Policy Optimized Text-to-Image Pipeline Design

Uri Gadot[1,2] Rinon Gal [2] Yftah Ziser [2] Gal Chechik [2] Shie Mannor [1,2]
[1]Technion [2]NVIDIA Research

Abstract

Text-to-image generation has evolved beyond single monolithic models to complex multi-component pipelines. These combine fine-tuned generators, adapters, upscaling blocks and even editing steps, leading to significant improvements in image quality. However, their effective design requires substantial expertise. Recent approaches have shown promise in automating this process through large language models (LLMs), but they suffer from two critical limitations: extensive computational requirements from generating images with hundreds of predefined pipelines, and poor generalization beyond memorized training examples. We introduce a novel reinforcement learning-based framework that addresses these inefficiencies. Our approach first trains an ensemble of reward models capable of predicting image quality scores directly from prompt-workflow combinations, eliminating the need for costly image generation during training. We then implement a two-phase training strategy: initial workflow vocabulary training followed by GRPO-based optimization that guides the model toward higher-performing regions of the workflow space. Additionally, we incorporate a classifier-free guidance based enhancement technique that extrapolates along the path between the initial and GRPO-tuned models, further improving output quality. We validate our approach through a set of comparisons, showing that it can successfully create new flows with greater diversity and lead to superior image quality compared to existing baselines.

1 Introduction

Recent advancements in generative AI have significantly improved the quality and diversity of text-to-image generation. Early models relied on monolithic architectures, where a single neural network directly translated textual prompts into visual outputs. However, as the field matured, it became clear that combining multiple specialized components—such as fine-tuned diffusion models, super-resolution modules, or specialized embeddings, into more sophisticated workflows leads to superior image quality and greater creative control [5, 40, 63]. This shift from monolithic models to modular workflows has been supported by user-friendly platforms such as ComfyUI[1], a popular open-source tool that allows users to visually construct complex generative pipelines through interconnected nodes represented in JSON format. ComfyUI has rapidly gained popularity due to its intuitive node-based interface, enabling users to assemble diverse generative models (e.g., Stable Diffusion, ControlNet, LoRAs) into flexible workflows tailored to specific image-generation tasks. Despite its accessibility, designing effective workflows remains challenging due to the vast space of possible component combinations and their prompt-dependent effectiveness. Consequently, crafting high-quality workflows typically requires considerable expertise and manual experimentation.

To address this challenge, recent work introduced ComfyGen [16], which uses large language models (LLMs) to automate the construction of prompt-adaptive workflows within ComfyUI. However, a key limitation of ComfyGen was its inability to generate genuinely novel workflow structures. At its core, their approach required synthesizing images using an extensive collection of pre-defined workflows

Preprint. Under review.

[1] https://www.comfy.org/

and prompts, an expensive process limiting their training set's size. Constrained by this small set, their approach essentially learned a classifier over existing flows rather than synthesizing original graph topologies or selecting novel model combinations. This limitation significantly constrains the potential creativity and adaptability of automated workflow generation systems and, as we later show — may also limit their downstream performance.

In parallel, reinforcement learning (RL) has emerged as a powerful paradigm for fine-tuning large language models (LLMs), enabling them to optimize their outputs directly based on reward signals derived from human preferences or other evaluative metrics. Techniques such as Reinforcement Learning from Human Feedback (RLHF) have demonstrated remarkable success in aligning model behaviors with human expectations by iteratively refining model parameters based on explicit reward feedback. Furthermore, recent developments like Group Relative Policy Optimization (GRPO) introduced memory-efficient RL algorithms capable of optimizing policies without separate value functions, making them particularly suitable for complex sequential decision-making tasks. Building on these advancements, we propose FlowRL, a novel extension that integrates reinforcement learning into the workflow prediction framework to overcome its originality limitations. Specifically, we formulate workflow generation as an RL problem where an LLM-based policy sequentially constructs workflow graphs by selecting nodes and connections conditioned on textual prompts. To efficiently guide this process without incurring prohibitive computational costs associated with direct image generation for each candidate workflow during training, we introduce a surrogate reward model trained to predict image quality scores directly from prompts and workflow structures.

Finally, we adopt GRPO combined with per-token reward attribution mechanisms to provide granular feedback during policy updates. This affords our RL agent greater precision in identifying decisions within a generated workflow that contribute positively or negatively toward overall image quality.

In summary, our contributions are as follows

- We introduce ComfyGen-RL, the first RL-based approach for generating genuinely novel ComfyUI workflows tailored to align with human preference feedback.
- We propose a surrogate human-preference reward model enabling efficient RL training without computationally expensive image generations.
- We integrate GRPO with per-token reward attribution for stable and memory-efficient policy optimization.

Through these innovations, FlowRL significantly advances automated workflow generation capabilities, enabling richer creativity and greater adaptability in text-to-image synthesis pipelines.

2 Related Work

Workflow Generation

A recent line of research explores the use of compound systems, where multiple models or modules are chained together, often yielding superior performance compared to isolated models. These multi-component systems have been applied across fields ranging from programming challenges [1] and olympiad-level mathematics [53] to medical diagnostics [38] and video generation [64]. However, building compound systems presents significant challenges. Models must be chosen not only for their individual strengths, but also for their ability to complement each other. Moreover, the parameters of the different components should be selected with the entire system in mind. To address these difficulties, recent work has explored meta-optimization frameworks, where the structure and parameters of entire pipelines are automatically tuned for downstream performance [28]. Others have adopted graph-based architectures allowing dynamic reconfiguration of component interactions [68].

In the realm of text-to-image generation, recent work explores the use of pipelines using agentic systems [67, 61, 23], genetic algorithms [51] or by fine-tuning LLMs using large flow datasets tagged with human preference scores [16]. Although the human preference-based framework has shown promising results, it relies on creating and ranking images using large sets of flows. This, in turn, leads to challenges in effectively scaling the dataset and to a lack of ability to synthesize unseen flows at inference time. Our work aims to address this challenge by leveraging a policy-optimization approach for more effective exploration of the flow parameter space, coupled with a surrogate reward function which avoids the need to generate and rank a large set of images.

