G^{3}AFT: Glance Guided Gradient Aligned Fine-Tuning for Visual Autoregressive Models

Jiayi Zhang[1,2], Xiefan Guo[1,2], Xinzhu Ma[2], and Di Huang[1,2⋆]

[1] State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing 100191, China
[2] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
{zhangjyi,xfguo,xinzhuma,dhuang}@buaa.edu.cn

Abstract. Autoregressive models in vision have gained widespread attention under the scalable next-token prediction paradigm, representing images as sequences of discrete tokens. Yet their optimization remains challenging due to the non-differentiability of discrete token sampling and the prohibitive memory costs of full-sequence fine-tuning. We propose G^{3}AFT, a novel end-to-end fine-tuning framework for discrete-token autoregressive models. Within this framework, we introduce two complementary techniques, *i.e.* Global Scope Gradient Recording (GSGR) and Noise Suppressed Gradient Projection (NSGP). GSGR selectively updates periodically sampled tokens by leveraging coarse global signals that summarize the overall layout and long-range structure of the image. This design alleviates the prohibitive memory overhead of full-sequence fine-tuning while maintaining long-range coherence. In parallel, NSGP enhances surrogate gradient flow by explicitly aligning updates with salient sampling alternatives and filtering out spurious signals introduced by softmax relaxations. By suppressing noisy gradient components and reinforcing causally consistent updates, NSGP provides a more faithful approximation to discrete sampling, ensuring faithful gradient propagation across long autoregressive trajectories. To the best of our knowledge, this is the first method under the next-token prediction paradigm for discrete-token autoregressive image generation that enables fully differentiable optimization across the entire generation process. Extensive experiments validate that G^{3}AFT achieves superior image quality and quantitative performance, providing a practical framework for fully differentiable fine-tuning of autoregressive image generation models. Code is available at https://github.com/Wotoosh/G3AFT.

Keywords: Text-to-Image Generation · Autoregressive Image Generation · End-to-End Fine-Tuning

1 Introduction

Image generation [3, 8, 47] aims to synthesize realistic or creative visual content from given inputs such as noise or text prompts. This task requires fine-grained

⋆ Corresponding author

Fig. 1: Visualization of fine-tuning on different objectives. Fine-tuning for aesthetic quality yields smoother, conventionally pleasing images with balanced composition, whereas fine-tuning for human-preference rewards emphasizes vivid color contrasts and heightened visual intensity, producing outputs that appear more striking and expressive.

control over fidelity, diversity, and semantic consistency to enable applications in digital art, gaming, scientific visualization, *etc.* Autoregressive visual models [40] have emerged as a promising approach for modeling images in discrete spaces. By employing quantized autoencoders to transform images into sequences of visual tokens, these models adopt the next-token prediction strategy that has proven highly effective in natural language processing [16, 26, 28]. This formulation provides a unified generative framework across modalities, enabling vision to inherit the scalability and reasoning consistency that autoregressive approaches have already demonstrated in language domains.

However, in autoregressive image generation, practical decoding pipelines often incorporate discrete sampling or diffusion-style iterative refinement to enhance perceptual quality, and these non-differentiable steps break the computational graph and prevent gradients from propagating across the full generative trajectory. Moreover, as each image is represented by a long sequence of hundreds of tokens, full-sequence fine-tuning incurs substantial memory and computational overhead, further constraining the practicality of end-to-end fine-tuning. Consequently, dependencies across distant positions in the token sequence cannot be effectively aligned through gradient-based optimization. While reinforcement learning [1, 18, 38] can in principle bypass broken gradients, since it does not rely

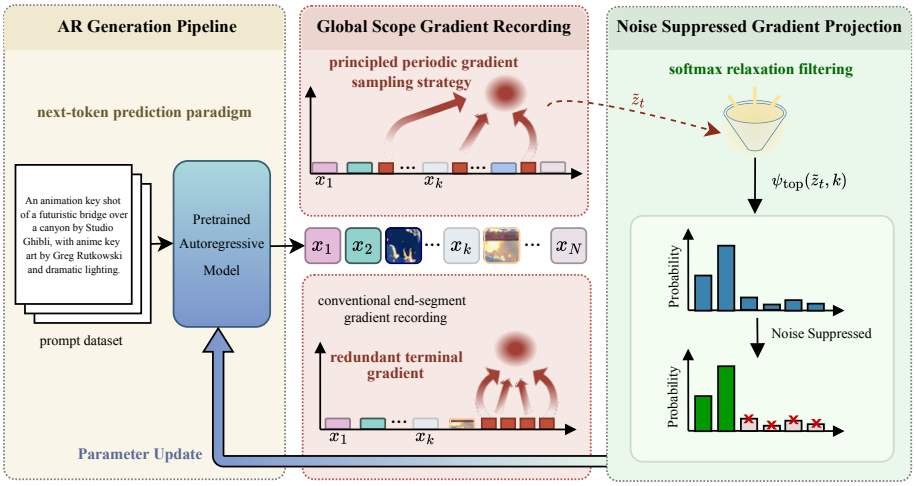

Fig. 2: An overview of G^{3}AFT. Text prompts are processed by an autoregressive model to produce image tokens, where gradients are selectively recorded through Global Scope Gradient Recording to balance local detail with global structural coherence. Meanwhile, Noise-Suppressed Gradient Projection refines surrogate gradients by detecting salient candidates and suppressing spurious updates, thereby providing a closer approximation to discrete sampling and ensuring causal consistency.

on end-to-end backpropagation but instead optimizes policies through sampled trajectories. However, in long-horizon visual generation, RL methods often suffer from high-variance credit assignment and slow convergence, which limits their practicality compared to direct gradient-based optimization.

