate this using a simple example.

Toy problem illustrating mode-recovering nature of sampling dynamics. Consider a toy example of a 1D Gaussian Mixture distribution with two modes as data distribution, i.e. $p_{\text{data}}(\mathbf{x}) = 0.5(\mathcal{N}(1, 0.05) + \mathcal{N}(-1, 0.05))$. We consider a 1000 forward diffusion steps for this problem. Top row in Fig. 9 shows samples from $q_t(\mathbf{x}_t)$ with the colors representing the mode of the original p_{data} distribution from which the sample of q_t is generated. As t increases, the samples mix with each other and become less distinguishable. To show the mode recovering behavior, we sample one data point from $q_t(\mathbf{x}_t)$ and plot the distribution of $p(\mathbf{x}_{t-10}|\mathbf{x}_t)$ using Langevin dynamics with the ground-truth score function. The intuition is that if $p(\mathbf{x}_{t-10}|\mathbf{x}_t)$ is high variance, then the subsequent samples from $p(\mathbf{x}_{t-20}|\mathbf{x}_{t-10})$ will discover different modes, and eventually the true data distribution. Middle row in Fig. 9 shows the samples from $p(\mathbf{x}_{t-10}|\mathbf{x}_t)$ for different t, using the ground-truth score function, and bottom row shows the samples from $p(\mathbf{x}_0|\mathbf{x}_t)$. The most high variance behavior is shown for larger values of t, and only a local finetuning to a particular mode of the data for smaller values of t. This motivates our regularization for Annealed Importance Guidance. During the earlier stages of reverse stage sampling (high values of t) when the mode-recovering behavior is the highest, we let the sampling dynamics be governed by the base model. This helps in early recovery of multiple modes of the data. During the later stages (smaller t), the score function from DRaFT dominates the convergence of these samples to the nearest high-reward samples.

A.7. Implementation Details

A.7.1 Training details

We initialize the model with learnable LoRA parameters in the UNet of the latent diffusion model, and keep all other components (text encoders, VAE decoder, reward model) frozen. All models are trained on 8 NVIDIA H100 GPUs. We use a micro

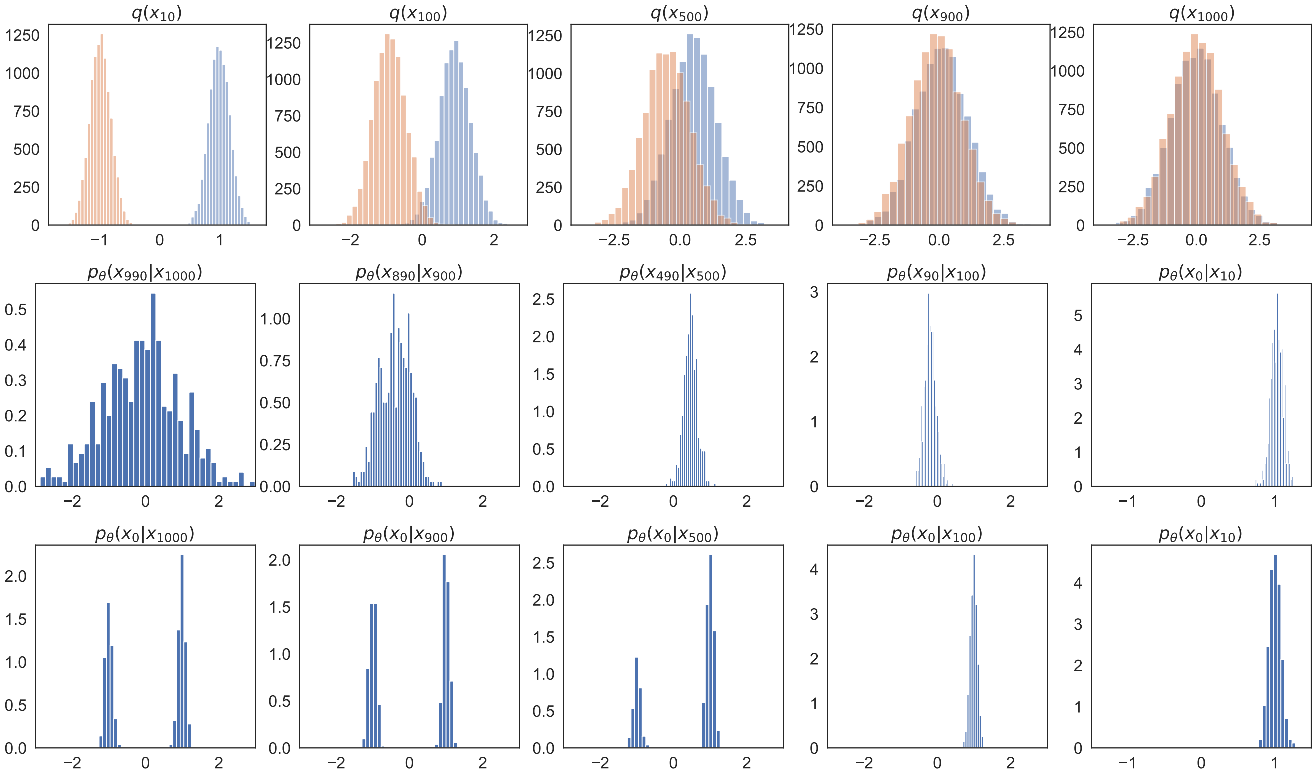
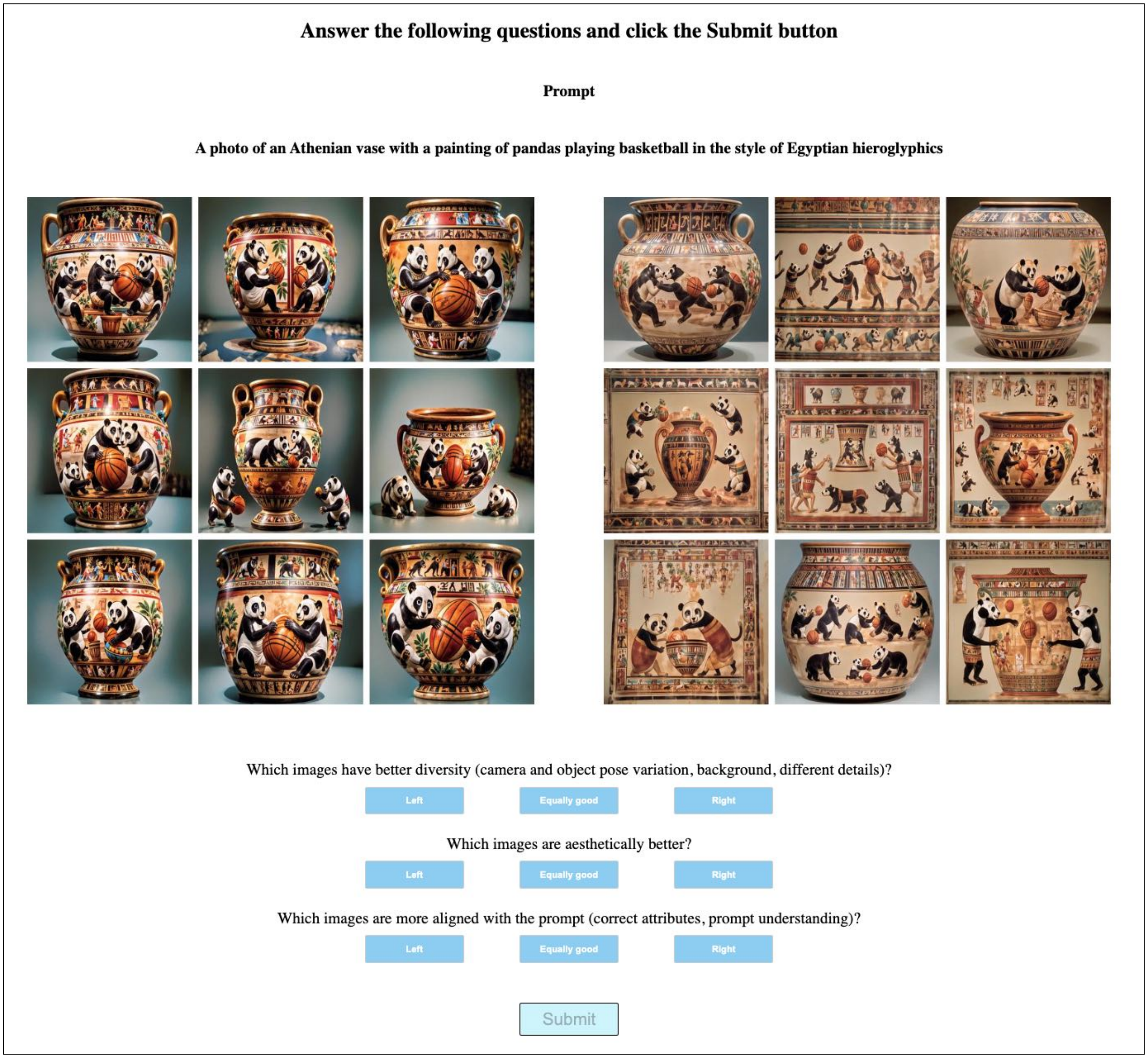


Figure 9. **Toy example showing the sample distribution during forward and reverse time sampling for a 1D problem**. Top row shows that data modes tend to mix more with larger t, indicating that mode-recovery behavior must emerge during later timesteps. Middle and bottom rows show that if a data is sampled from q_t for higher t, then multiple modes of the data are covered from Langevin dynamics, motivating the use of the base model to guide the initial phase of sampling from the reverse-time SDE, and using DRaFT for local finetuning.

batch size of 1 with 4 gradient accumulation steps. For each model, we generate images of the recommended resolution, i.e. images with resolution 512×512 for SDv1.4 and images with resolution 1024×1024 for SDXL. For all models, a constant learning rate of 2.5e-4 is used, without any warmup, annealing, decay or warm restarts. We use the AdamW optimizer for all experiments with $\beta_1 = 0.9, \beta_2 = 0.999$, and a gradient clipping parameter of 0.1. To save memory, all models are trained with BF16 mixed precision training, with DDP level parallelism.

