Fine-Tuning Next-Scale Visual Autoregressive Models with Group Relative Policy Optimization

Matteo Gallici [1] **Haitz Sáez de Ocáriz Borde** [1]

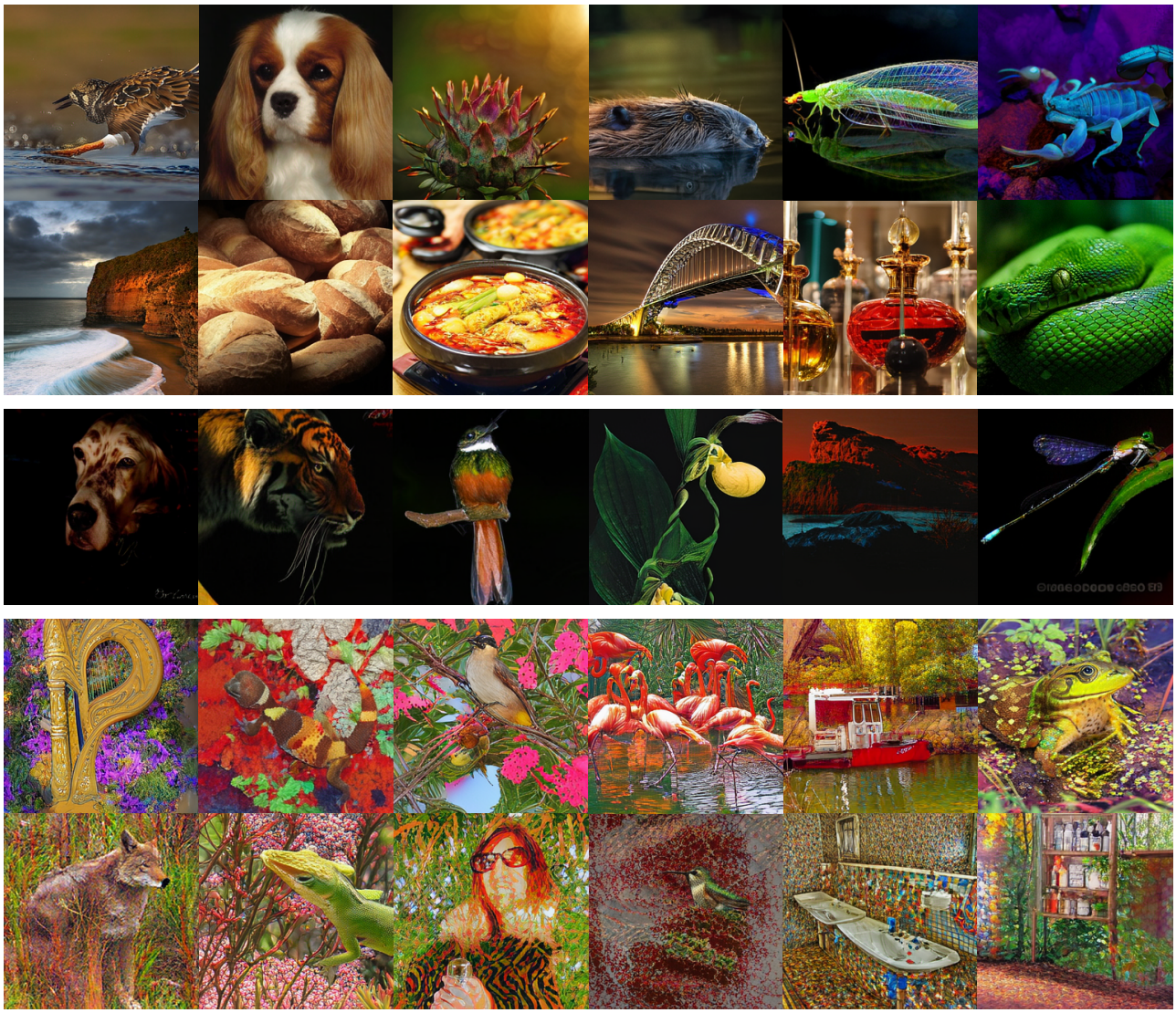

Figure 1. Images from VAR-*d*30 (Tian et al., 2024) fine-tuned with group relative policy optimization. At the top, samples from the model fine-tuned with an aesthetic reward. In the middle, samples from the model fine-tuned to produce paintings optimized with CLIP score. At the bottom, the model is fine-tuned with a combination of the two rewards. Given the notable scarcity of artworks in the ImageNet dataset used for pre-training VAR, the artistic patterns observed in the images should primarily be attributed to discoveries made during fine-tuning through reinforcement learning.