In this paper, we introduce G^{3}AFT, a novel fine-tuning framework for autoregressive image generation within the next-token prediction paradigm. An overview of G^{3}AFT is shown in Figure 2. Our approach addresses three central challenges in autoregressive fine-tuning: the loss of gradient flow caused by non-differentiable discrete sampling, the relaxation-induced artifacts that arise when softmax surrogates are used to restore differentiability, and the prohibitive memory cost of full-sequence optimization. Within this framework, Global Scope Gradient Recording (GSGR) introduces a softmax-based relaxation to recover differentiability and selectively updates periodically sampled tokens using coarse global signals, mitigating the computational overhead of full-sequence optimization while preserving structural consistency. Noise Suppressed Gradient Projection (NSGP) further refines surrogate gradient flow by aligning updates with salient sampling alternatives and removing relaxation-induced artifacts, supporting robust credit assignment across long autoregressive trajectories.

To validate our approach, we conduct experiments on autoregressive image generation models that follow the next-token prediction paradigm, using both HPSv2 [39] reward and aesthetic reward [35] signals. The qualitative differences between these objectives are illustrated in Figure 1, which highlights

how different reward signals steer the visual characteristics of generated images. Fine-tuning on aesthetic quality produces smoother and more compositionally balanced outputs, whereas human-preference rewards tend to amplify contrast and visual intensity. Beyond these qualitative trends, extensive experiments across multiple benchmarks demonstrate that our G^3AFT framework consistently outperforms existing baselines and achieves state-of-the-art performance in reward-based fine-tuning for discrete-token autoregressive image generation. Our contributions are summarized as three-fold:

- We introduce the first end-to-end fine-tuning framework for discrete-token autoregressive image models within the next-token prediction paradigm, allowing gradients to propagate through the entire generative trajectory.
- We introduce two complementary modules that make this framework practical and effective, i) Global Scope Gradient Recording (GSGR), which performs selective gradient updates on periodically sampled tokens to enable memory-efficient optimization over long autoregressive sequences, and ii) Noise Suppressed Gradient Projection (NSGP), which suppresses spurious gradients arising from softmax-based relaxations and preserves reliable gradient signals along the autoregressive trajectory.
- Extensive experiments demonstrate that our approach enables stable, fully differentiable fine-tuning for autoregressive image generation models following the next-token prediction paradigm, consistently improving reward alignment and visual quality over existing approaches.

2 Related Work

2.1 Text-to-Image Generation

Text-to-Image Generation (T2I) [14,31,48] aims to synthesize realistic images directly from natural language descriptions. As a fundamental task in multimodal learning [29,41,46], it bridges vision and language, enabling intuitive human–AI interaction and supporting applications in art design, content creation, education, and accessibility. Early T2I methods relied on template-based generation and feature matching, which were limited in realism and diversity. The introduction of Generative Adversarial Networks (GANs) [11] marked a breakthrough in T2I. Early GAN-based works such as Obj-GAN [20], DM-GAN [52], and SD-GAN [22] significantly improved image fidelity, semantic alignment, and controllability, laying the foundation for subsequent transformer and diffusion-based approaches.

More recently, diffusion models have emerged as the dominant paradigm in text-to-image generation, delivering state-of-the-art performance in both visual quality and diversity. Earlier diffusion approaches typically operated in pixel space, which required substantial computational resources. To overcome this limitation, Latent Diffusion Models (LDM) [34] perform diffusion in the latent space of pre-trained encoders, enabling denoising in a low-dimensional representation while mitigating high-variance latent spaces through regularization. Building

upon this idea, GLIDE [25] introduced classifier-free guidance to improve controllability and sample quality. More recently, eDiff-I [2] advanced text-to-image diffusion by leveraging an ensemble of expert models, achieving higher fidelity and diversity in generation.

2.2 Discrete Token Image Modeling

A central development in bridging vision and language modeling is the discretization of continuous image signals into compact token representations. Instead of operating directly on raw pixels or latent feature maps, discrete token image modeling typically partitions an image into local patches, each of which is mapped to a discrete index from a finite codebook. In this way, the entire image can be reformulated as a sequence of tokens, analogous to words in a sentence. This patch-based tokenization not only compresses high-dimensional visual data but also provides a structured representation that can be seamlessly integrated into sequence modeling frameworks, thereby establishing a unified token-based interface across modalities.

Early work such as VQ-VAE [37] pioneered the use of vector quantization to map image patches into discrete indices, effectively compressing high-dimensional inputs while preserving semantic fidelity. Subsequent extensions, including VQ-VAE-2 [32], introduced hierarchical tokenization schemes to capture multi-scale structures, while VQ-GAN [44] integrated adversarial objectives and perceptual losses to improve reconstruction quality. These advances demonstrated that discrete tokenization not only reduces redundancy but also provides a structured representation amenable to large-scale generative modeling.