A.7.2 Reward Models

The HPSv2 [68] model is trained on the Human Preference Dataset v2. HPDv2 is a large-scale dataset with 798k binary preference choices for 434k images. Each pair contains two images generated by different models using the same prompt, and is annotated with a binary choice made by one annotator. The prompts are collected from DrawBench and DiffusionDB containing user-written prompts, the latter of which is 'sanitized' using ChatGPT to remove biases arising due to style words and leads to a reduced NSFW score. The PickScore [30] model is trained on the Pick-a-Pic dataset. The Pick-a-Pic dataset was created using a web application where users can write a prompt and are presented with two generated images, and they are asked to select their preferred option or indicate a tie if they have no strong opinion about either image. Although moderation is done to remove users who generate NSFW images or make judgements at a rapid pace (indicating low quality or random preference), the Pick-a-Pic model contains a lot of NSFW prompts. Consequently, we observe more NSFW generated images when finetuned with the Pickscore model compared to the HPSv2 model, even when prompts are not NSFW. However, the Pickscore model also generates more aesthetically pleasing images.



Answer the following questions and click the Submit button

Prompt

A photo of an Athenian vase with a painting of pandas playing basketball in the style of Egyptian hieroglyphics

Which images have better diversity (camera and object pose variation, background, different details)?

Which images are aesthetically better?

Which images are more aligned with the prompt (correct attributes, prompt understanding)?

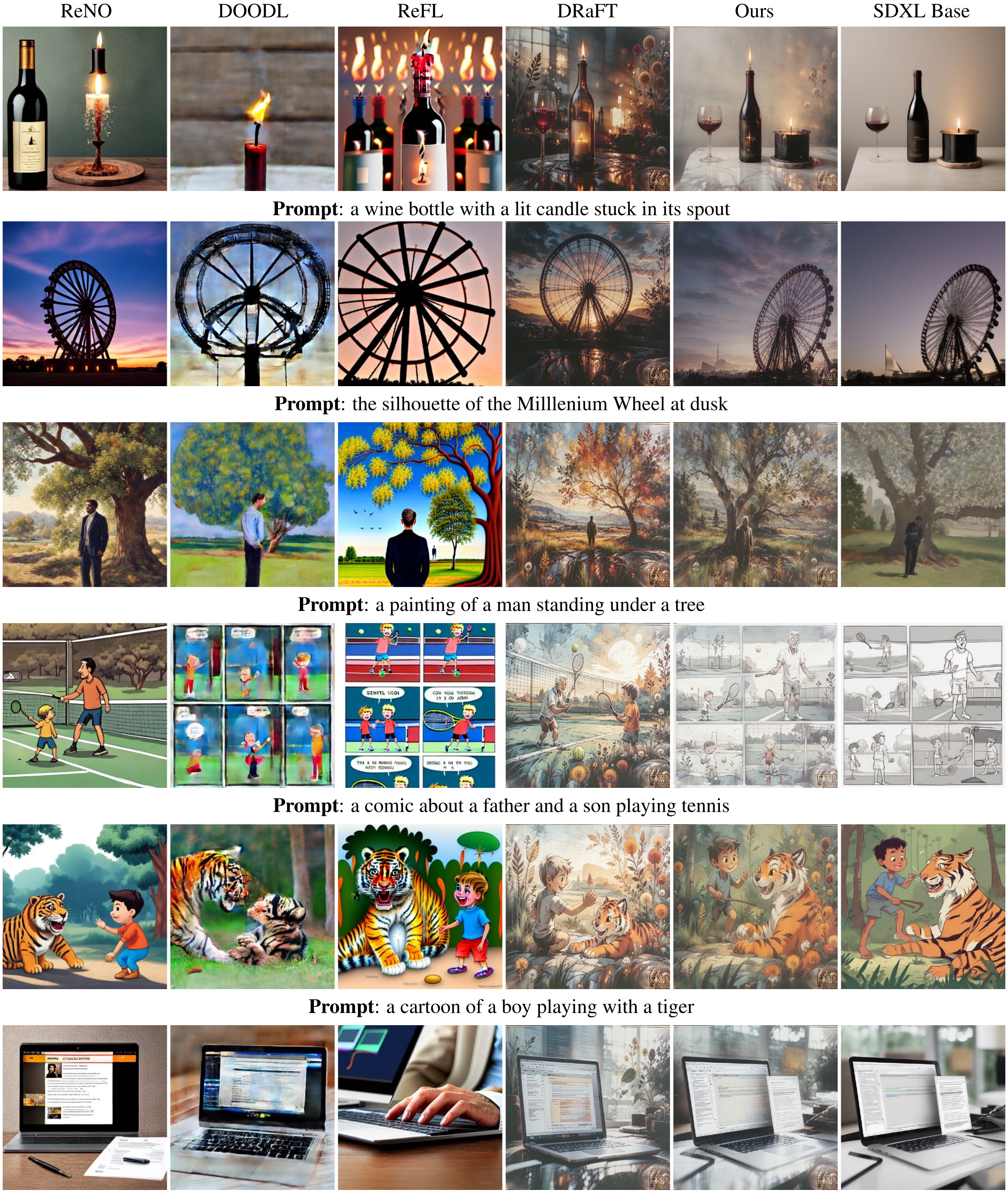
Figure 10. **Minimalistic web app designed for user study.** A web app built with Flask dynamically selects a prompt from the coverage dataset, selects nine random indices from the 50 generated images without replacement, randomly shuffles the order, and displays the images side by side. This is followed by three questions. Users click on either option and hit Submit. Upon hitting submit, the vote is recorded and a new set of images are shown.

A.7.3 Choice of γ

Unless the KL parameter λ or LoRA scaling parameter α' that are scalar quantities, AIG requires a function $\gamma(t)$. In this paper, we consider the family of functions

$$\gamma_{p,T}(t) = 1 - \left(\frac{T-t}{T}\right)^p$$

For $p = 1$, the weighing is simply linear, i.e. $\gamma_{1,T}(t) = \frac{t}{T}$. For $p > 1$, the power term quickly vanishes and the sampling dynamics are governed by the base model for more timesteps. For $p < 1$, the power term remains close to 1, therefore diminishing the effect of the base model in the earlier timesteps. We consider $p = 1, 1.25, 1.5, 2, 3, 4, 5$ for the ablations in the paper, and $p = 2$ for the qualitative studies. However, we note that more sophisticated γ scheduling is possible, i.e. $\gamma_{p,T}(t) = H\left(\frac{t}{T} - p\right)$, or $\gamma_{\kappa,T}(t) = \sigma\left(\kappa(t - T/2)\right)$ where H is the Heaviside step function, and σ is the sigmoid function. We leave exploration of these sophisticated scheduling functions to future work.