[1]Supermodel. Correspondence to: Haitz Sáez de Ocáriz Borde <ocariz@supermodel.ai>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

Abstract

Fine-tuning pre-trained generative models with Reinforcement Learning (RL) has emerged as an effective approach for aligning outputs more closely with nuanced human preferences. In this paper, we investigate the application of Group Relative Policy Optimization (GRPO) to fine-tune next-scale visual autoregressive (VAR) models. Our empirical results demonstrate that this approach enables alignment to intricate reward signals derived from aesthetic predictors and CLIP embeddings, significantly enhancing image quality and enabling precise control over the generation style. Interestingly, by leveraging CLIP, our method can help VAR models generalize beyond their initial ImageNet distribution: through RL-driven exploration, these models can generate images aligned with prompts referencing image styles that were absent during pre-training. In summary, we show that RL-based fine-tuning is both efficient and effective for VAR models, benefiting particularly from their fast inference speeds, which are advantageous for online sampling—an aspect that poses significant challenges for diffusion-based alternatives.

1. Introduction

For Artificial Intelligence (AI) systems to integrate successfully into real-world applications and products, it is crucial to design intelligent systems that support effective user interaction, align with human preferences, and exhibit ethical, user-centric behaviors. This alignment is particularly challenging in generative modeling; although recent breakthroughs have dramatically advanced image synthesis, ensuring generated images adhere to nuanced human aesthetic and semantic criteria remains difficult. In this context, Reinforcement Learning (RL)-based fine-tuning has emerged as a powerful tool for aligning pre-trained generative models with human preference, gaining prominence first in language (Ouyang et al., 2022), and more recently in the image domain (Yang et al., 2024; Black et al., 2024).

Current methods predominantly apply RL to diffusion-based models due to their state-of-the-art generation quality. However, these models typically incur substantial computational costs during sampling, significantly hampering their efficiency for online RL fine-tuning. Autoregressive models, particularly those leveraging a next-scale approach such as VAR (Tian et al., 2024), provide a promising alternative by enabling much faster inference while maintaining competitive image quality. Motivated by the advantages of autoregressive modeling, we explore Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a newly introduced RL variant based on Proximal Policy Optimization (PPO) (Schulman et al., 2017), to fine-tune VAR. GRPO eliminates the need for a separate value function estimator through grouped sampling, enhancing both efficiency and training stability. Concretely, our contributions are the following:

- We propose the *first framework integrating GRPO with next-scale visual autoregressive models (VAR)*, leveraging the computational efficiency and sampling speed of VAR.
- We empirically demonstrate that VAR models fine-tuned with GRPO significantly *improve image aesthetics and alignment* with complex reward signals derived from CLIP and human-rated aesthetic preferences.
- We provide evidence that GRPO fine-tuning enables VAR models to *generalize beyond the original ImageNet training distribution*, effectively aligning to prompts referencing unseen concepts and aesthetic criteria through RL-driven exploration.

2. Related Work

Before presenting our approach, we first discuss related work in image generative modeling and RL-based alignment of generative models, both in the context of language and image modeling.

2.1. Image Generative Models

Autoregressive (van den Oord et al., 2017), diffusion (Sohl-Dickstein et al., 2015), and flow matching models (Lipman et al., 2023), distinguish themselves from other generative modeling approaches, such as Variational Autoencoders (VAEs) (Kingma & Welling, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), by simplifying the task of generating samples from complex data distributions in a single forward pass into a series of smaller, more tractable intermediate predictions.

Diffusion and Flow Matching Diffusion (Sohl-Dickstein et al., 2015) and flow matching (Lipman et al., 2023) (or rectified flows (Liu et al., 2023)) have been highly successful in image generation using large scale models (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024). However, they require generating intermediate trajectories between the base and target distributions with the same dimensionality as their final output, which can become increasingly computationally expensive as the dimensionality of the data they are training on grows (even when operating in the latent space of an autoencoder). This can also make RL fine-tuning challenging due to slow sampling, as we will discuss later.