Beyond compression, discrete token image modeling establishes a principled foundation for unified models [42,49] that jointly address understanding and generation. By reformulating images into sequences of discrete tokens, this paradigm aligns the representation of vision with that of language, enabling both modalities to be processed within a single sequence modeling framework. Such a unified token-based interface paves the way for models that can seamlessly integrate comprehension and synthesis, supporting tasks ranging from text-to-image generation to image captioning [10, 15, 24] and visual question answering [9, 17, 23].

2.3 Autoregressive Image Generation

Building upon discrete tokenization, autoregressive image generation formulates synthesis as a sequence modeling problem. In this paradigm, an image is represented as an ordered sequence of discrete tokens, and the generative model predicts each token conditioned on its predecessors. This approach directly parallels language modeling, where words are generated one by one in a coherent sequence, thereby enabling the reuse of powerful autoregressive architectures such as Transformers for visual generation. Beyond image synthesis, autoregressive models have also proven effective across diverse generative domains, underscoring their versatility and broad applicability [12, 21, 33, 43].

Early autoregressive models demonstrated the feasibility of this formulation by applying sequence prediction to quantized image tokens. Systems such as DALL·E [30] and CogView [7] exemplified the scalability of this approach, showing that large-scale training on multimodal corpora could yield models capable of producing diverse and semantically aligned images from textual prompts. LlamaGen [36] advances this direction by adapting LLaMA-style autoregressive transformers to discrete visual token modeling, providing a simple and scalable framework for image generation. Building on this trajectory, Janus-Pro [5] extends the paradigm with larger capacity and refined training strategies, achieving stronger stability and instruction-following in text-to-image generation.

However, the sequential nature of autoregression poses inherent challenges. Capturing long-range dependencies across hundreds of tokens often results in optimization difficulties, as gradients must traverse extended autoregressive trajectories. In addition, discrete tokenization introduces non-differentiable sampling operations that disrupt end-to-end gradient flow. These limitations motivate our proposed G^{3}AFT framework, which establishes the first fully differentiable fine-tuning paradigm for discrete-token autoregressive image generation. Within this framework, we introduce Global Scope Gradient Recording (GSGR) and Noise Suppressed Gradient Projection (NSGP), two complementary techniques that improve efficiency and ensure faithful gradient propagation.

3 Method

3.1 Preliminaries

Recent advances in generative modeling have explored discrete representation spaces as a unified paradigm for both understanding and generation. Instead of operating directly on continuous pixels, images are first mapped into sequences of discrete tokens through vector quantization [45, 51] or learned codebooks [19, 50]. This tokenization allows the application of sequence modeling techniques originally developed for natural language processing, and provides a compact, interpretable representation that bridges vision and language.

Autoregressive models provide a principled framework for generative modeling in discrete spaces. For a sequence of discrete tokens $x = (x_1, x_2, \ldots, x_N)$, the joint distribution can be factorized into a product of conditional probabilities:

$$p(x) = \prod_{i=1}^{N} p(x_i \mid x_{<i}; \theta), \tag{1}$$

where $x_{<i} = (x_1, \ldots, x_{i-1})$ denotes all previously generated tokens and θ represents the model parameters. The conventional training objective is to minimize the negative log-likelihood (NLL):

$$L(\theta) = -\sum_{i=1}^{N} \log p_\theta(x_i \mid x_{<i}). \tag{2}$$

Fig. 3: Comparison between periodic gradient sampling in GSGR and conventional end-segment gradient recording. Traditional practices collect gradients only from the final portion of the sequence, leading to redundant adjacent signals. GSGR instead samples gradients at regular intervals across the trajectory, capturing broader context with reduced redundancy.

This objective supervises each token prediction independently under teacher forcing, rather than evaluating the quality of the final generated image. As a result, the optimization signal does not propagate through the full generative trajectory.

This paradigm has been successfully adapted to vision by mapping images into one-dimensional sequences of discrete tokens, such as pixels, patches, or latent codes. Through vector quantization or learned codebooks, images are discretized into compact sequences, which are then modeled by decoder-only Transformers that predict each token step by step. During inference, sampling from $p_\theta(x)$ proceeds sequentially: the model generates tokens one by one until a complete sequence is obtained. The generated sequence is subsequently decoded back into the image space, enabling high fidelity synthesis. Because both text and images can be represented as discrete token sequences, autoregressive models naturally support multimodal tasks and provide a unified framework for understanding and generation.

3.2 Global Scope Gradient Recording

In prior works [6, 27] on continuous diffusion-based image generation, it is often assumed that later timesteps encode richer semantic information, and gradients are therefore recorded primarily from the final steps. However, while this assumption is well-motivated in continuous diffusion models, where timesteps correspond to progressively refined noise levels, it does not naturally extend to discrete-token autoregressive generation, in which each step produces a single token and adjacent tokens often convey overlapping information. Selecting only the final portion of the generative trajectory is equivalent to sampling a contiguous block of tokens near the sequence end, which are typically adjacent and therefore highly redundant, failing to capture global context.

Based on the need to capture global information while reducing redundancy inherent in conventional gradient recording, we present Global Scope Gradient Recording (GSGR), as illustrated in Figure 3. GSGR adopts a principled periodic gradient sampling strategy that allocates gradient collection across the entire generative trajectory, thereby yielding representative optimization signals with reduced memory overhead. Unlike approaches [6, 27] that concentrate solely on the final steps, GSGR records gradients at regular intervals, providing a global snapshot of the trajectory that conveys coarse yet representative signals. By distributing gradient collection in this manner, GSGR not only mitigates local redundancy but also enhances optimization stability by aligning training with the overall structure of the generated image.