Fine-Tuning LLMs with RL: Reinforcement learning (RL) has become increasingly central to the development of large language models (LLMs), playing a key role in aligning model outputs with user preferences and enhancing task-specific capabilities. A prominent example is Reinforcement Learning from Human Feedback (RLHF) [39], which fine-tunes models using reward signals derived from human preferences to better align with communicative goals and social norms [8, 24]. Beyond alignment, RL has shown promise in improving LLMs' performance on domains requiring precise reasoning, such as mathematics [54, 56, 34] and code generation [30, 33]. Recently, [48] proposed Group Relative Policy Optimization (GRPO) as a scalable alternative to Proximal Policy Optimization (PPO). GRPO removes the need for a critic model by optimizing contrastive objective based on intra-group ranking, yielding better sample efficiency, improved stability, and reduced computational complexity [36, 46]. GRPO-trained LLMs demonstrated state-of-the-art performance in mathematical problem solving and code generation, highlighting its effectiveness on tasks requiring structured reasoning and adherence to correctness [48].

Improving Text-to-Image Generation Quality The rapid adoption of text-to-image models [45, 37, 44, 13, 41] has led to many research efforts focused on improving their image quality and better matching human preferences. Some works focus on inference-time modifications, either optimizing noise seeds towards better behaving regions of the diffusion space [14, 43] or applying self-guidance and frequency-based modulations [21, 49, 35] to the generated features.

More commonly, models are tuned to provide better quality outputs. This is often done through carefully selected high-quality datasets or better captioning methods [9, 3, 47]. Another approach uses reward models [29, 59, 60, 31] to guide the generation process. These reward models can be used with reinforcement learning [4, 11, 15, 66], or through direct optimization [6, 42, 55].

Finally, recent methods explore the use of LLMs to improve text-to-image generation [62], commonly by using them to construct workflows featuring multiple models or chained editing tools [67, 51, 16]. Our work similarly uses LLMs to construct workflows, but better aligns them to human preferences through the use of reward models coupled with a reinforcement-learning feedback mechanism.

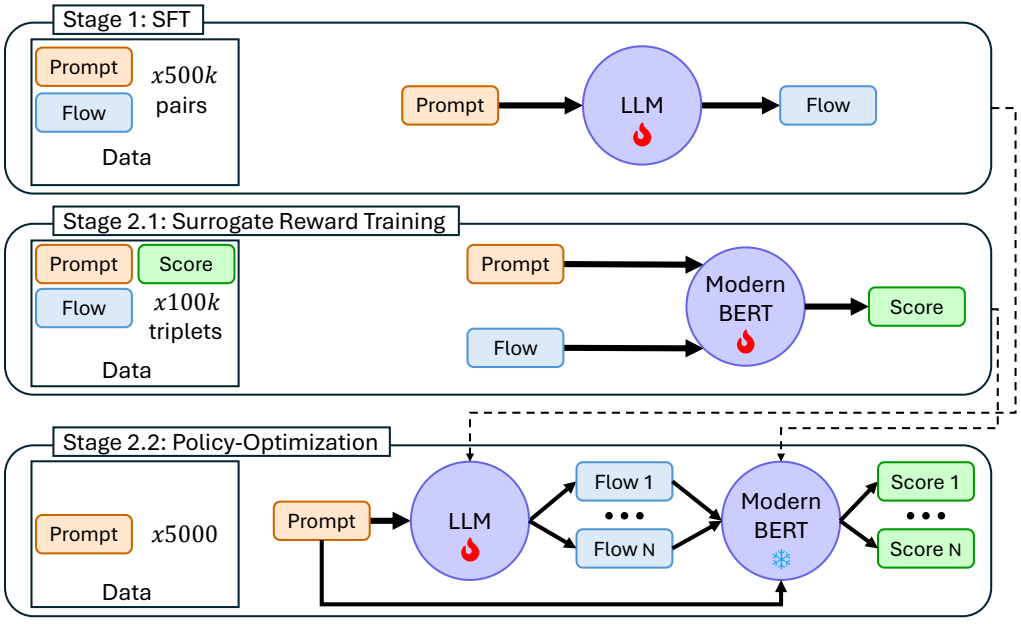

Figure 1: Pipeline overview. Step 1: Finetune LLM for general flow generation (SFT, 500K prompt-flow pairs). Step 2.1: Train reward model (100K prompt-flow-score triplets). Step 2.2: Optimize for quality using GRPO. 🔥 = learning, ❄ = frozen.

3 Methodology

Our goal is to enable efficient training of a human-preference based, prompt-to-workflow prediction system. Ideally, this system should be able to innovate and produce novel, unseen flows. Prior work struggled with this aspect, primarily due to their reliance on scoring images generated with a large set of fixed flows, whose parameters were sampled uniformly from a predefined set of options. To overcome this hurdle, we propose a two-phase training strategy. In the first, we pre-train on a large set of un-scored flows. This avoids the need to generate and score images, allowing us to use a much larger set to teach the LLM the structure of flows and the available components. Then, we perform a second tuning stage, where we leverage human-preference predictor models jointly with recent reinforcement-learning ideas (GRPO [48]) to drive the model towards better-performing subsets of the flow space. As training progresses, more samples are drawn from these regions, and hence, less computation is wasted on inefficient exploration.

However, generating and scoring images during LLM training is itself a costly process, which requires an order of a minute for every training step. Hence, we draw on ideas from the autonomous driving literature, where costly simulations are often replaced by faster predictors trained to replicate simulation outputs [2, 25, 26]. Here, we apply this idea by learning surrogate reward models that predict the final image score directly from the prompt and workflow pair. Notably, prior work has observed that such surrogates are susceptible to reward-hacking solutions [17, 50, 58]. Motivated by findings that ensembles can mitigate reward hacking [7, 65], we train an ensemble of such models and use their variance as a measure of uncertainty, allowing us to filter out samples that optimize for any individual surrogate reward model. Below we present these core components in greater detail and provide an overview of additional design choices or components that allow us to increase efficiency further or refine our results. An overview of our training pipeline is shown in Figure 1.

3.1 Training Data

To train our model we use the flow and prompt dataset of ComfyGen [16]. This set contains 33 human-created flows that define an overall graph structure, further augmented by randomly sampling novel parameter choices for existing blocks such as different base models, differnet LoRAs, diffusion samplers or even the number of steps and guidance scale. Since we do not need to score images for our first stage, we can apply more extensive augmentations and create $2,000$ variants from each baseline flow structure (compared with ComfyGen's 100). The set also contains 10000 prompts taken from the generation sharing website CivitAI.com. We keep the 500 prompts used to test ComfyGen as a holdout, and train using the rest.