Autoregressive Image Models Autoregressive modeling (van den Oord et al., 2017), on the other hand, relies on a language-modeling-style next-token prediction approach, which can make inference substantially faster. Autoregression achieves this by representing the data as a token sequence and predicting each element successively. For instance, the original DALL-E model (Ramesh et al., 2021) was based on such an approach. In particular, a next-scale based variant of the original autoregressive paradigm has gained recent attention in the literature with the introduction of class-conditioned image models such a VAR (Tian et al., 2024), as well as text-to-image models like Infinity (Han et al., 2024), and Switti (Voronov et al., 2024), to name a few. This will be the focus of our paper.

2.2. Generative Model Alignment with Reinforcement Learning

Using RL to steer the behavior of large pre-trained models has proven very successful in language and has only recently gained traction in other data modalities.

Large Language Model Alignment Defining what a human considers to be a 'good' text sample can be very challenging to do via a loss function. Instead, LLMs are first pre-trained on a vast corpus of internet-scraped text data, minimizing a cross-entropy loss to acquire their base knowledge, and then later aligned through fine-tuning on instructions. Since the introduction of Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) to turn base language models into chatbots (e.g. GPT-3 (Brown et al., 2020) into ChatGPT), there has been a plethora of follow-up research that has ultimately lead to using RL to teach LLMs how to reason. RLHF leverages a reward model, that is, a neural network that takes in text tokens and returns a scalar value that numerically represents human preference. In this setup, the policy is the language model itself, generating text sequences based on a given prompt. Its action space is the token vocabulary size, and its observation space is all possible input token sequences. Finally, the reward function guides the learning process, combining feedback from the aforementioned reward model with a mechanism to prevent drastic changes to the policy, often based on the Kullback–Leibler (KL) divergence. We will be working with a similar setup in Section 4.

Image Model Alignment Aligning image diffusion models with similar techniques presents a challenge: their intractable exact likelihood computation makes it difficult to directly apply many conventional RL algorithms. While recent literature has proposed diffusion-specific frameworks—such as reward-weighted likelihood maximization (Lee et al., 2023), Direct Preference for Denoising Diffusion Policy Optimization (D3PO) (Yang et al., 2024), Denoising Diffusion Policy Optimization (DDPO) (Black et al., 2024), and very recently, GRPO-based approaches like Flow-GRPO (Liu et al., 2023) and Dance-GRPO (Xue et al., 2025)—we argue that this overcomplication would be removed if the image generative model were autoregressive. Additionally, the latter significantly improves the computational amenability of RL fine-tuning due to its fast sampling speed. Although recently (Wang et al., 2025) explored the use of GRPO in a simple autoregressive multi-modal model, the application of RL in next-scale autoregressive prediction remains unexplored.

3. Preliminaries

In this section, we review the next-token prediction paradigm and the next-scale formulation for image generation.

Next-token Prediction We use the term *token* to denote the index of an entry in the codebook (or vocabulary used by our model). Consider a sequence of tokens $[e_1, e_2, \ldots, e_T]$, where T denotes the sequence length. The next-token (causal) autoregressive model formulation assumes that the probability of a token e_t is conditioned solely on the preceding tokens $[e_1, e_2, \ldots, e_{t-1}]$, establishing a one-way dependency among them. This allows the likelihood of the entire sequence to be factorized as: $p([e_1, e_2, \ldots, e_T]) = \prod_{t=1}^{T} p(e_t \mid [e_1, e_2, \ldots, e_{t-1}])$. Autoregressive models learn to model data by maximizing the likelihood of observed sequences during training, $p_\theta([e_1, e_2, \ldots, e_T])$. Specifically during large-scale pre-training, the model's parameters θ are adjusted to minimize the negative log-likelihood, typically through the use of cross-entropy loss: $\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p_\theta([e_1, e_2, \ldots, e_T]_i)$, where $[e_1, e_2, \ldots, e_T]_i$ refers to the i-th training sequence out of a total of N. By minimizing $\mathcal{L}(\theta)$, the model learns to predict each token in the sequence based on its preceding context, a process commonly known as *next token prediction*. After training, the model can generate new sequences based on the learned probability distribution. This formulation is especially natural in language modeling, where the text exhibits a clear causal structure.