Our goal is to fine-tune the parameters θ of a pre-trained model such that optimization reflects global structure rather than redundant local signals. Concretely, given a token sequence $\{t_1, \ldots, t_N\}$, GSGR aggregates gradients at regularly spaced positions determined by the subsampling set:

$$\mathcal{S}_k^r = \{\, i \mid i = mk + r,\ m = 0, \ldots, \lfloor (N - r)/k \rfloor \,\}, \quad 0 \le r < k. \tag{3}$$

For each training sample, we randomly draw an offset r and take $\mathcal{S}_k$ to be the corresponding subsampling set $\mathcal{S}_k^r$. Formally, the aggregated objective is defined as:

$$J(\theta) = \frac{1}{|\mathcal{S}_k|} \sum_{i=1}^{N} \mathbf{1}_{\{i \in \mathcal{S}_k\}} \cdot \nabla_\theta L(t_i; \theta), \tag{4}$$

where $\mathbf{1}_{\{i \in \mathcal{S}_k\}}$ is an indicator function that selects indices according to the periodic rule. This objective achieves a balance between efficiency and coverage. By subsampling gradients across the trajectory, GSGR reduces memory overhead while ensuring that optimization is guided by signals representative of the entire generative process.

It is worth noting that the periodic subsampling rule defined above represents only one possible instantiation of gradient aggregation. The choice of $\mathcal{S}_k^t$ can be adapted to alternative criteria beyond uniform strides. While strided selection offers simplicity and efficiency, other strategies may emphasize tokens that carry higher semantic uncertainty or richer contextual information. We further investigate such alternatives in Section 4.1, where entropy-driven selection is introduced as a complementary approach.

In practice, GSGR improves fine-tuning by mitigating overfitting to redundant local patterns and fostering coherence between local token generation and global image structure. This mechanism is particularly effective for long sequences, where naive gradient recording incurs prohibitive memory costs and fails to capture the global information distributed across the trajectory.

3.3 Noise Suppressed Gradient Projection

Noise Suppressed Gradient Projection (NSGP) is motivated by the fundamental challenge of discrete token sampling in autoregressive image generation. Each

Algorithm 1: G^{3}AFT (Glance Guided Gradient Aligned Fine-Tuning)

Input: Pre-trained autoregressive model weights θ, prompt dataset $\mathcal{C}$, tokenizer $\mathcal{T}$, stride parameter s, candidate threshold τ

while *not converged* **do**

- Sample prompt $c \sim \mathcal{C}$;
- Generate token sequence $x_{1:T} = \text{Model}(c; \theta)$;
- `// Global Scope Gradient Recording (GSGR)`
- Initialize gradient buffer $\hat{G} \leftarrow \emptyset$;
- Define selection function $\phi_{\text{select}}(s)$ to determine recording steps;
- Define candidate function $\psi_{\text{top}}(\tau)$ to select salient candidates;
- **for** $t = 1$ **to** T **do**
  - Compute loss $\mathcal{L}_t = \ell(x_t, c)$;
  - **if** $t \in \phi_{select}(s)$ **then**
    - Compute raw gradient $g_t = \nabla_\theta \mathcal{L}_t$;
    - `// Noise-Suppressed Gradient Projection (NSGP)`
    - Compute relaxed logits $\tilde{z}_t = \text{softmax}(z_t)$;
    - Select salient set $\mathcal{S}_t = \psi_{\text{top}}(\tilde{z}_t, \tau)$;
    - Retain gradient components $g_t^{\mathcal{S}} = g_t[\mathcal{S}_t]$;
    - Append $g_t^{\mathcal{S}}$ to $\hat{G}$;
- Aggregate salient gradients $\hat{g} = \sum_t g_t^{\mathcal{S}}$;
- Update model parameters: $\theta \leftarrow \theta - \eta \hat{g}$;

step in the sequence requires drawing a token from a categorical distribution, and this sampling operation is inherently non-differentiable, which prevents gradients from being propagated through the sampling step. In this work, we introduce a softmax-based relaxation that serves as a differentiable surrogate for discrete sampling, allowing gradients to flow through the sampling operation and making backpropagation feasible.

While this relaxation enables differentiability, it also introduces a mismatch between the surrogate gradient and the causal structure of the sampling process. Gradients are distributed across all logits, including those that are almost impossible to be sampled, which results in spurious and non-causal updates. Such noise destabilizes optimization and obscures the correct credit assignment along long autoregressive trajectories. Our innovation arises from the need to better approximate the gradient flow of discrete token sampling itself, ensuring that differentiability does not come at the cost of causal consistency.