Prompt: a wine bottle with a lit candle stuck in its spout

Prompt: the silhouette of the Milllenium Wheel at dusk

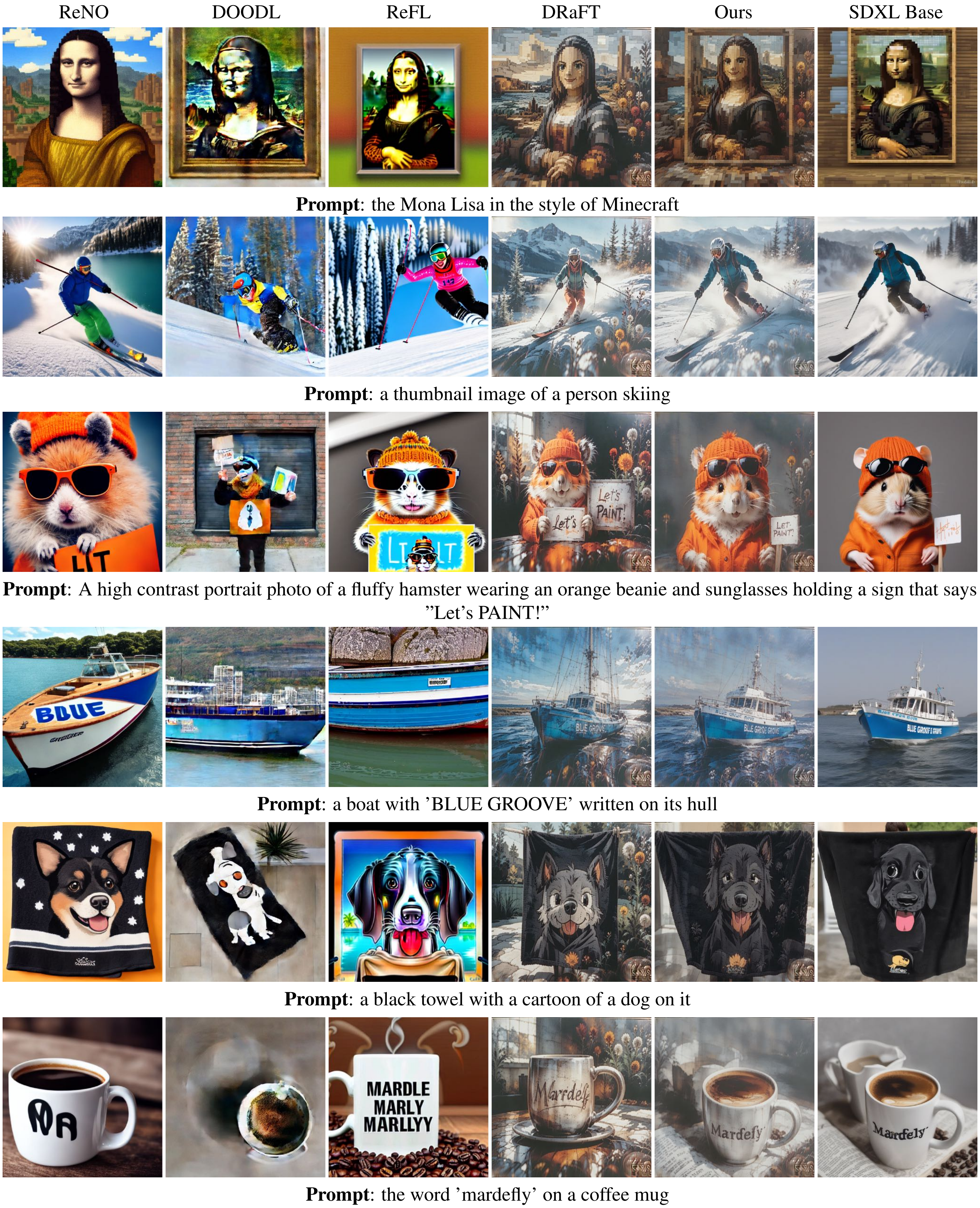
Prompt: a painting of a man standing under a tree

Prompt: a comic about a father and a son playing tennis

Prompt: a cartoon of a boy playing with a tiger

Prompt: a laptop screen showing a document being edited

Figure 11. **Qualitative comparison of reward finetuning methods on PartiPrompt prompt dataset**. Qualitatively, our model inherits the large-scale details from the base model, inheriting its diversity, but generated images follow stylistic aspects of the DRaFT model.



Prompt: the Mona Lisa in the style of Minecraft

Prompt: a thumbnail image of a person skiing

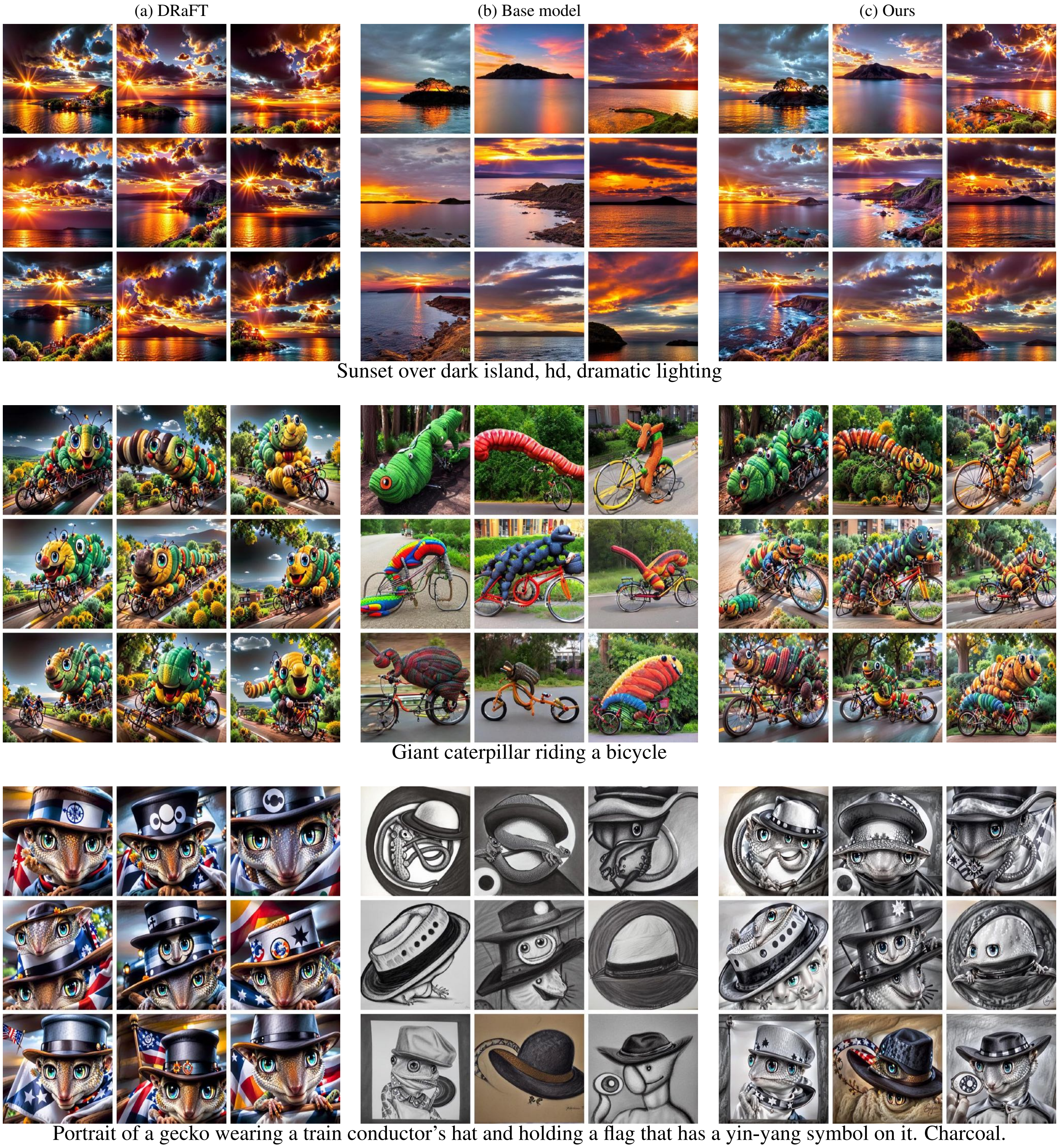
Prompt: A high contrast portrait photo of a fluffy hamster wearing an orange beanie and sunglasses holding a sign that says ”Let’s PAINT!”

Prompt: a boat with ’BLUE GROOVE’ written on its hull

Prompt: a black towel with a cartoon of a dog on it

Prompt: the word ’mardefly’ on a coffee mug

Figure 12. **Qualitative comparison of reward finetuning methods on PartiPrompt prompt dataset**. Qualitatively, our model inherits the large-scale details from the base model, inheriting its diversity, but generated images follow stylistic aspects of the DRaFT model.



(a) DRaFT

(b) Base model

(c) Ours

Sunset over dark island, hd, dramatic lighting

Giant caterpillar riding a bicycle

Portrait of a gecko wearing a train conductor’s hat and holding a flag that has a yin-yang symbol on it. Charcoal.

Figure 13. **Qualitative comparison of DRaFT and AIG**. Three columns of rows show set of nine images generated from the same seeds by the (a)DRaFT, (b)Base model, and (c)Our model. Our method preserves the diversity of details of different images, while adding aesthetic quality leading to both high rewards and high user preference.