3.2 Stage 1: Supervised Fine-Tuning on Flow Dataset

The first stage involves supervised fine-tuning (SFT) an LLM on a dataset of prompt-flow pairs without explicit score labels. At this stage, our goal is to teach the LLM the appropriate vocabulary and flow structure while maintaining output diversity. Our flow dataset D_{SFT} consists of pairs (p_i, f_i) where p_i represents a randomly sampled prompt and f_i represents a randomly sampled flow. We tune the model to take the sampled prompt p_i and return its matching flow f_i. The full LLM query is shown in the supplementary. After fine-tuning, we evaluate the model's perplexity on D_{SFT}, achieving a score of 1.9, which reflects strong alignment with the encoded workflows structural patterns.

Efficient Flow Representation Scheme While prior work [16] directly predicts ComfyUI JSON representations, we note that these JSONs typically contain thousands of tokens, leading to long generation times and increasing memory requirements. An inspection of the tokenized JSONs shows that many tokens are wasted on maintaining the JSON format (e.g., on brackets or quotation marks) or on breaking down model or component names. Hence, to improve training efficiency and reduce token usage, we propose to modify the encoding scheme, using a novel structured representation that captures essential components while reducing token count. Additionally, we introduce specialized tokens to represent key elements of the flow. (e.g. tokens for ComfyUI node names or for model choices). An example of the difference between the two tokenization methods is outlined in Figure 2.

This new encoding scheme yields significant practical advantages resulting in substantial improvements in both computational efficiency and memory utilization. Quantitatively, the 86.7% reduction

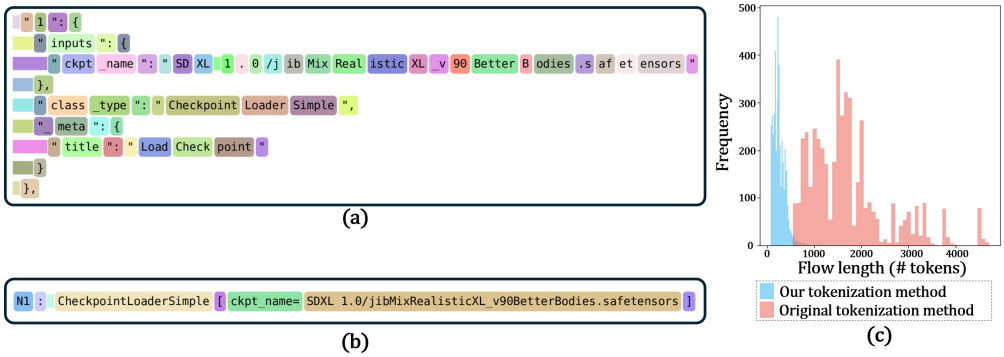

Figure 2: An example of a single ComfyUI node tokenized. **(a)** displays the original JSON input as tokenized by the standard Llama tokenizer. **(b)** shows our custom encoding, with introducing additional tokens to explicitly represent relevant components within workflow. Colored segment corresponds to a different token. **(c)** histogram of flows length (in token) of all training-set

in the average token length ($1500 \rightarrow 200$ tokens per workflow) enabled a $16\times$ batch size increase ($2 \rightarrow 32$ samples/batch) during the first-stage training. Ultimately reaching a $3\times$ time improvement over the original tokenization. These enhancements make it feasible to train complex models and apply memory-intensive algorithms, such as GRPO.

3.3 Stage 2: Reward-based policy-optimization

In the second stage, our goal is to tune the workflow-prediction LLM to better align it with flows that produce high quality outputs for a given prompt. To do so, we propose to leverage the recently introduced Group Relative Policy Optimization (GRPO) approach, which estimates advantages by comparing responses within groups of similar prompts, rather than relying on a separate value function. Using GRPO has two main benefits: (1) it eliminates the need to learn a separate value function, enabling better memory utilization during training and (2) its group-based reward normalization encourages greater exploration and diversity in generated workflows. However, the use of this approach requires us to score and rank the different candidate flows generated for each input prompt at training time. Naively, we could simply generate images with each such flow and score them using the human-preference predictors used by ComfyGen [16]. However, for complex flows, creating the images might take an order of a minute, greatly limiting the speed of training. Hence, we propose to avoid this lengthy generation step and instead train a surrogate reward model that will directly estimate the final reward from a pair of prompt and flow inputs.

Surrogate Reward Model Training We implement the surrogate reward model R_ϕ on top of a ModernBert [57] backbone, with a novel output head trained to map the CLS token into a score. To tune the model, we feed it with strings containing a prompt and flow pair, and task it to predict the human-preference score for the image produced by this pair. For data, we use the ComfyGen dataset D_R, which contains triplets of prompt p_i, flow f_i and score s_i. The surrogate's loss is then:

$$L_R(\phi) = \sum_{(p_i,f_i,s_i)\in D_R} MSE(R_\phi(p_i, f_i), s_i). \tag{1}$$

Although the construction of the original ComfyGen dataset still required generating images and scoring them, we find that the surrogate reward is much more sample efficient, performing well with just the 330 post-augmentation flows of ComfyGen (compared with our own $80k$ unscored flows).

3.3.1 Component-Aware Hybrid Reward Formulation

Since downstream flow performance can be heavily influenced by relatively few tokens (model choices, existence of specific blocks), we propose to further refine our surrogate model with a prefix-prediction score that is better able to assign credit to specific components. Specifically, we tune an additional reward model R_ϕ^{prefix} to predict the generated image score even when presented only with

randomly sampled prefixes of the flow:

$$L_{R^{pre}}(\phi) = \sum_{(p_i, f_i[1:j], s_i) \in D_R} MSE(R_\phi^{pre}(p_i, f_i[1:j]), s_i). \tag{2}$$

Our final reward design combines these two complementary signals to assign a different reward to each token t, depending on both the expected performance of the full flow, as well as a prefix ending with its component:

$$R(t) = R_\phi(p, f) + \sum_{j=1}^{J} \mathbb{1}_{t \in T_j} \cdot R_\phi^{pre}(p, f_{1:j}), \tag{3}$$

where T are the tokens comprising the same flow component as t, and we sum over the contribution of the entire component.