To achieve this, NSGP refines the surrogate gradient by projecting it onto the causal subspace defined by the effective support of the sampling distribution. In practice, this entails suppressing gradient contributions from logits that are irrelevant to the sampled outcome and preserving only those aligned with the sampling mechanism. The resulting gradients are cleaner and more faithful to the discrete process they are intended to approximate. Intuitively, NSGP functions as a filter in the optimization pipeline, ensuring that while softmax relaxations

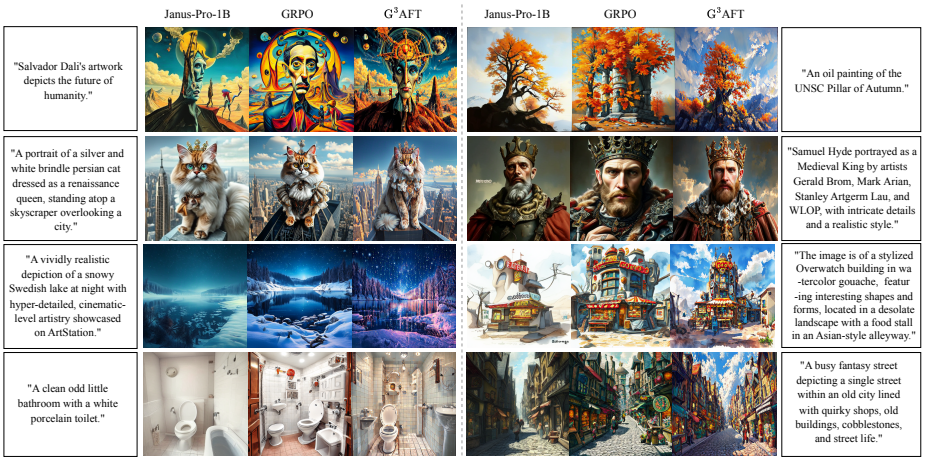

Fig. 4: Visualizations on the HPDv2 dataset comparing Janus-Pro-1B [5], GRPO [13], and our proposed G^{3}AFT under identical prompts. Each row corresponds to a distinct prompt, while columns present generations from different models. The comparison shows that G^{3}AFT yields more detailed and aesthetically consistent results across a diverse set of prompts.

provide differentiability, the gradient signals that pass through remain causally consistent and statistically efficient.

By aligning differentiability with causal faithfulness, NSGP provides a principled refinement to the softmax-based relaxation mechanism. Whereas GSGR focuses on mitigating redundancy in gradient coverage by distributing updates across the generative trajectory, NSGP addresses the complementary issue introduced by softmax-based relaxations by suppressing gradient directions that stem from relaxation-induced artifacts rather than the model's true discrete dynamics, thereby ensuring that optimization signals remain both representative and reliable. Taken together, GSGR and NSGP form the two pillars of the proposed G^{3}AFT framework, which enhances efficiency by reducing memory overhead and improves stability by filtering noisy gradients. Beyond these immediate benefits, the unified design also facilitates scalable adaptation to diverse autoregressive tasks, offering a general recipe for fine-tuning models under constrained compute budgets. The complete training pipeline is outlined in Algorithm 1, which illustrates how GSGR and NSGP interact to deliver coherent and faithful optimization for visual autoregressive models.

4 Experiments

In the main paper, we evaluate the proposed G^{3}AFT framework on the Janus-Pro-1B [5] model under multiple reward formulations to examine its adaptability to diverse alignment objectives. The evaluation encompasses quantitative metrics and qualitative analyses, providing a comprehensive assessment of model

Table 1: Comparison of HPSv2 rewards across different baselines.

Model	Anime	Concept-Art	Paintings	Photo	Average
Janus-Pro-1B	26.56	26.46	26.51	25.68	26.30
GRPO	28.05	27.59	27.73	**27.42**	27.70
Ours	**28.52**	**28.37**	**28.59**	26.64	**28.03**

performance. Ablation studies are further conducted to isolate the contribution of individual components and to clarify the design choices underlying the framework. Additional experimental results across different model variants and resolution settings, together with detailed experimental configurations and supplementary visualizations, are provided in the supplementary material.

4.1 Main Results

Human Preference Guided Fine-Tuning. To examine the effectiveness of the proposed framework, we conduct experiments on the Human Preference Score v2 (HPSv2) [39] benchmark. HPSv2 is a large-scale evaluation suite designed to quantify alignment with human aesthetic and stylistic preferences, covering diverse domains such as anime, concept art, paintings, and photorealistic images.

Fine-tuning in our framework is performed using HPDv2 training prompts, and performance is reported as the mean reward across the four test categories. The results in Table 1 indicate that our approach surpasses both Janus-Pro-1B and GRPO [13] baselines, achieving higher scores in most categories and yielding a superior overall average. Importantly, while reinforcement learning methods such as GRPO emphasize the balance between exploration and exploitation, our framework remains more faithful to the alignment objective itself, directly optimizing for human preference rewards rather than proxy exploration signals. This explains why improvements in the Photo category are less pronounced compared to RL-based methods; However, the gains in other domains are significantly stronger, reflecting the reliability of our approach and echoing phenomena observed in prior works [6] that fine-tune models on HPSv2.

In addition to quantitative improvements, qualitative comparisons in Figure 4 demonstrate that G^{3}AFT produces generations with greater detail and enhanced aesthetic consistency across diverse prompts. These results highlight the adaptability of our framework to preference-guided objectives and underscore its advantage in producing visually compelling outputs that align more closely with human judgments.