Visual Autoregressive Modeling Traditionally, the next-token prediction paradigm has been adapted for image generation by flattening visual samples, predominantly using a raster-scan approach, where image patches are arranged into a sequence (van den Oord et al., 2017). This disrupts the local spatial correlations in images and other visual data, such as videos, hinders the model's ability to learn bidirectional correlations within the sequence, and may also negatively affect zero-shot generalization. For instance, it restricts lower-right image patches to depend on upper-left patches but not vice versa. Visual AutoRe-

gressive modeling (VAR) (Tian et al., 2024) suggests using a more natural multi-scale, coarse-to-fine representation for images. In VAR the likelihood of the data sample is modeled as follows: $p([r_1, r_2, \ldots, r_K]) = \prod_{k=1}^{K} p(r_k \mid [r_1, r_2, \ldots, r_{k-1}])$, where r_k is used to denote an autoregressive unit at each scale k of the visual data representation consisting of $h_k \times w_k$ tokens, where $h_k, w_k \in \mathbb{N}_+$ and $h_K > h_{K-1} > ... > h_1$ and $w_K > w_{K-1} > ... > w_1$. In short, in this formulation we impose a causal autoregressive dependency at the resolution (or scale) level, rather than at the token level. That is, r_k is a collection of tokens instead of a single token. This, in turn, requires to construct an alternative latent space, where the continuous embedding is represented using quantized multi-scale residual embeddings. Similar to next-token prediction, once the transition probabilities have been learned, it can be used to generate new data samples via the token-to-codebook mappings and passing the final embedding reconstruction to a decoder.

4. Method

As mentioned, we specifically focus on applying RL to next-scale autoregressive image models. Inspired by successful RL applications in language modeling, we observe a natural alignment since both language models and autoregressive image generators follow a next-token prediction paradigm. Additionally, the superior efficiency and speed of autoregressive models make them particularly well-suited to online RL, where online sampling typically presents the main computational challenge.

4.1. Group Relative Policy Optimization (GRPO)

Among the techniques that were recently proposed to fine-tune LMMs with RL, *Group Relative Policy Optimization* (GRPO) (Shao et al., 2024) has emerged as a particularly effective method, offering improved training efficiency and greater stability by replacing the PPO value function estimator with group sampling. More specifically, given a class label c and a set of G outputs $\{o_1, o_2, \ldots, o_G\}$ sampled from the policy $\pi_{\theta_{old}}$, the objective of GRPO is to update π_θ by maximizing the following function:

$$\mathcal{J}(\theta) = \mathbb{E}_{o_i \sim \pi_{\theta_{old}}} \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min \Bigg(\frac{\pi_\theta(o_i \mid c)}{\pi_{\theta_{old}}(o_i \mid c)}, \text{clip}\Big(\frac{\pi_\theta(o_i \mid c)}{\pi_{\theta_{old}}(o_i \mid c)}, 1-\epsilon, 1+\epsilon \Big) A_i \Bigg) - \beta D_{\text{KL}} \Bigg], \tag{1}$$

where ϵ is the clipping threshold hyperparameter (Schulman et al., 2017), β controls the strength of the (per token) KL divergence regularization term with respect to the pre-trained policy $\pi_{\theta_{ref}}$ $D_{\text{KL}}(\pi_\theta \mid\mid \pi_{\theta_{ref}}) = \frac{\pi_{\theta_{ref}}(o_i|c)}{\pi_\theta(o_i|c)} - \log(\frac{\pi_{\theta_{ref}}(o_i|c)}{\pi_\theta(o_i|c)}) - 1$, and A_i represents the advantage estimate computed with the rewards obtained within each output group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, ..., r_G\})}{\text{std}(\{r_1, r_2, ..., r_G\})}. \tag{2}$$

4.2. GRPO for Next-Scale Autoregressive Image Models

Extending GRPO to next-scale autoregressive modelling is straightforward by keeping the Vector Quantized Variational Autoencoder (VQ-VAE) (van den Oord et al., 2017) decoder frozen and fine-tuning only the VAR model operating in its discrete latent space. Indeed, in this context, VAR functions exactly like an LLM, except that at each inference step it produces logits for a set of tokens (the next-scale tokens, r_k) rather than just one. Therefore, the same GRPO formalization used for LLMs can be applied directly. Specifically, in this context, $\pi_{\theta_{\text{ref}}}$ denotes the pre-trained VAR, while π_θ and $\pi_{\theta_{\text{old}}}$ represents the VAR model that we are currently optimizing and sampling from, respectively. The outputs $\{o_1, o_2, \ldots, o_G\}$ correspond to the set of tokens produced autoregressively by $\pi_{\theta_{\text{old}}}$ during sampling, which are subsequently decoded into images by the VQVAE and evaluated by a specific image reward model.