3.3.2 Uncertainty-Aware Reinforcement Learning

Finally, prior work [17, 50, 58] observed that the use of surrogate reward models can lead to reward hacking. To avoid this pitfall, we train an ensemble of N surrogate models $\{R_{\phi_1}, R_{\phi_2}, ..., R_{\phi_N}\}$, each using a different split of our training data. The ensemble provides us with both a more robust mean prediction, as well as with an uncertainty estimate:

$$\mu(p, f) = \frac{1}{N} \sum_{i=1}^{N} R_{\phi_i}(p, f); \quad \sigma(p, f) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (R_{\phi_i}(p, f) - \mu(p, f))^2}. \tag{4}$$

We can then define an uncertainty-aware reward function:

$$R(p, f) = \begin{cases} \mu(p, f) & \sigma(p, f) \leq \tau \\ 0 & \sigma(p, f) > 0 \end{cases}$$

where τ is a threshold parameter. This pessimistic approach assigns zero reward to prompt-flow pairs with high uncertainty, preventing the model from optimizing specific subsets of the reward ensemble, or from drifting to regions where the surrogate's predictions are unreliable.

3.4 Dual model guidance

As an additional step, we propose that results may be further improved through the use of a novel inference mechanism inspired by classifier-free guidance (CFG, [20]). Specifically, we draw on recent work on image generation [27] which demonstrate that diffusion models can be guided by extrapolating the predicted scores along the direction from an under-trained version of the model, and the fully trained one. We propose to apply a similar idea here, where we consider both our policy-optimized model ($\mathcal{M}_{GRPO}$, stage 2) and its "undertrained" SFT version ($\mathcal{M}_{SFT}$, stage 1). At inference time, generations are sampled by interpolating the logits of of both models:

$$\log p_{CFG}(f_j | f_{<j}, p) = \log p_{SFT}(f_j | f_{<j}, p) + \gamma\big(\log p_{GRPO}(f_j | f_{<j}, p) - \log p_{SFT}(f_j | f_{<j}, p)\big) \tag{5}$$

where $\gamma \geq 0$ controls the guidance strength. Unless otherwise noted, we use $\gamma = 1.5$.

4 Experiments

4.1 Comparisons

We follow [16] and compare our approach to a set of baselines across two main metrics: (1) The GenEval [18] benchmark which measures prompt-adherence by using object detection and classification modules to evaluate correct object generation, placement, and attribute binding. (2) Human preference, using the CivitAI prompt-set of ComfyGen [16]. For the latter, we evaluate our approach using both an automated preference metric (HPS v2, [59]) as well as a user study.

We compare our approach against the following types of baselines: (1) Fixed, monolithic models including: SDXL, popular fine-tuned versions thereof, and SDXL-DPO, which was directly

fine-tuned with human preference data. (2) Fixed, popular workflows, where we use the same workflow to generate all images regardless of the prompt. (3) Prior pipeline construction approaches, including agentic workflows that select and use off-the-shelf editing tools to correct generated content (GenArtist, [67]) and reward-based fine-tuned LLMs (ComfyGen [16]).

Prompt adherence: As summarized in Table 1, FlowRL demonstrates strong performance on the GenEval benchmark despite not being explicitly trained for prompt adherence. It achieves an overall score of 0.61, matching the best-performing baseline, ComfyGen. Notably, our approach outperforms other methods in the "two objects" (0.85 vs. 0.82) and "binding" (0.38 vs. 0.29) categories, indicating improved capability in handling complex compositional prompts. A representative qualitative example illustrating prompt adherence is provided in Figure 4.

Figure 3: Example of generations with FlowRL

Visual Quality: To automatically evaluate the visual quality of FlowRL's outputs, we follow [55, 43, 16] and use a pair-wise comparison of HPS v2 [59] score between FlowRL and each baseline and report the average win rate. These comparisons use the full CivitAI test set of [16]. The win-rate of each baseline over FlowRL is reported in Table 1. Additionally, we conducted a user study were we show users 35 randomly sampled prompts and the images generated for each, using FlowRL and one of the baselines. Here, we focus on the best performing baseline from each category, as well as ComfyGen [16]. We then ask them to select the image that they prefer, taking both prompt adherence and visual quality into account. We report the aggregated win percentage in figure 5, and add more details in the supplementary. This experiment demonstrates FlowRL's capability to create more performant ComfyUI workflows for the given input prompts. Representative qualitative comparisons highlighting these improvements are provided in Figure 4, where our outputs consistently exhibit better prompt alignment and structural coherence compared to baseline generations.

Model	Single object	Two object	Counting	Colors	Position	Attribute binding	Overall	HPSv2 winrate vs. FlowRL
SDXL	0.98	0.74	0.39	0.85	0.15	0.23	0.55	2% ± 0.6%
JuggernautXL	**1.00**	0.73	0.48	0.89	0.11	0.19	0.57	5%± 1%
DreamShaperXL	0.99	0.78	0.45	0.81	**0.17**	0.24	0.57	3%± 0.6%
DPO-SDXL	**1.00**	0.81	0.44	**0.90**	0.15	0.23	0.59	5%± 1%
Most Popular Flow	0.95	0.38	0.26	0.77	0.06	0.12	0.42	13%± 1%
2nd Most Popular Flow	**1.00**	0.65	**0.56**	0.86	0.13	0.34	0.59	14%± 1%
GenArtist	0.94	0.41	0.40	0.72	**0.24**	0.07	0.47	5% ± 1%
RPG-DiffusionMaster	**1.00**	0.64	0.21	0.89	0.20	0.35	0.55	3%± 0.8%
ComfyGen	0.99	0.82	0.50	**0.90**	0.13	0.29	**0.61**	40% ± 2%
FlowRL (Ours)	**1.00**	**0.85**	0.44	0.86	0.11	**0.38**	**0.61**	-

Table 1: GenEval and HPS v2 comparisons. FlowRL is on-par with ComfyGen on GenEval and outperforms all other baseline approaches in overall score. On human preference metrics, FlowRL significantly outperforms prior methods. CIs are calculated as one standard deviation from the mean.

Novelty of generated flows: A key advantage of our approach lies in its capacity to generate workflows that are not merely copies of those seen during training. To quantify this novelty, we generate 500 flows using the CivitAI test set, and calculate the normalized Levenshtein distance (NLD) [52, 32] between each generated workflow and its nearest training sample. We further normalize these values by the NLD between training samples, giving us a measure of what fraction of the variance in training data we manage to preserve. Additionally, we report how many generated flows exist "as-is" in the training data, and how many unique flows were created in the 500 output set.