Aesthetic Quality Guided Fine-Tuning. To evaluate the transferability of our fine-tuning framework across distinct reward formulations, we incorporate an aesthetic reward derived from the LAION aesthetic predictor V2 [35]. The LAION aesthetic predictor is a regression model trained on large-scale human annotations of image aesthetics, designed to assign continuous scores reflecting

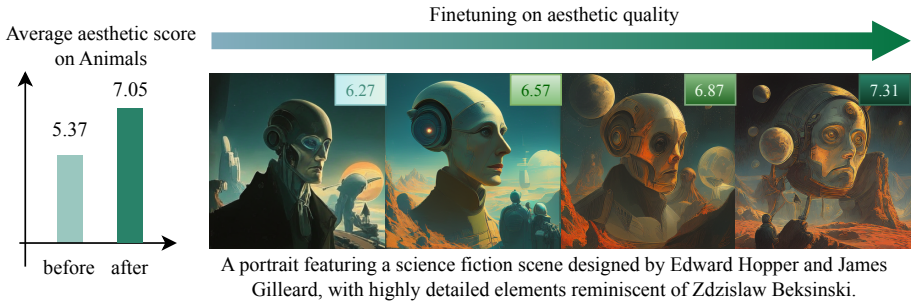

Fig. 5: Aesthetic quality improvements achieved through fine-tuning with aesthetic rewards. The left panel summarizes quantitative improvements on the Animals benchmark, while the right panel presents qualitative examples from the HPDv2 training setting, illustrating enhanced stylistic coherence and visual appeal.

Table 2: Comparison of HPSv2 rewards under different gradient aggregation strategies. GRR denotes the gradient record ratio, i.e., the proportion of tokens whose gradients are retained during optimization.

Strategy	GRR	Anime	Concept-Art	Paintings	Photo	Average
Entropy-Driven	33.3%	28.49	28.54	28.69	26.49	28.05
Strided	50%	28.52	28.37	28.59	26.64	28.03

perceived visual appeal. It has been widely adopted as a standardized proxy for aesthetic evaluation in generative modeling, offering a reliable measure of stylistic quality beyond task-specific metrics.

To assess the generality of our framework across aesthetic-reward training settings of different scales and data distributions, we evaluate G^{3}AFT under two representative fine-tuning scenarios. When fine-tuned on the Animals benchmark introduced in DDPO [4], which consists of 45 category-level prompts covering common animal classes, our method delivers clear quantitative gains, with the average aesthetic score increasing from 5.37 to 7.05. Complementing this small-scale benchmark setting, fine-tuning on the much larger HPDv2 dataset yields notable qualitative improvements, as illustrated in the right panel of Figure 5, where the generated images exhibit enhanced stylistic coherence, smoother compositions, and more balanced color usage. Taken together, these results demonstrate that G^{3}AFT remains robust across distinct aesthetic-reward training conditions, achieving both quantitative improvements in benchmark scores and qualitative gains in visual appeal.

Delving into Alternative Gradient Aggregation Strategies. Within the G^{3}AFT framework, our default policy is Strided Selection, which records gradients at evenly spaced positions. This interval-based approach is simple, efficient, and avoids the need for additional preprocessing, making it well-suited for large-scale experiments. To probe the potential of more information-aware strate-

Table 3: Ablation study on NSGP and discrete sampling approximations. ✓ denotes enabled, × denotes disabled. The column *Approximation* indicates whether discrete sampling is simulated via the Gumbel-softmax or directly by softmax (Pure).

Model	NSGP	Approximation	Anime	Concept-Art	Paintings	Photo	Avg.
Janus-Pro-1B	×	Gumbel	28.20	28.15	27.95	26.59	27.72
	×	Pure	28.39	28.27	28.39	26.32	27.84
	✓	Pure	**28.52**	**28.37**	**28.59**	**26.64**	**28.03**

gies, we further explore Entropy-Driven Selection, which introduces a pre-selection phase in which entropy values are computed for each token during a preliminary pass, and tokens with relatively high entropy are retained for gradient recording. The rationale behind this design is that high-entropy tokens correspond to positions of greater uncertainty and richer information content, and thus aggregating gradients at these locations may better capture global structure. However, under the same gradient record ratio, this method incurs additional computational overhead and higher peak memory usage, since later steps in the computation graph are more costly to store.

While our main and ablation experiments adopt the strided strategy for efficiency, we also investigate whether selecting more information-rich tokens for gradient recording can provide additional benefits. The entropy-driven variant embodies this idea by prioritizing tokens with higher uncertainty, and the results in Table 2 demonstrate its promise. Despite retaining gradients for only 33.3% of the tokens, it achieves performance comparable to strided selection that retains 50% of the tokens, underscoring the potential of entropy-based criteria as a principled alternative that balances efficiency with information density.

4.2 Ablation Studies

Ablation Study on Noise Suppressed Gradient Projection. Table 3 summarizes the ablation results comparing NSGP under different surrogate sampling approximations. The findings show that NSGP consistently improves performance when combined with pure softmax, achieving the strongest results across all evaluation subsets. In contrast, the commonly used Gumbel-Softmax approximation injects random perturbations into all logits, regardless of their relevance to the sampled outcome. This sample-independent noise weakens the alignment between gradient signals and the causal structure of the sampling process, leading to less faithful optimization. Pure softmax avoids such extraneous noise, allowing NSGP to more effectively emphasize informative logits and suppress irrelevant ones.

Taken together, the ablation results show that NSGP combined with pure softmax provides the most faithful surrogate for discrete sampling. This configuration yields cleaner optimization signals by suppressing irrelevant gradient components, maintains causal consistency by aligning updates with the sampled outcome. By avoiding this extraneous noise, our choice of pure softmax enables

Table 4: Ablation experiments on the effectiveness of Global Scope Gradient Recording, Memory indicates peak GPU usage on a single GPU.