It is important to note that during training, we use simple multinomial sampling from the VAR output logits scaled with a temperature value τ of 0.7 to balance exploration and exploitation. We deliberately avoid using classifier-free guidance (CFG) (Ho & Salimans, 2022), top-p or top-k sampling, beam search, or other techniques that could improve sample quality in order to maintain a tractable sampling distribution during optimization. During inference, however, we do use CFG along with top-p or top-k sampling. More concretly, we employ the same hyperparameters as those used in the original VAR model (Tian et al., 2024).

Since rewards are typically computed with respect to the final image output, we encounter a credit assignment problem: the individual contribution of each token to the final score is unknown. To address this, we compute the advantages for each sample and apply the same advantage to all tokens. This approximation enables us to maintain a per-token KL penalty, which we found essential for preserving the generation quality of the original model.

5. Experiments

For our experiments, we employed the original VAR pre-trained on ImageNet. This model can generate images conditioned on one of the 1,000 original ImageNet labels (it is a class-conditioned image model). We focused on the smallest and largest variants of the model, both of which generate images at a resolution of 256×256: VAR-$d16$ (310M parameters) and VAR-$d30$ (2B parameters).

Unless otherwise explicitly stated, all experiments were

conducted by sampling a batch of random labels from the original 1k ImageNet labels and using 8 Nvidia H100 GPUs. Additionally, for all experiments, the following hyperparameters were used: a batch size of 32 labels for sampling, 16 groups, minibatch size of 32 during optimization, a learning rate of 10^{-4}, $\beta = 0.2$, temperature of 0.7, and $\epsilon = 0.2$.

5.1. Preliminary Experiments: Brightness

In order to test our implementation, we designed a simple toy experiment where we used GRPO to align the model to generate images with high or low brightness. Brightness is straightforward and inexpensive to compute, and it provides a stable and clear reward signal: $\frac{1}{HW}\sum_{i,j}(0.2989R_{i,j} + 0.5870G_{i,j} + 0.1140B_{i,j})$ where $H = h_K$ and $W = w_K$ denote image height and width, respectively, and R, G, B are the RGB channel values.

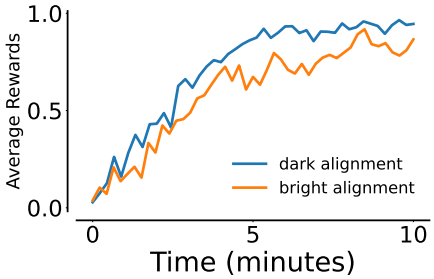

Figure 2. Fine-Tuning VAR-$d16$ with GRPO on dark/bright alignment requires only 10 minutes with a single H100 GPU.

To align the model to generate images with high brightness, we set up a scalar reward giving a value of 1 when brightness is greater than or equal to 0.8, and 0 otherwise. For aligning the model to generate dark images, we give a reward of 1 when brightness is lower than 0.2. As shown in Figure 2, using GRPO we are able to perfectly align VAR to generate bright or dark images in less than 10 minutes using a single GPU. Despite the simplicity of the task, the method exhibits surprisingly strong sample efficiency and stability, reinforcing our hypothesis that the faster inference speed of VAR makes it well-suited for online RL in image generation. The results of this experiment are shown in Figure 3.

5.2. Improving Image Aesthetics

To evaluate the capability of GRPO to align the VAR model to more complex scores, we employ LAION's aesthetic predictor V2 (Schuhmann et al., 2022) (AES) as a proxy for human preferences. AES combines CLIP embeddings with a multi-layer perceptron and was trained on 176,000 images rated by users on an aesthetic scale from 1 to 10, with a score of 10 marking an image as a work of art. For this task, we trained both VAR-$d16$ and VAR-$d30$ for 40,000 steps, which corresponds to a single epoch of VAR's supervised

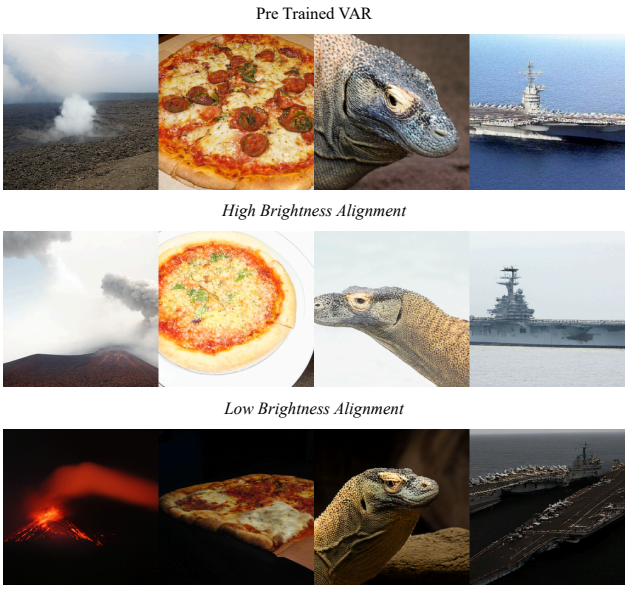

Figure 3. Visual results of the toy experiment aligning the model to generate only bright or dark images.