The results are reported in Table 2. Our experiments confirm the findings of [16] which report that their approach learned to copy flows from the training data. FlowRL meanwhile achieves significantly higher novelty, demonstrating the ability to generalize to new parameter combinations. These results

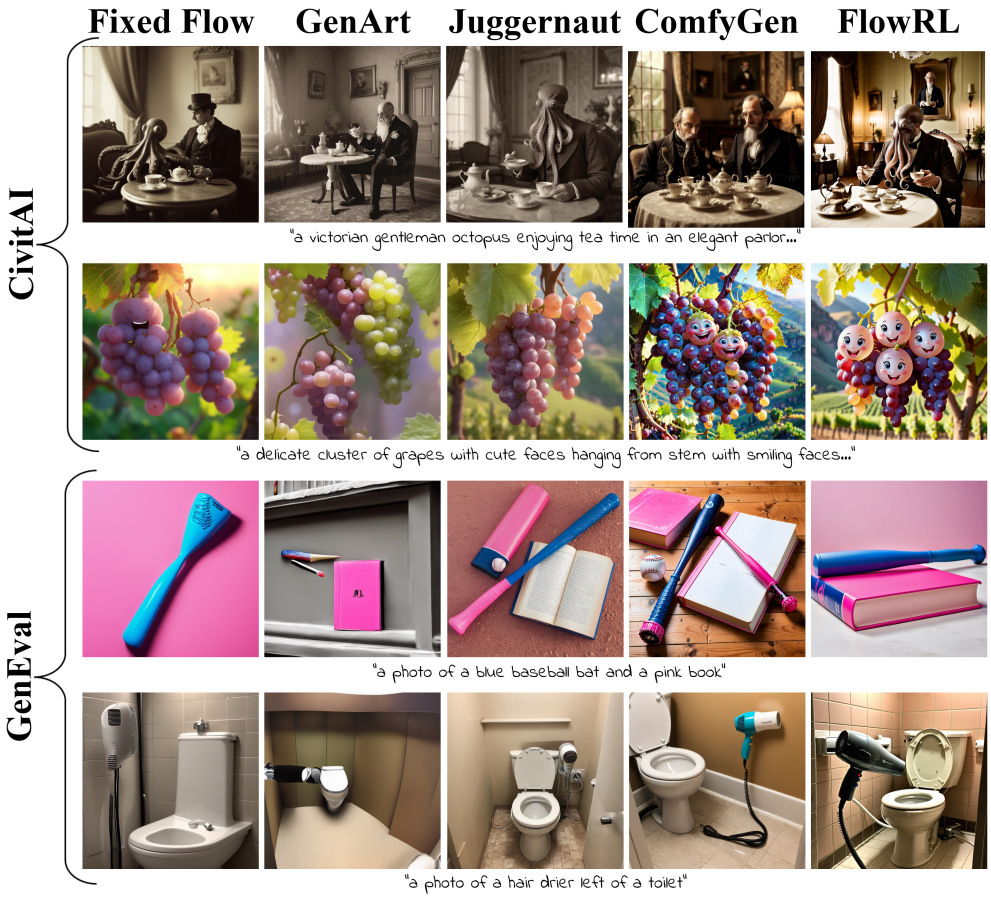

Figure 4: Qualitative results on CivitAI and GenEval prompts.

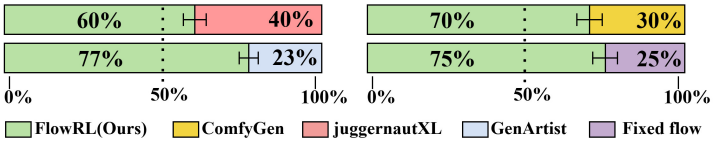

Figure 5: Human study win rate of FlowRL vs other relevant baselines

highlight the effectiveness of our reinforcement learning framework in encouraging the LLM to explore and produce a broader range of complex workflows.

Effects of Dual model guidance: Next, we investigate the impact of our dual-model guidance approach. Specifically, prior work [12] highlighted the ability of guidance-based methods to trade diversity for performance (or recall for precision). We show that similar behavior can be observed here. As shown in Table 2, while increasing guidance strength (γ) improves HPS v2 scores win-rate vs ComfyGen, it significantly impacts the structural diversity of the generated workflows. At $\gamma = 1.5$, our method maintains the uniqueness of generated flows. However, as γ increases, we observe a dramatic reduction in the uniqueness ratio to just 8%.

Notably, all our FlowRL variants maintain near-zero overlap with training data (0-1% "exists in data" vs ComfyGen's 94%), and the NLD ratio actually improves with guidance (from 0.6 without CFG to 0.75 at $\gamma = 2$). This pattern suggests that stronger guidance pushes the model to consistently generate a smaller subset of high-performing workflows, effectively concentrating probability mass on patterns that maximize reward but reducing exploration of the solution space. Conceptually, this mirrors observations in image generation with CFG, where higher guidance strengths produce higher-quality but less diverse output.

Method	unique ratio (%)	exists in data (%)	NLD ratio	HPSv2 win-rate Vs ComfyGen
ComfyGEN	7%	94%	0	-
FlowRL (w/o CFG)	41%	1%	0.6	59%
FlowRL + CFG($\gamma = 1.5$)	41%	0%	0.74	60%
FlowRL + CFG ($\gamma = 2$)	8%	0%	0.75	63%

Table 2: Comparison of originality of flow generation models

4.2 Ablation study

To quantify the impact of individual components in FlowRL, we conducted an ablation study comparing variants with and without our key improvements. We evaluated the following modifications: (1) removing the component-aware reward model, (2) removing the uncertainty ensemble cutoff, (3) varying number of BERT models in our reward ensemble, (4) dropping the SFT step (stage 1), and (5) dropping the GRPO-tuning step. For (5), we instead use the stage-1 model to sample five flows per prompt, and use our reward ensemble to score them in relation to the prompt. Then, we generate an image with the highest scoring flow. Finally, to ensure that our benefits are not grounded in the novel encoding scheme, we also evaluate a baseline ComfyGen [16] model trained on this new representation. We compare all scenarios against both the original ComfyGEN and against our full model, using HPSv2 scores on the CivitAI prompt set. The errors reported are the $1 - \sigma$ Wald interval.

win ratio	w/o prefix reward	w/o reward cutoff	Ensemble of 1 Berts	Ensemble of 3 Berts	Ensemble of 5 Berts	ComfyGen (+ encoded)	SFT only	w/o SFT stage
vs ComfyGen(%)	55 ±2.22	57 ±2.22	55 ±2.22	56 ±2.21	56 ±2.22	37 ±2.16	29 ±2.02	0 -
vs ours (%)	42 ±2.21	45 ±2.21	33 ±2.1	34 ±2.12	36 ±2.15	26 ±1.96	19 ±1.75	0 -

Table 3: The win ratio on the HPSv2 score for each component of our method compared to (1) the ComfyGen baseline and (2) the full ComfyGenRL model, using head-to-head comparisons.