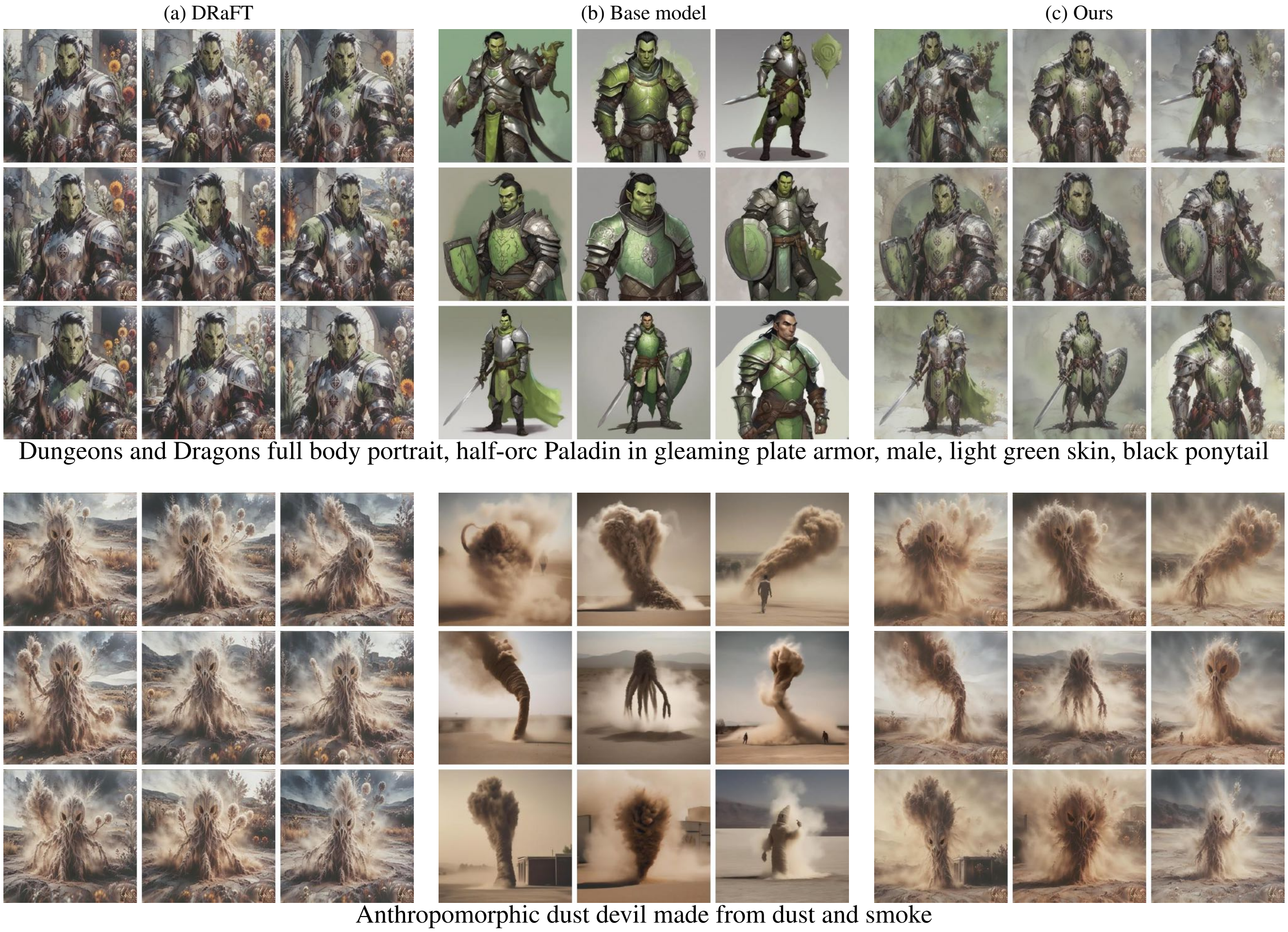
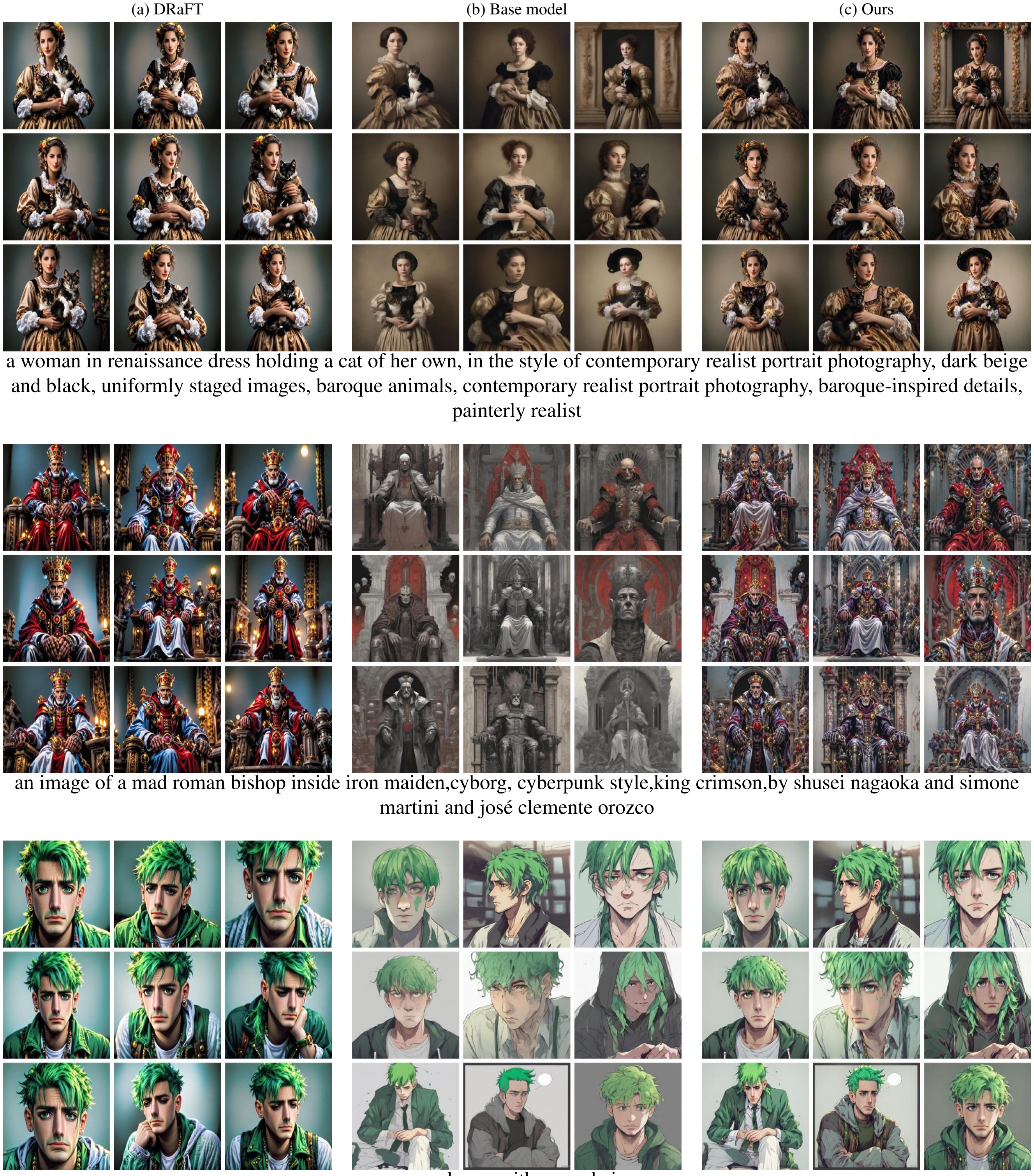

(a) DRaFT
(b) Base model
(c) Ours
Dungeons and Dragons full body portrait, half-orc Paladin in gleaming plate armor, male, light green skin, black ponytail
Anthropomorphic dust devil made from dust and smoke

Figure 14. **Qualitative comparison of DRaFT and AIG**. Three columns of rows show set of nine images generated from the same seeds by the (a)DRaFT, (b)Base model, and (c)Our model. Our method preserves the diversity of details of different images, while adding aesthetic quality leading to both high rewards and high user preference.



(a) DRaFT

(b) Base model

(c) Ours

a woman in renaissance dress holding a cat of her own, in the style of contemporary realist portrait photography, dark beige and black, uniformly staged images, baroque animals, contemporary realist portrait photography, baroque-inspired details, painterly realist

an image of a mad roman bishop inside iron maiden,cyborg, cyberpunk style,king crimson,by shusei nagaoka and simone martini and josé clemente orozco

a sad man with green hair

Figure 15. **Qualitative comparison of DRaFT and AIG**. Three columns of rows show set of nine images generated from the same seeds by the (a)DRaFT, (b)Base model, and (c)Our model. Our method preserves the diversity of details of different images, while adding aesthetic quality leading to both high rewards and high user preference.