training on ImageNet. This amounted to 16 and 40 hours of training, respectively, using 8 NVIDIA H100 GPUs. As shown in Figure 4, both models improve their AES scores by approximately 1 point, with VAR-$d30$ increasing from 4.80 (pre-trained) to 5.80. A progression of the outputs obtained during the optimization process is shown in Figure 5.

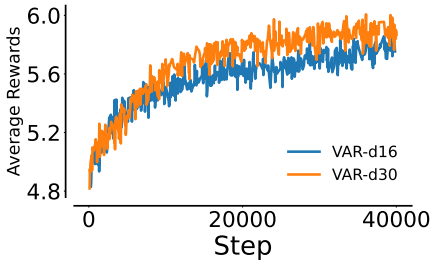

Figure 4. Fine-Tuning VAR-$d16$ and VAR-$d30$ with GRPO using Aesthetic Reward.

In order to validate that the fine-tuned models are still capable of generating images corresponding to the ImageNet labels, we can no longer rely on the FID score, as the RL optimization has altered the generative distribution of ImageNet. Instead, we employ a standard ImageNet classifier (ResNet50 IMAGENET1K_V2 (Vryniotis, 2021)) and observe the difference in accuracy between samples from the reference and fine-tuned models. In Table 1, we report the Aesthetic Score computed from 10,000 samples (10 per class) for the pre-trained and fine-tuned models, along with the ResNet50 top-5 accuracy. Notably, the models fine-tuned with GRPO maintain an accuracy above 90% while generating images with an Aesthetic Score approximately 1

point higher than that of the ImageNet validation set. This observation confirms that the model did not diverge significantly from the ImageNet distribution during training.

In Table 1, we also report the scores for VAR-$d16$ fine-tuned using only 50% of the ImageNet labels. In this case, the AES score increased more moderately by 0.5 points. However, it is noteworthy that the AES score also increased by 0.5 points for the labels not seen during training, suggesting that the model generalized effectively during the fine-tuning process.

Table 1. Comparison of models by Aesthetic Score and ResNet50 Accuracy@5. Notice that the very high accuracy observed for the pre-trained model reflects its ability to reproduce the distribution of the ImageNet training set, where ResNet50 IMAGENET1K_V2 likely achieves near-perfect accuracy. VAR-$d16$ GRPO (seen 50% labels) was fine-tuned using 50% of the ImageNet classes, while the other GRPO models used all possible labels.

Model	Aesthetic Score	ResNet50 Accuracy@5
ImageNet Validation Set	4.91	95.43%
VAR-$d16$ (Tian et al., 2024)	4.64	97.8%
VAR-$d16$ GRPO (seen 50% labels)	5.10	97.7%
VAR-$d16$ GRPO (unseen 50% labels)	5.09	97.9%
VAR-$d16$ GRPO (seen 100% labels)	**5.55**	95.92%
VAR-$d30$ (Tian et al., 2024)	4.80	99.8%
VAR-$d30$ GRPO (seen 100% labels)	**5.80**	90%

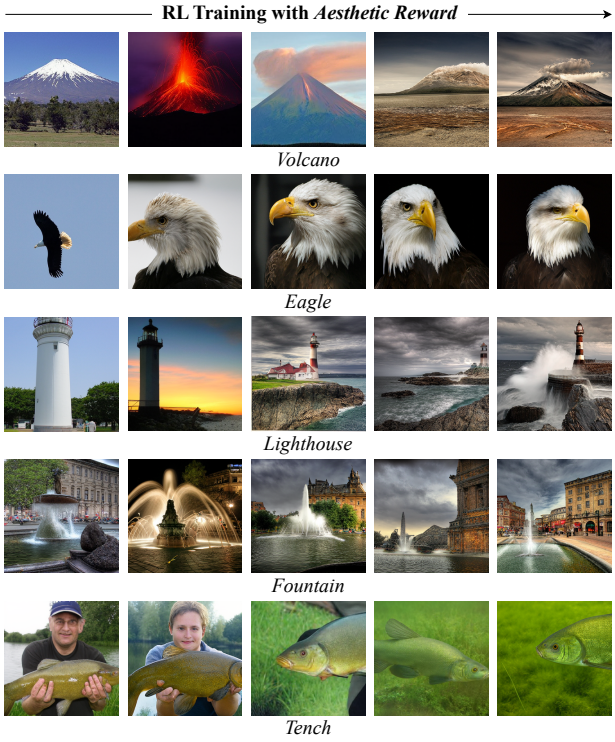

Figure 5. Sampling progress of VAR-$d30$ during GRPO optimization with aesthetic reward.