Model	GRR	GSGR	Memory	Anime	Concept-Art	Paintings	Photo	Avg.
Janus-Pro-1B	33.3%	×	67.05 GB	27.42	27.44	27.49	26.29	27.16
		✓	53.90 GB	27.88	27.71	28.05	26.07	27.43

NSGP to fully realize its role as a principled refinement, delivering both stronger empirical results and more robust optimization procedure.

Ablation Study on Global Scope Gradient Recording. A common assumption in diffusion-based fine-tuning is that later steps carry richer information, since they implicitly encode the cumulative influence of earlier training signals. Consequently, prior work often selects gradients from the later steps for fine-tuning. However, this practice has two drawbacks in the context of autoregressive image generation. First, the computation graphs of later steps consume substantially more GPU memory, limiting scalability. Second, focusing solely on later steps biases optimization toward local autoregressive dependencies, which undermines the preservation of global information. To address these issues, we introduce Global Scope Gradient Recording (GSGR), which aggregates gradient signals across earlier steps to retain global context while reducing memory overhead.

The ablation results in Table 4 validate the effectiveness of GSGR. When GSGR is disabled, the model instead selects the same proportion of tokens from later steps, leading to higher memory usage (67.05 GB) and weaker performance across evaluation subsets. Enabling GSGR reduces peak memory consumption to 53.90 GB while simultaneously improving average performance from 27.16 to 27.43. This demonstrates that GSGR not only alleviates the computational burden but also strengthens optimization by balancing local autoregressive fidelity with global contextual consistency.

5 Conclusion

In this work, we introduce the first end-to-end fine-tuning framework for discrete token autoregressive image generation within the next-token prediction paradigm, enabling fully differentiable optimization along the entire generative trajectory. Our framework, G^{3}AFT, directly addresses the fundamental challenges posed by non-differentiable sampling and prohibitive memory costs, which have long hindered optimization in autoregressive visual models. Within this paradigm we introduce two complementary techniques. Global Scope Gradient Recording provides representative global optimization signals with reduced memory overhead, and Noise Suppressed Gradient Projection refines surrogate gradients to ensure causal consistency. The two modules are designed to be mutually reinforcing and together forming a coherent optimization pipeline.

Beyond efficiency and fidelity, G^{3}AFT establishes a generalizable methodology for discrete-token visual autoregressive models, offering a scalable recipe for adapting large generative models under constrained compute. Extensive experiments validate the effectiveness of our approach, demonstrating consistent gains in both quantitative metrics and qualitative coherence. We believe this framework opens new directions for scalable differentiable optimization not only in visual autoregression but also in broader multimodal generative tasks.

Acknowledgements

This work is partly supported by the National Natural Science Foundation of China (82441024), the Beijing Natural Science Foundation (L251073), the Research Program of State Key Laboratory of Complex and Critical Software Environment, and the Fundamental Research Funds for the Central Universities.

References

1. Arulkumaran, K., Deisenroth, M.P., Brundage, M., Bharath, A.A.: Deep reinforcement learning: A brief survey. IEEE Signal Processing Magazine **34**(6), 26–38 (2017) 2
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., et al.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022) 5
3. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research **55**, 409–442 (2016) 1
4. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023) 12
5. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025) 6, 10
6. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. In: The Twelfth International Conference on Learning Representations 7, 8, 11
7. Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al.: Cogview: Mastering text-to-image generation via transformers. Proceedings of Advances in Neural Information Processing Systems **34**, 19822–19835 (2021) 6
8. Elasri, M., Elharrouss, O., Al-Maadeed, S., Tairi, H.: Image generation: A review. Neural Processing Letters **54**(5), 4609–4646 (2022) 1
9. de Faria, A.C.A.M., Bastos, F.d.C., da Silva, J.V.N.A., Fabris, V.L., Uchoa, V.d.S., Neto, D.G.d.A., Santos, C.F.G.d.: Visual question answering: A survey on techniques and common trends in recent literature. arXiv preprint arXiv:2305.11033 (2023) 5
10. Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: A review. ACM Computing Surveys **56**(3), 1–39 (2023) 5