The results are presented in table 3. These demonstrate the vital contribution of each component to overall performance. The full model consistently outperforms all ablations, with particularly significant drops observed when removing the SFT stage entirely (0% win rate against ComfyGen and our full model). This emphasizes the critical nature of proper initialization before applying reinforcement learning methods. Looking at specific components, "prefix reward" proves the most beneficial, showing the importance of assigning more granular rewards. The "ComfyGen (+encoded)" variant, which uses our encoding scheme but lacks reinforcement learning, achieves only a 37% win rate against the original ComfyGen, highlighting that our encoding improvements work synergistically with the GRPO training approach.

5 Discussion

This paper presents a novel approach for fine-tuning LLMs using a combination of supervised learning on flow data, surrogate reward modeling, and uncertainty-aware reinforcement learning. Our method addresses several key challenges in LLM fine-tuning, including reward hacking, distribution shifts, and training efficiency. The results demonstrate that our approach outperforms existing baselines across multiple metrics. Importantly, compared to prior workflow generation work, our approach demonstrates greater output diversity and successfully generalizes to novel flows that did not exist in the training data.

Although it improves on the current state-of-the-art in multiple aspects, our approach still maintains many of their limitations. First, it remains focused on text-to-image workflows, with no support for editing tasks or video modules. Second, introducing new workflow components to the LLM would require retraining our entire stack. In the future, we hope to explore more efficient ways of adapting to novel models or blocks.

By enabling reliable and diverse automated workflow generation, our work advances generative AI systems that adapt to human preferences. We hope it will help foster more collaborative innovation by streamlining the integration of independently trained, specialized modules.

References

[1] Google DeepMind AlphaCode Team. Alphacode 2 technical report. `https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2_Tech_Report.pdf`, 2024.

[2] Halil Beglerovic, Michael Stolz, and Martin Horn. Testing of autonomous vehicles using surrogate models and stochastic optimization. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6. IEEE, 2017.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Yujie Cao, Azhan Abdul Aziz, Wan Nur Rukiah Mohd Arshard, Chang Xu, Muhamad Abdul Aziz Bin Ab Gani, and Issarezal Bin Ismail. Ai 2d-3d generation in architectural design: A divine hand or a pandora's box. In *2024 IEEE 22nd Student Conference on Research and Development (SCOReD)*, pages 203–208. IEEE, 2024.

[6] Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

[8] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

[9] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[10] Daniel, Michael, and the Unsloth Community. Unsloth: Fast, memory-efficient llm fine-tuning library. `https://github.com/unslothai/unsloth`, 2024. Version 2.0. Accessed: 2025-05-20.

[11] Fei Deng, Qifei Wang, Wei Wei, Tingbo Hou, and Matthias Grundmann. Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7423–7433, 2024.

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*.

[14] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *arXiv preprint arxiv:2406.04312*, 2024.

[15] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[16] Rinon Gal, Adi Haviv, Yuval Alaluf, Amit H Bermano, Daniel Cohen-Or, and Gal Chechik. Comfygen: Prompt-adaptive workflows for text-to-image generation. *arXiv preprint arXiv:2410.01731*, 2024.

[17] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[18] Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[21] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7462–7471, 2023.

[22] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

[23] Oucheng Huang, Yuhang Ma, Zeng Zhao, Mingrui Wu, Jiayi Ji, Rongsheng Zhang, Zhipeng Hu, Xiaoshuai Sun, and Rongrong Ji. Comfygpt: A self-optimizing multi-agent system for comprehensive comfyui workflow generation, 2025.

[24] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[25] Keyur Joshi, Chiao Hsieh, Sayan Mitra, and Sasa Misailovic. Gas: Generating fast and accurate surrogate models for autonomous vehicle systems. *arXiv preprint arXiv:2208.02232*, 2022.

[26] Maria Kalweit, Gabriel Kalweit, Moritz Werling, and Joschka Boedecker. Deep surrogate q-learning for autonomous driving. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1578–1584. IEEE, 2022.

[27] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

[28] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[29] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[30] Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

[31] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[32] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[33] Zeyuan Li, Yangfan He, Lewei He, Jianhui Wang, Tianyu Shi, Bin Lei, Yuchen Li, and Qiuwu Chen. Falcon: Feedback-driven adaptive long/short-term memory reinforced coding optimization system. *arXiv preprint arXiv:2410.21349*, 2024.

[34] Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*, 2023.

[35] Yang Luo, Yiheng Zhang, Zhaofan Qiu, Ting Yao, Zhineng Chen, Yu-Gang Jiang, and Tao Mei. Freeenhance: Tuning-free image enhancement via content-consistent noising-and-denoising process. *arXiv preprint arXiv:2409.07451*, 2024.

[36] Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

[37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[38] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[39] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[40] Iván J Pérez-Colado, Manuel Freire-Morán, Antonio Calvo-Morata, Víctor M Pérez-Colado, and Baltasar Fernández-Manjón. Ai asyet another tool in undergraduate student projects: Preliminary results. In *2024 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–7. IEEE, 2024.

[41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024.

[42] Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. Aligning text-to-image diffusion models with reward backpropagation. *arXiv preprint arXiv:2310.03739*, 2023.

[43] Zipeng Qi, Lichen Bai, Haoyi Xiong, et al. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[46] S Sane. Hybrid group relative policy optimization: A multi-sample approach to enhancing policy optimization. arxiv, 2025.

[47] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*, 2023.

[48] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[49] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4733–4743, 2024.

[50] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*, 2023.

[51] Dominik Sobania, Martin Briesch, and Franz Rothlauf. Comfygi: Automatic improvement of image generation workflows, 2024.

[52] Keiichiro Tashima, Hirohisa Aman, Sousuke Amasaki, Tomoyuki Yokogawa, and Minoru Kawahara. Fault-prone java method analysis focusing on pair of local variables with confusing names. In *2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 154–158. IEEE, 2018.

[53] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 2024.