5.3. Learning to Paint with CLIP Alignment

To further evaluate the capabilities of our method, we employ the CLIP Score (Radford et al., 2021) to measure the similarity between the generated samples and a fixed text prompt. In this experiment, our goal is to assess the ability of GRPO to align a generative model with respect to a different modality (text), even though this modality does not directly condition the generative process. Specifically, the VAR model remains conditioned solely on the input labels, while the CLIP Score is computed relative to a fixed text prompt.

We begin by conditioning the CLIP score on the prompt *"an old picture"*. Since the ImageNet training set contains images of this type, the model quickly generates samples that align with the style. We therefore move to a more challenging prompt: *"a painting"*. Note that the original ImageNet dataset discouraged paintings or drawings, since annotators were instructed to collect: "photos only, no painting, no drawings, etc." (Deng, 2012). This means that, in order to align with this kind of prompt, the model should in principle generate samples that are out of the training distribution. Nevertheless, as shown in Figure 6 after some iterations, the policy finds ways to increase the "painting" CLIP Score. This suggests that, at least in part, the model really learned something new through RL exploration and exploitation, i.e., *it learned artistic patterns without demonstrations*. Obviously, this kind of hypothesis should be explored more deeply, but the preliminary results obtained in this context seem to indicate that online RL could, in principle, align an image generative model even further than what has been seen in the training data.

We display the results in Figure 9. These experiments were performed with VAR-$d30$ fine-tuned for 10k steps, corresponding to around 10 hours of compute in 8 H100 GPUs.

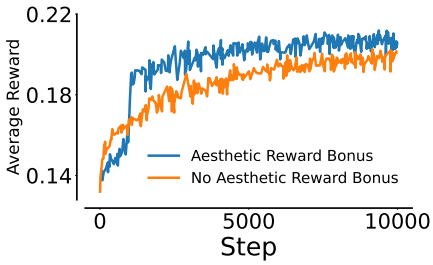

Figure 6. CLIP Score measuring alignment with the prompt "A painting" during fine-tuning, with and without the application of the Aesthetic Reward bonus (excluded from the plots).

5.4. Ablations

Finally, we provide a short discussion on hyperparameter ablations which may be useful for future researchers.

KL Penalty The importance of the KL penalty in GRPO for image generation is illustrated in Figure 7. Without sufficient regularization, the model easily finds a reward-hacking policy and loses the ability to generate images properly conditioned on the labels. The KL term also acts as an entropy bonus, discouraging the model from generating identical or highly similar images even for different labels.

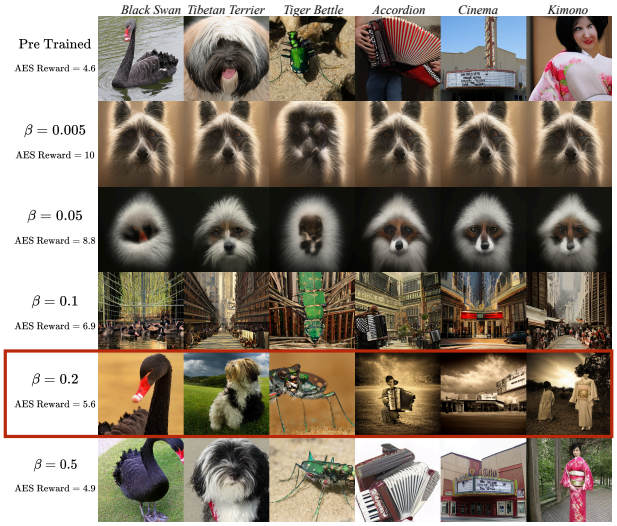

Figure 7. Effect of β KL coefficient when training with GRPO.

Group Optimization In Figure 8, we demonstrate how group optimization effectively aids training. Fixing a compute time budget, we find that increasing the number of groups during sampling monotonically enhances the performance of GRPO, at least up to 16 groups.