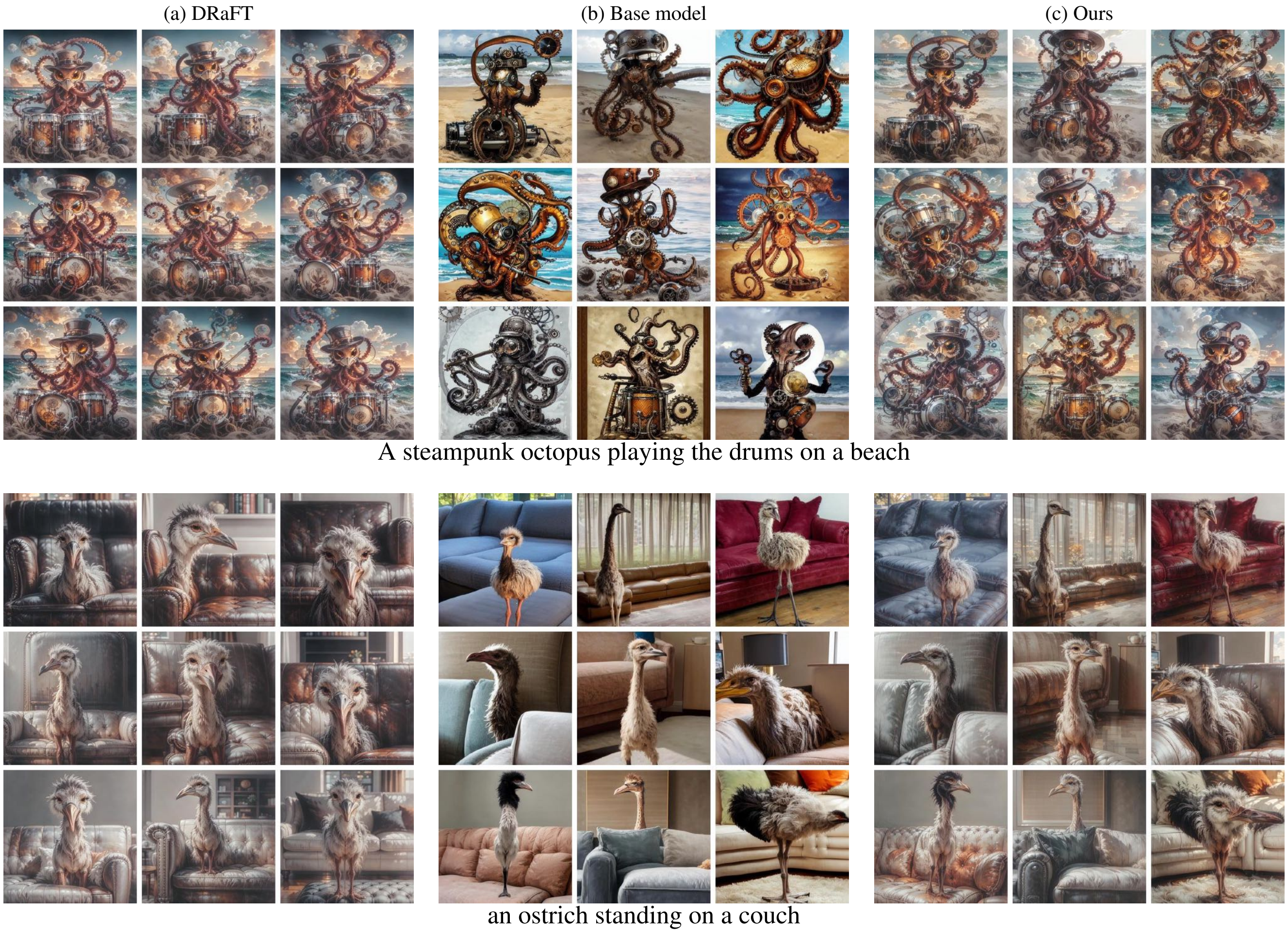

(a) DRaFT
(b) Base model
(c) Ours
A steampunk octopus playing the drums on a beach
an ostrich standing on a couch

Figure 16. **Qualitative comparison of DRaFT and AIG**. Three columns of rows show set of nine images generated from the same seeds by the (a)DRaFT, (b)Base model, and (c)Our model. Our method preserves the diversity of details of different images, while adding aesthetic quality leading to both high rewards and high user preference.

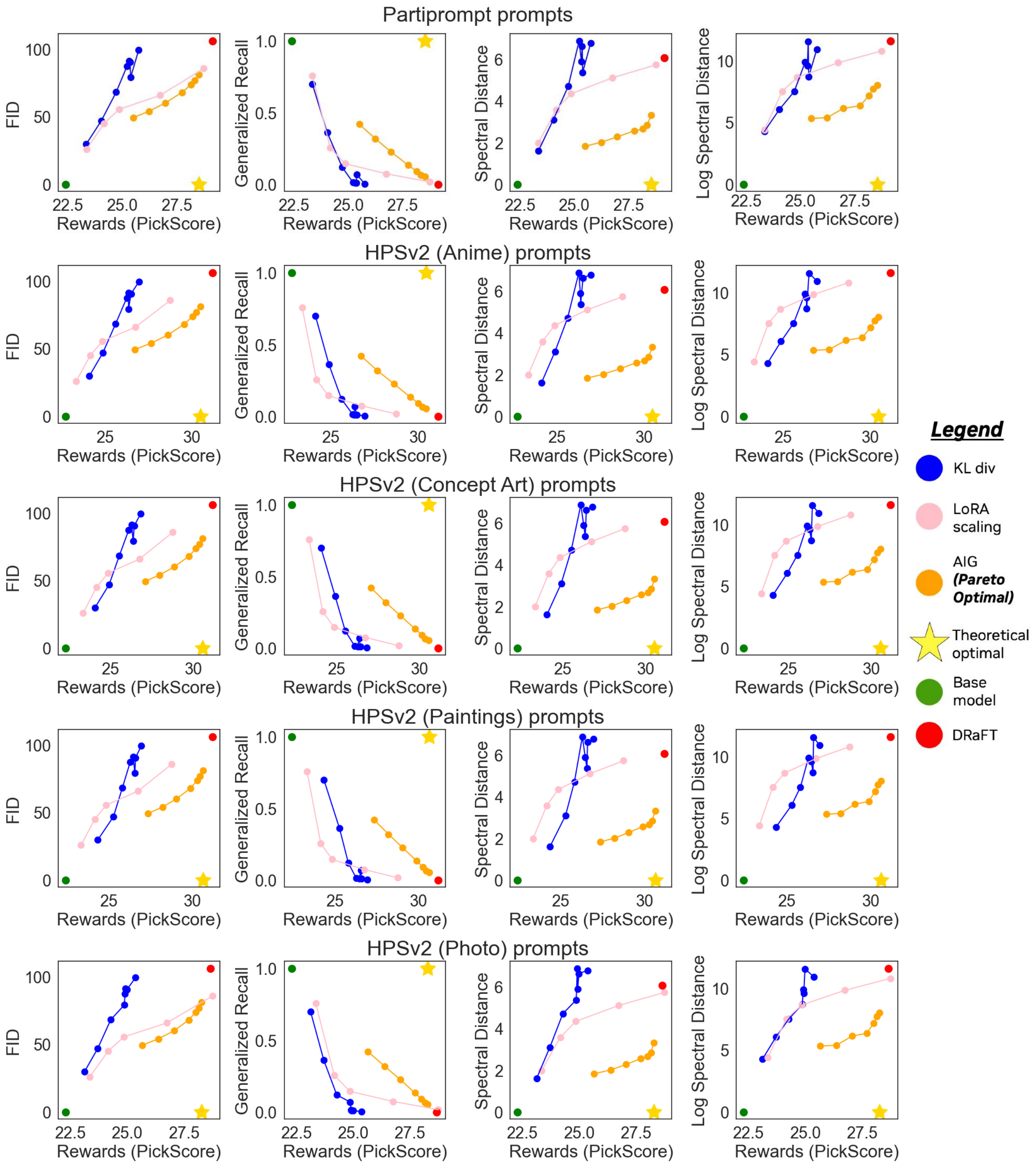


Figure 17. **Reward-diversity tradeoff for SDXL trained on PickScore**: **Green** represents the base model, **Red** represents DRaFTwith no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. An ideal baseline would achieve the highest reward (represented by DRaFT) as well as a complete match with the base distribution. For all measures for both PartiPrompt and HPSv2 subset prompts, AIG achieves Pareto-optimality.