[54] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

[55] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[56] Peiyi Wang, Lei Li, Zhihong Shao, RX Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *arXiv preprint arXiv:2312.08935*, 2023.

[57] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

[58] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R Bowman, He He, and Shi Feng. Language models learn to mislead humans via rlhf. *arXiv preprint arXiv:2409.12822*, 2024.

[59] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[60] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[61] Xiangyuan Xue, Zeyu Lu, Di Huang, Zidong Wang, Wanli Ouyang, and Lei Bai. Comfybench: Benchmarking llm-based agents in comfyui for autonomously designing collaborative ai systems, 2024.

[62] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *International Conference on Machine Learning*, 2024.

[63] Honglin Yao. Improving performance of face generation for specific individuals based on lora. In *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)*, pages 77–80. IEEE, 2024.

[64] Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024.

[65] Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles. *arXiv preprint arXiv:2401.00243*, 2023.

[66] Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. Large-scale reinforcement learning for diffusion models. *arXiv preprint arXiv:2401.12244*, 2024.

[67] Wang Zhenyu, Li Aoxue, Li Zhenguo, and Liu Xihui. Genartist: Multimodal llm as an agent for unified image generation and editing. *arXiv preprint arXiv:2407.05600*, 2024.

[68] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024.

A Appendix

A.1 Broader impact statement

Our work offers a new path to improve text-to-image generation, but this improvement is not without possible social impacts. Text-to-image models can be used to create harmful or misleading content, and improving their output can increase this risk.

Moreover, our work relies on the abundance of community-created, fine-tuned specialized models and adapters. These are rarely developed with safety in mind, and do not typically undergo red team assessments. Hence, they may increase the risk of the user generating biased or unsafe content. However, this can be mitigated by carefully curating the generative models seen during training, or by black-listing specific models in the output flow strings.

Future work may be able to further refine the reward model used at training to also align it with safety, for example by reducing the score for content deemed unsafe by a detector.

A.2 ComfyUI Overview

ComfyUI is a popular ($77,300$ stars on GitHub at the time of this writing), open-source workflow engine designed for flexible and extensible automation of generative AI tasks. Its node-based interface allows users to visually construct and execute complex processing pipelines, with users across the community often implementing and sharing new nodes to accommodate the changing landscape of generative tools. The pipelines constructed in ComfyUI can be exported to a JSON format, which we then map to a more compact representation and use for both our training data and our LLM output representation. To run our generated workflows, we convert them back to the JSON format and run them through the ComfyUI API. A dedicated user could also load these workflows through the UI and further manually refine them.

A.3 Additional Qualitative results

Here, we give more qualitative examples of FlowRL generations.

Figures 6, 7 and 8 provide additional generations of CivitAI prompts using FlowRL. We give a detailed list of the relevant prompt (ordered by appearance order, from top-left to bottom right)

We provide additional qualitative comparisons between FlowRL and the baselines in Figure 9 for both CivitAI prompts and GenEval prompts.

In addition in figure 10 we give a qualitative comparison between FlowRL with and w/o the usage of the dual model guidance mechanism (CFG).

List of prompts for example generations

1. "Amazing detailed photography of a cute adorable samurai kitten holding Katana with 2 paws, Cherry Blossom Tree petals floating in air, high resolution, piercing eyes, lifelike fur, Anti-Aliasing, FXAA, De-Noise, Post-Production, SFX, insanely detailed & intricate, hypermaximalist, elegant, ornate, hyper realistic, super detailed, noir coloration, serene, 16k resolution, full body"
2. "masterpiece, best quality, high quality, intricate, absurdres, very aesthetic, no humans, landscape, outdoors, mountain tops, wind, windy, wind lines, clouds, above clouds, cliff, wind magic, aurora, ultra wide angle shot, cinematic style, highly detailed, extremely detailed, sharp detail, majestic, shallow depth of field, movie still, soft light, circular polarizer, colorful, wallpaper, professional illustration, anime"
3. "pixar style of turtle, as a pixar character, tinny cute, luminous, wearing hawaiian hat, at the sea shore, tropical beach, smile, high detailed, photorealistic, 8k"
4. "Medieval German castle, surrounded by mountains, high fantasy, epic, digital art."
5. "style of Edvard Munch, Piercing, sagacious eyes, mirage-like, the Sandswept dreamdweller, a trickster of dunes, clad in a wind-whispered turban, eternally smirking, sandswaggling over a dune-freckled miragepath in an ancient zephyr-twisted cactidle wilderness of towering

dustfrond phantasmagorias, paying no heed to the sun-scorched skyripples above, Arid, Sand-whirled, Mirage, Cacti, Mystical Desert, oasis illusions. Edvard Munch style"