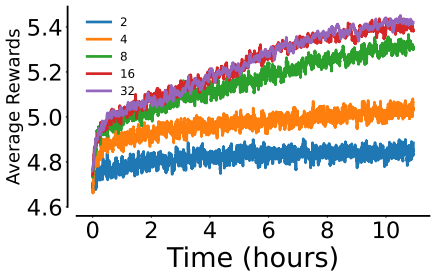

Figure 8. Effect of varying the number of groups during the sampling phase of GRPO (Aesthetic Reward optimization).

6. Conclusion

In this work, we have demonstrated the effectiveness and efficiency of GRPO for aligning next-scale visual autoregressive models. A natural next step is to extend this method to text-to-image VAR-based models (Han et al., 2024; Voronov et al., 2024).

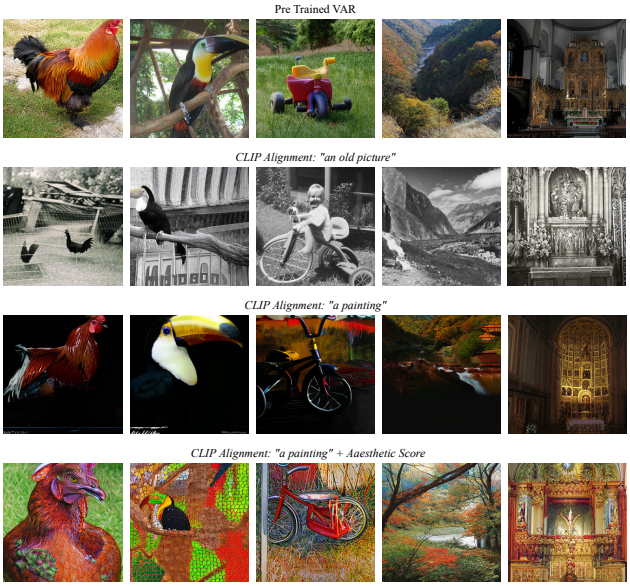

Figure 9. Experiments conducted to align VAR with a fixed prompt using the CLIP Score. In this setup, VAR itself is not conditioned on the prompt; only the reward model is. Since VAR was trained on ImageNet, the training data predominantly consists of photographs (no paintings, no drawings). This means that the artistic effects observed in the samples generated when aligning the model with the caption *"a painting"* should be attributed primarily to RL-driven exploration and optimization, as if the model effectively "learned to paint" during fine-tuning. Notice how the model adopts a simple yet effective approach to increase the "picture" reward by incorporating a technique reminiscent of Caravaggio's chiaroscuro. This effect is suppressed when an Aesthetic Score is added to the reward, resulting in more colorful and stylistically diverse images.

References

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning, 2024. URL https://arxiv.org/abs/2305.13301.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Deng, J. *Large scale visual recognition*. PhD thesis, Princeton University, 2012.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., and Rombach, R. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Han, J., Liu, J., Jiang, Y., Yan, B., Zhang, Y., Yuan, Z., Peng, B., and Liu, X. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2412.04431.

Ho, J. and Salimans, T. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback, 2023. URL https://arxiv.org/abs/2302.12192.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Sohl-Dickstein, J. N., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *ArXiv*, abs/1503.03585, 2015. URL https://api.semanticscholar.org/CorpusID:14888175.

Tian, K., Jiang, Y., Yuan, Z., Peng, B., and Wang, L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. 2024.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. *ArXiv*, abs/1711.00937, 2017. URL https://api.semanticscholar.org/CorpusID:20282961.

Voronov, A., Kuznedelev, D., Khoroshikh, M., Khrulkov, V., and Baranchuk, D. Switti: Designing scale-wise transformers for text-to-image synthesis. *arXiv preprint arXiv:2412.01819*, 2024.

Vryniotis, V. How to train state-of-the-art models using torchvision's latest primitives. *PyTorch*, 2021.

Wang, J., Tian, Z., Wang, X., Zhang, X., Huang, W., Wu, Z., and Jiang, Y.-G. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025.

Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., and Luo, P. Dancegrpo: Unleashing grpo on visual generation, 2025. URL https://arxiv.org/abs/2505.07818.

Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Li, Q., Shen, W., Zhu, X., and Li, X. Using human feedback to fine-tune diffusion models without any reward model, 2024. URL https://arxiv.org/abs/2311.13231.