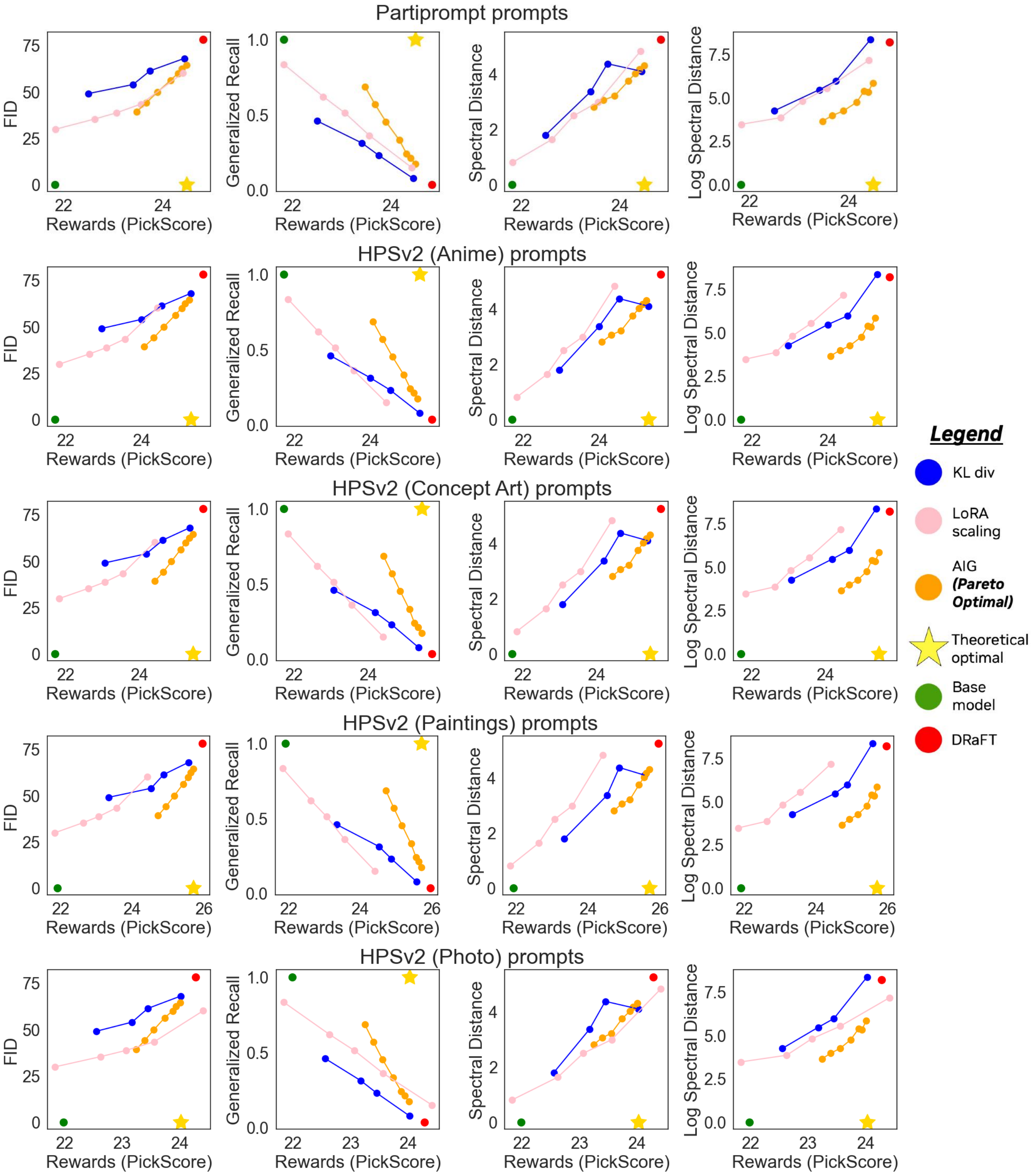


Figure 18. **Reward-diversity tradeoff for SDv1.4 trained on PickScore**: **Green** represents the base model, **Red** represents DRaFTwith no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. An ideal baseline would achieve the highest reward (represented by DRaFT) as well as a complete match with the base distribution. For all measures for both PartiPrompt and HPSv2 subset prompts, AIG achieves Pareto-optimality.

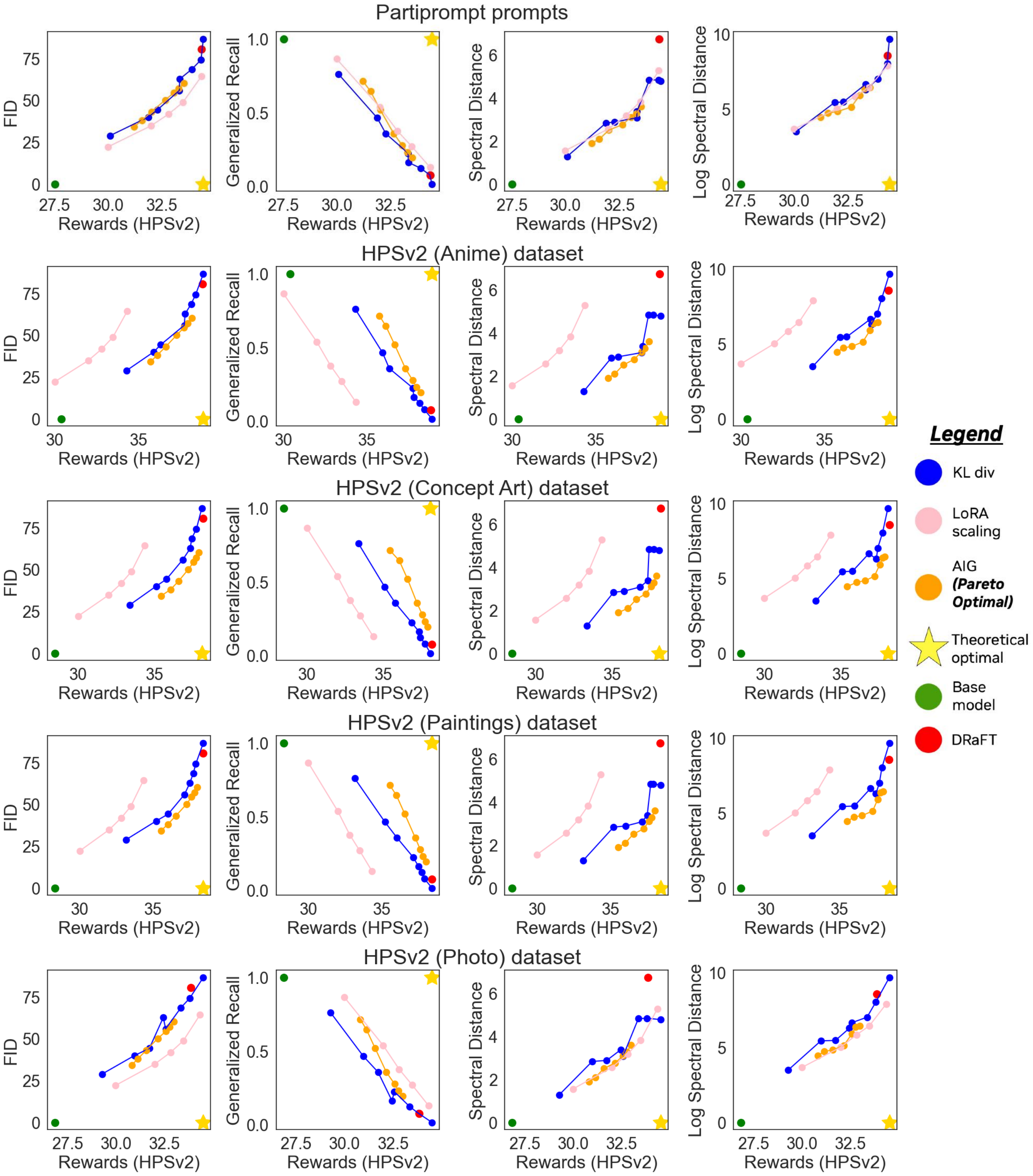


Figure 19. **Reward-diversity tradeoff for SDXL trained on HPSv2**: **Green** represents the base model, **Red** represents DRaFTwith no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. An ideal baseline would achieve the highest reward (represented by DRaFT) as well as a complete match with the base distribution. For all measures for both PartiPrompt and HPSv2 subset prompts, AIG achieves Pareto-optimality.