11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020) 4
12. Guo, B., Zhang, X., Wu, H., Wang, Y., Zhang, Y., Wang, Y.F.: Lar-sr: A local autoregressive model for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1909–1918 (2022) 5
13. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025) 10, 11
14. He, X., Deng, L.: Deep learning for image-to-text generation: A technical overview. IEEE Signal Processing Magazine **34**(6), 109–116 (2017) 4
15. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys **51**(6), 1–36 (2019) 5
16. Khurana, D., Koli, A., Khatter, K., Singh, S.: Natural language processing: state of the art, current trends and challenges. Multimedia tools and applications **82**(3), 3713–3744 (2023) 2
17. Kim, B.S., Kim, J., Lee, D., Jang, B.: Visual question answering: A survey of methods, datasets, evaluation, and challenges. ACM Computing Surveys **57**(10), 1–35 (2025) 5
18. Le, N., Rathour, V.S., Yamazaki, K., Luu, K., Savvides, M.: Deep reinforcement learning in computer vision: a comprehensive survey. Artificial Intelligence Review **55**(4), 2733–2819 (2022) 2
19. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11523–11532 (2022) 6
20. Li, W., Zhang, P., Zhang, L., Huang, Q., He, X., Lyu, S., Gao, J.: Object-driven text-to-image synthesis via adversarial training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12174–12182 (2019) 4
21. Li, Z., Cheng, T., Chen, S., Sun, P., Shen, H., Ran, L., Chen, X., Liu, W., Wang, X.: Controlar: Controllable image generation with autoregressive models. arXiv preprint arXiv:2410.02705 (2024) 5
22. Ma, J., Zhang, L., Zhang, J.: Sd-gan: Saliency-discriminated gan for remote sensing image superresolution. IEEE Geoscience and Remote Sensing Letters **17**(11), 1973–1977 (2019) 4
23. Manmadhan, S., Kovoor, B.C.: Visual question answering: a state-of-the-art review. Artificial Intelligence Review **53**(8), 5705–5745 (2020) 5
24. Ming, Y., Hu, N., Fan, C., Feng, F., Zhou, J., Yu, H.: Visuals to text: A comprehensive review on automatic image captioning. IEEE/CAA Journal of Automatica Sinica **9**(8), 1339–1365 (2022) 5
25. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 5
26. Otter, D.W., Medina, J.R., Kalita, J.K.: A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems **32**(2), 604–624 (2020) 2
27. Prabhudesai, M., Goyal, A., Pathak, D., Fragkiadaki, K.: Aligning text-to-image diffusion models with reward backpropagation (2023) 7, 8

28. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: A survey. Science China technological sciences **63**(10), 1872–1897 (2020) 2
29. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. IEEE Signal Processing Magazine **34**(6), 96–108 (2017) 4
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 6
31. Ramzan, S., Iqbal, M.M., Kalsum, T.: Text-to-image generation using deep learning. Engineering Proceedings **20**(1), 16 (2022) 4
32. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. Proceedings of Advances in Neural Information Processing Systems **32** (2019) 5
33. Ren, S., Huang, X., Li, X., Xiao, J., Mei, J., Wang, Z., Yuille, A., Zhou, Y.: Medical vision generalist: Unifying medical imaging tasks in context. arXiv preprint arXiv:2406.05565 (2024) 5
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022) 4
35. Schuhmann, C., Beaumont, R.: Laion-aesthetics, 2022. URL https://laion.ai/blog/laion-aesthetics/. Accessed pp. 11–10 (2023) 3, 11
36. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. arXiv preprint arXiv:2406.06525 (2024) 6
37. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Proceedings of Advances in Neural Information Processing Systems **30** (2017) 5
38. Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., Dai, B., Miao, Q.: Deep reinforcement learning: A survey. IEEE Transactions on Neural Networks and Learning Systems **35**(4), 5064–5078 (2022) 2
39. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023) 3, 11
40. Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., Yao, Y., Shen, H., Wan, Z., Huang, J., et al.: Autoregressive models in vision: A survey. Transactions on Machine Learning Research 2
41. Xu, P., Zhu, X., Clifton, D.A.: Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(10), 12113–12132 (2023) 4
42. Yang, Y., Tian, H., Shi, Y., Xie, W., Zhang, Y.F., Dong, Y., Hu, Y., Wang, L., He, R., Shan, C., et al.: A survey of unified multimodal understanding and generation: Advances and challenges. Authorea Preprints (2025) 5
43. Yao, K., Gao, P., Yang, X., Sun, J., Zhang, R., Huang, K.: Outpainting by queries. In: European Conference on Computer Vision. pp. 153–169. Springer (2022) 5
44. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved vqgan. arXiv preprint arXiv:2110.04627 (2021) 5
45. Yu, Q., Weber, M., Deng, X., Shen, X., Cremers, D., Chen, L.C.: An image is worth 32 tokens for reconstruction and generation. Proceedings of Advances in Neural Information Processing Systems **37**, 128940–128966 (2024) 6

46. Yuan, Y., Li, Z., Zhao, B.: A survey of multimodal learning: Methods, applications, and future. ACM Computing Surveys **57**(7), 1–34 (2025) 4
47. Zhan, F., Yu, Y., Wu, R., Zhang, J., Lu, S., Liu, L., Kortylewski, A., Theobalt, C., Xing, E.: Multimodal image synthesis and editing: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **4** (2023) 1
48. Zhang, C., Zhang, C., Zhang, M., Kweon, I.S., Kim, J.: Text-to-image diffusion models in generative ai: A survey. arXiv preprint arXiv:2303.07909 (2023) 4
49. Zhao, S., Zhang, X., Guo, J., Hu, J., Duan, L., Fu, M., Chng, Y.X., Wang, G.H., Chen, Q.G., Xu, Z., et al.: Unified multimodal understanding and generation models: Advances, challenges, and opportunities. arXiv preprint arXiv:2505.02567 (2025) 5
50. Zheng, C., Vuong, T.L., Cai, J., Phung, D.: Movq: Modulating quantized vectors for high-fidelity image generation. Proceedings of Advances in Neural Information Processing Systems **35**, 23412–23425 (2022) 6
51. Zhu, L., Wei, F., Lu, Y., Chen, D.: Scaling the codebook size of vq-gan to 100,000 with a utilization rate of 99%. Proceedings of Advances in Neural Information Processing Systems **37**, 12612–12635 (2024) 6
52. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5802–5810 (2019) 4