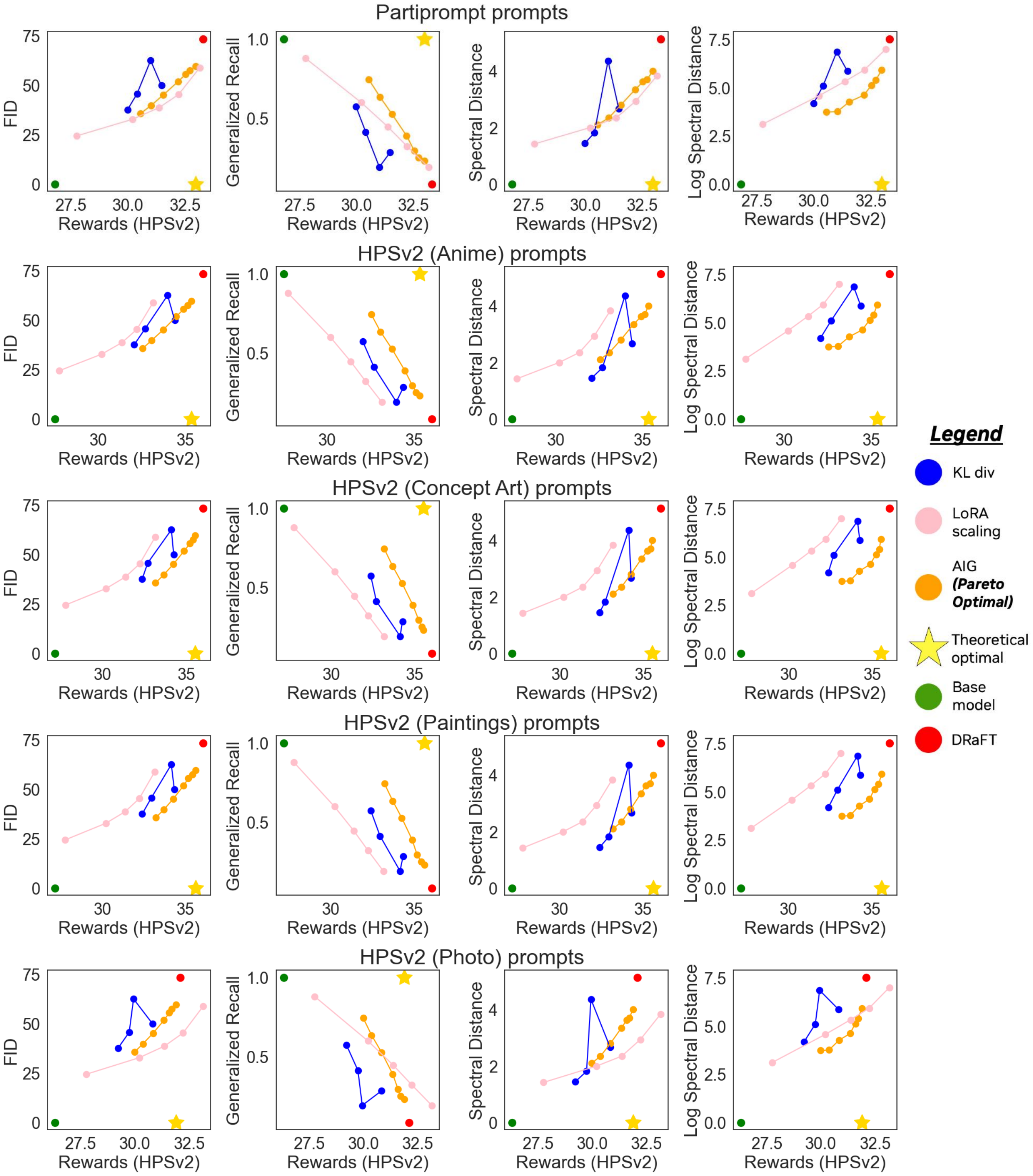
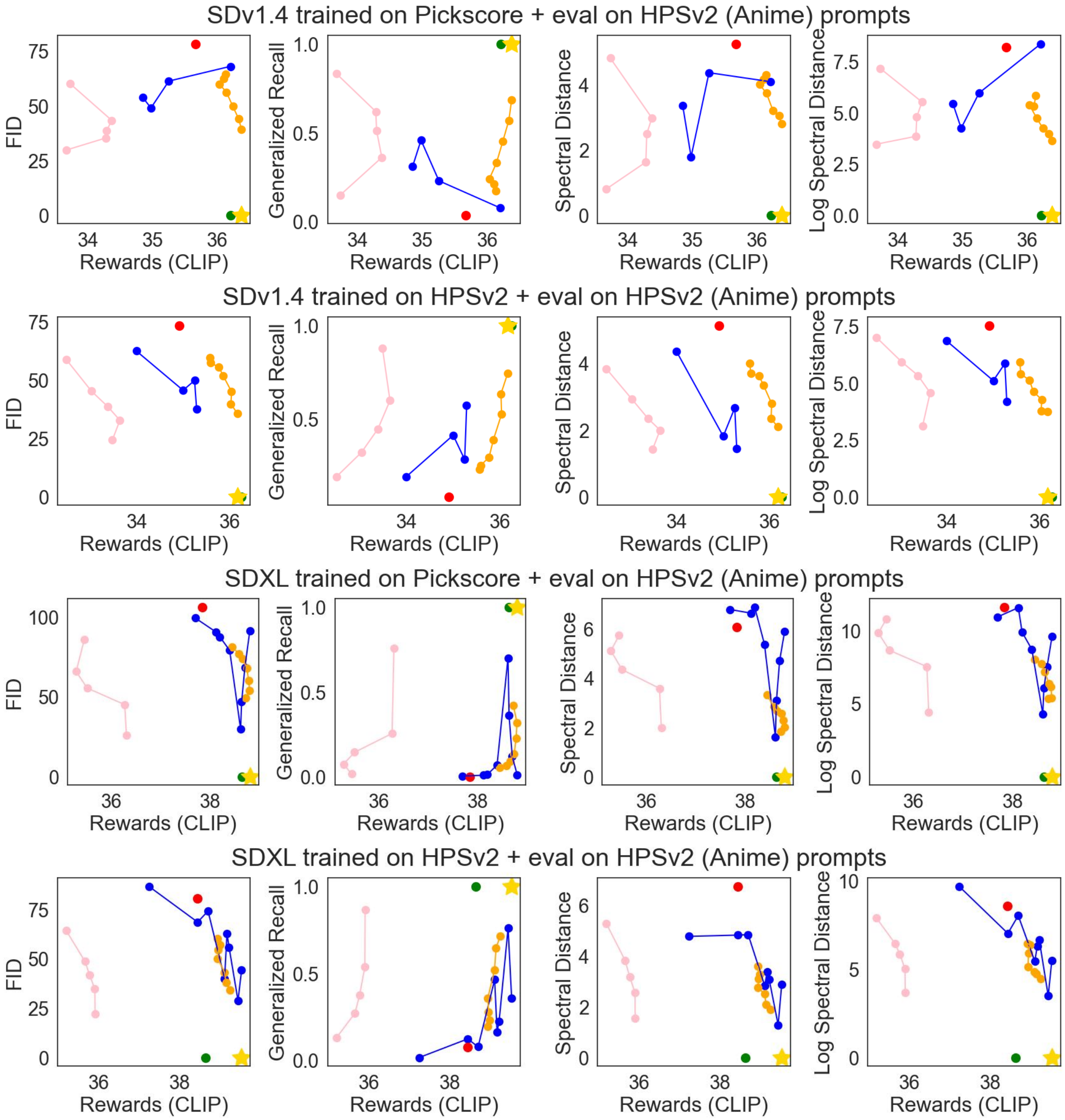


Figure 20. **Reward-diversity tradeoff for SDv1.4 trained on HPSv2**: **Green** represents the base model, **Red** represents DRaFTwith no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. An ideal baseline would achieve the highest reward (represented by DRaFT) as well as a complete match with the base distribution. For all measures for both PartiPrompt and HPSv2 subset prompts, AIG achieves Pareto-optimality.


SDv1.4 trained on Pickscore + eval on HPSv2 (Anime) prompts

SDv1.4 trained on HPSv2 + eval on HPSv2 (Anime) prompts

SDXL trained on Pickscore + eval on HPSv2 (Anime) prompts

SDXL trained on HPSv2 + eval on HPSv2 (Anime) prompts

Figure 21. **CLIP-Diversity tradeoff for configurations on HPSv2 anime prompts**: **Green** represents the base model, **Red** represents DRaFT with no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. AIG consistently outperforms LoRA scaling in CLIP alignment.

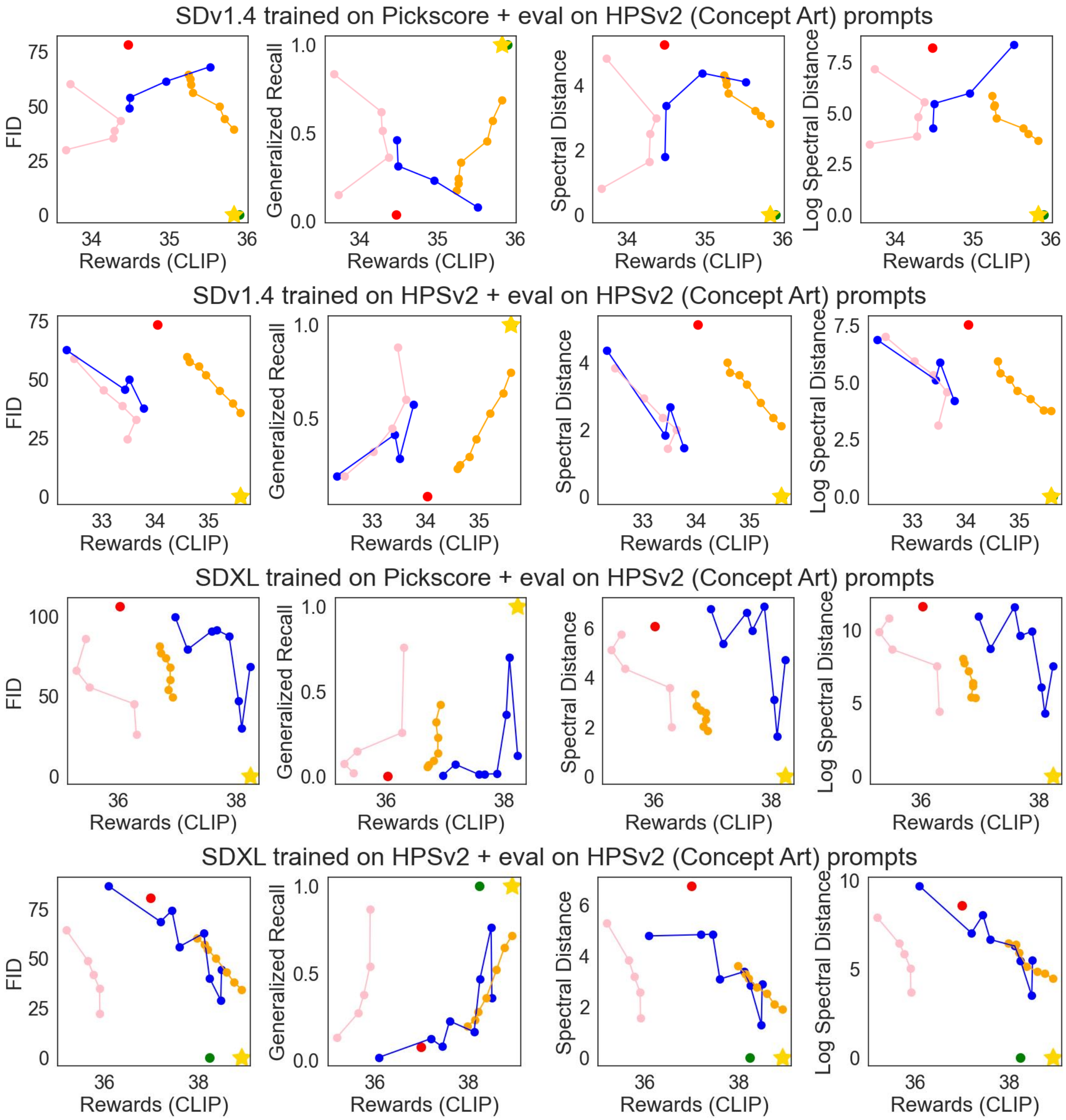
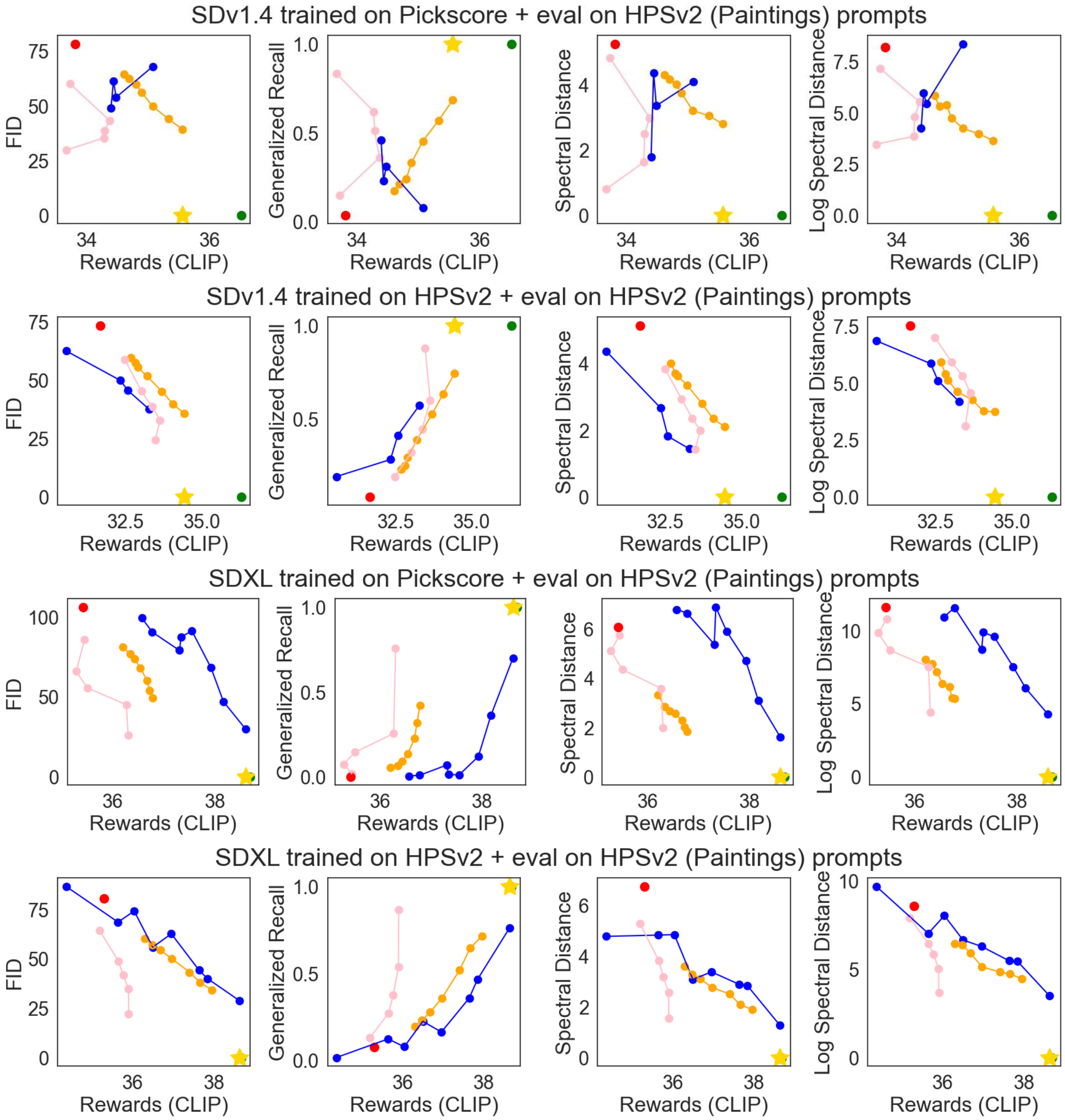


Figure 22. **CLIP-Diversity tradeoff for configurations on HPSv2 concept art prompts**: **Green** represents the base model, **Red** represents DRaFT with no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. AIG consistently outperforms LoRA scaling in CLIP alignment.



SDv1.4 trained on Pickscore + eval on HPSv2 (Paintings) prompts

SDv1.4 trained on HPSv2 + eval on HPSv2 (Paintings) prompts

SDXL trained on Pickscore + eval on HPSv2 (Paintings) prompts

SDXL trained on HPSv2 + eval on HPSv2 (Paintings) prompts

Figure 23. **CLIP-Diversity tradeoff for configurations on HPSv2 painting prompts**: **Green** represents the base model, **Red** represents DRaFT with no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. AIG consistently outperforms LoRA scaling in CLIP alignment.

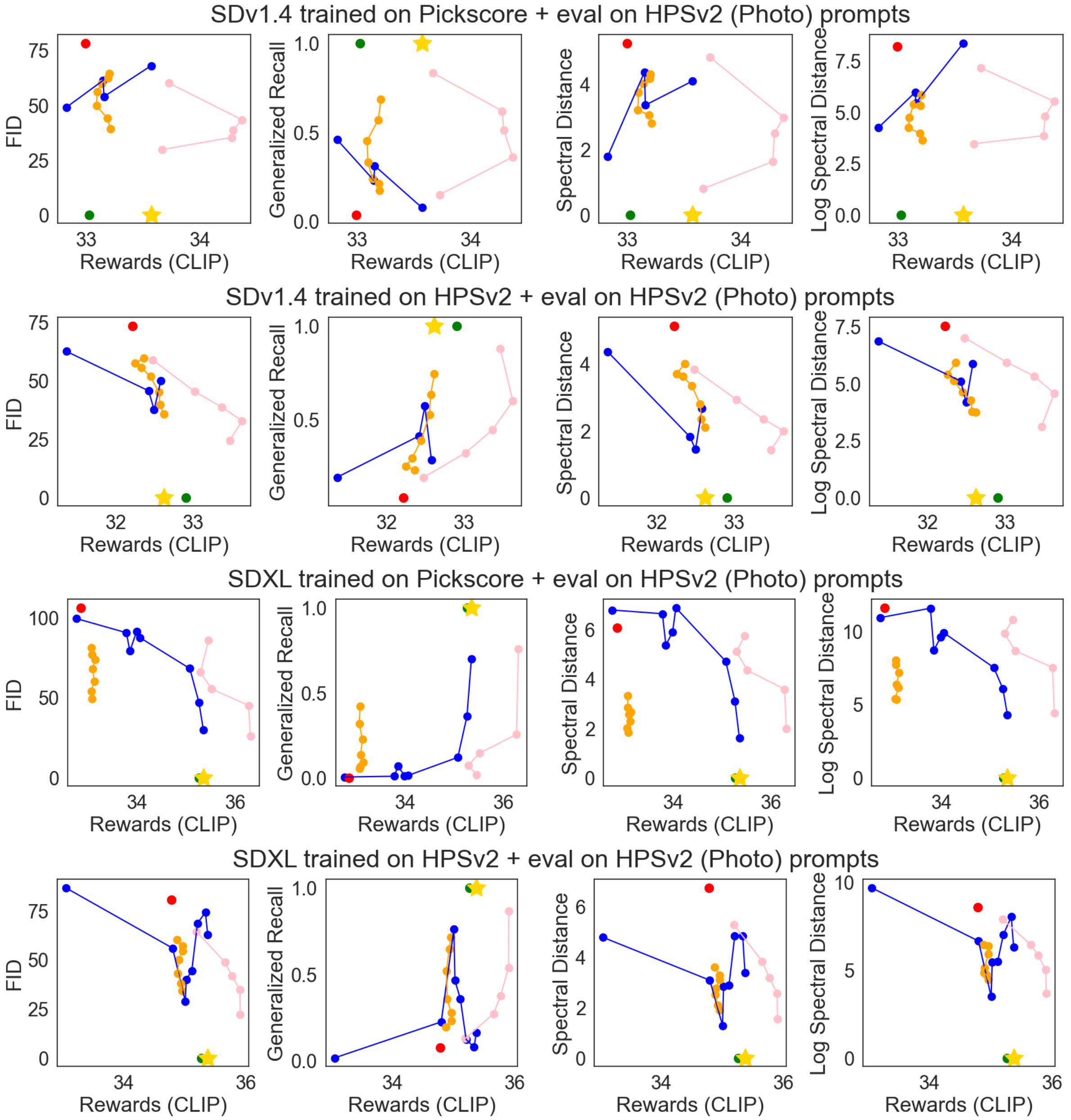


Figure 24. **CLIP-Diversity tradeoff for configurations on HPSv2 photo prompts**: **Green** represents the base model, **Red** represents DRaFT with no regularization, **Gold** star represents the ideal score. **Blue** represents different models with different KL regularization coefficients λ, **Pink** represents different amounts of LoRA scaling, and **Orange** represents different $\gamma(t)$ for AIG. AIG underperforms LoRA scaling in CLIP alignment.