6. "full body, Fat cats at Elrond's council from the movie Lord of the Rings, fluffy paws, background action-packed"
7. "detailed, vector art, thick lines, oil painting, vibrant, colorful, candy pink, scarlet red, orange, smooth coloring, nature, landscape, stone pillars, long wild trees, moody streaks sky, natural lighting, river, reflections, best composition, background"
8. a woman with red hair and a white shirt is shown in this painting style photo with a pink background, Charlie Bowater, stanley artgerm lau, a painting, fantasy art masterpiece, best quality, depth of field, backlighting, intricate details"
9. "cinematic shot of stone giant walking in lush forest, dappled sunlight, high resolution"
10. "Majestic jagged rocky mountains, red mesas, wind eroded colorful rock formations, twighlight, starry night, petrified forest national park, arizona, astrophotography"
11. "Cubist inspiration, A landscape represented with planes and flat colors. The landscape could show a field, forest or city, and flat planes and colors could be used to create a sense of depth and perspective, surrealism, aesthetic, bold gorgeous colours, high definition, super clear resolution, iridescent watercolor ink, acid influence, fantastic view, crisp quality, complex background, medium: old film grain, tetradic colors, golden hour, rust style, vantablack aura, golden ratio, rule of thirds, cinematic lighting Dark realism and magical. Complementary poisonous colors with deep zoom Memphis style abstract bokeh background with deep zoom"
12. "FrostedStyle Highly detailed Dynamic shot of a transparent frosted ruby reindeer, glowing with rage from within extremely detailed"
13. "vertical symmetry, vntblk, movie poster art, blood moon, red moon, darkest night, stonehenge, low angle:famous artwork by caspar david friedrich and stephan martiniere, perfectly round scifi portal, ominous dark surreal and unique landscape with towering obelisks piercing the sky, glowing ornate lovecraftian artifact, jagged rock formations, night sky, mysterious, ethereal, deserted, dark corners, burgundy, anthrazit grey, crimson, sunset orange, yellow, teal:16, ultra detailed"
14. "by Peter Holme III and Roger Dean and Vitaly Golovatyuk and Mark Lovett, cinematic, shallow depth of field"
15. "grainy, extremely detailed, intricate detail, dynamic lighting, photorealistic, filmg, natural lighting, low light, cat, slime, red glowing eyes, :P, fluffy, hairy, fluff, glowing stripes, raining, wet, dark theme, open mouth, lot of teeth, abyss, lurking in shadow"
16. "The art of Origami, Paper folding, Swan on a lake, Amazing colours, Intricate details, Painstaking Attention to Details, UHD"
17. "amateur analog photo, The creature monster brown fur Easter bunny character covered in yeast, evil, creepy, in dark forest, fine textures, high quality textures of materials, volumetric textures, natural textures"
18. "In a wondrously gleaming futuristic realm composed entirely of ripe peaches, a towering palace made of glistening peach flesh and pitted stone stands as the focal point of the image. The palace's walls are adorned with intricate carvings of peach vines and blossoms, while peach juice flows like streams through the city streets. This vivid and surreal painting captures the ethereal beauty of a world where nature and architecture are seamlessly intertwined, every detail rendered with unparalleled precision and depth, making viewers feel as if they could reach out and touch the succulent fruit structures."
19. "high-contrast palette, cinematic quality, fashion photography, chimp wearing a black suit with a black shirt with a black vest with a black necktie with black Rayban style sunglasses, natural skin texture, realistic skin texture, skin pores, skin oils"
20. "faistyle, retro artstyle, painting medium, lake, mountain, forest"
21. "close up Portrait photo of muscular bearded guy in a worn mech suit, light bokeh, intricate, steel metal rust, elegant, sharp focus, photo by greg rutkowski, soft lighting, vibrant colors, masterpiece, streets, detailed face"

22. "detailed ink, pen and ink, mail art, best quality, detailed epic ice transparent ethereal otherworldly ghost castle in the blue sky, clouds, smoke, fog, detailed landscape, ghost figures, lake, boat, green forest, detailed flying dragon at the sky, detailed scales, warm lights, glittering, Craola, Dan Mumford, Andy Kehoe, 2d, flat, art on a cracked paper, patchwork, stained glass, cute, adorable, fairytale, storybook detailed illustration, cinematic, ultra highly detailed, tiny details, beautiful details, mystical, luminism, vibrant colors, complex background"
23. "crystal scorpion"
24. "the image portrays a tranquil scene of a boat floating gently on the water, surrounded by an expansive landscape. the moon, full and glowing with a warm, reddish orange hue, casts a mystical ambiance over the entire scene. its reflection shimmers off the surface of the water, adding to the serene atmosphere. in the distance, mountains loom under the moon's soft glow, their peaks partially obscured by the low hanging clouds. they appear majestic yet gentle, as if watching over the peaceful night below. trees line the shore in the foreground, their silhouettes faintly visible against the darkening sky. this picturesque setting evokes a sense of calm and tranquility, inviting viewers to take a moment and appreciate the beauty of nature. it is a symphony of colors and shapes, each element working harmoniously together to create a visually captivating and emotionally soothing composition."
25. "hyper detailed, elusive, exotic, angelic, luminescent, by James Gilleard and by Alice Pasquini, point-of-view shot, fisheyes view, sunbeams lighting"
26. "A colossal majestic tiger, looms over a cavern's silhouette, gazing intently at a small human figure with a stance of curiosity, the figure a silhouette against a backdrop bathed in the warm oranges and yellows of a sun, faces the tiger unafraid, with floating embers dancing around them both, scene of serene confrontation amidst the enveloping dusk"
27. "illustration, solo, animal skull head, screaming, long split tongue, fangs, head closeup, black leather coat, horror atmosphere, side view"
28. "mysterious silhouette of woman from the enchanted pond, abstract art, by Minjae Lee, Carne Griffiths, Emily Kell, Geoffroy Thoorens, Aaron Horkey, Jordan Grimmer, Greg Rutkowski, extraordinary depth, masterpiece, surreal, geometric patterns, extremely detailed, bokeh, perfect balance, deep and thin edges, artistic photorealism, smoothness, excellent masterpiece by the head of rapid engineering, white background: 1.2"
29. ""Struggling to breathe, like being held under water" beautiful inner light, deep shadows, extraordinary detail"
30. "a fantasy landscape at dawn covered in magical flowers"
31. "vintage, shabby, morning, dawn, cozy world, Kruskamp, Monge, Kincaid, Potter, Dali, Burton, oil, coal, provence, house by the sea, cozy and beautiful landscape, double composition, drama, tragedy, the core of magic"
32. "fine art, oil painting, best quality, dark tales, illustration, each color adds depth, and the entire piece comes together to create a breathtaking spectacle of motion and tranquility., while the ball is adorned with an array of stripes in various hues. the figurine, while her right hand delicately holds a small, epic splash cover art in the van gogh style, starry sky, dan mumford, andy kehoe, 2d, flat, delightful, vintage, art on a cracked paper, patchwork, stained glass, fairytale, storybook detailed illustration, cinematic, ultra highly detailed, tiny details, beautiful details, mystical, luminism, vibrant colors, complex background"
33. "Envision a breathtaking waterfall cascading into a crystal-clear pool surrounded by lush greenery. The pool is home to magical water creatures, including playful water sprites and elegant swans with feathers that shimmer in shades of silver and gold. Mist rises from the waterfall, creating rainbows in the sunlight. Curious frogs with iridescent skin leap from rock to rock, while dragonflies with jeweled wings flit above the water. The sanctuary is a hidden paradise, inviting all who enter to experience the tranquility and magic of this enchanting world"
34. "futuristic building, surface from a alien planet, mountains in the background, sci fi, fantasy, space art, galaxy background, shotting stars, dynamic angle, intricate "
35. "Image is a digital artwork featuring a futuristic samurai robot. The robot has a sleek, metallic body with intricate mechanical details and a predominantly black and silver color

scheme. It wears a large, red, conical hat and a matching red cape that flows behind it. The robot's face is obscured by a mask, giving it a mysterious appearance. It holds a red and black katana in its right hand, ready for combat. The background is a gradient of dark grey, with a circular, smoky effect behind the robot, adding to the dramatic and intense atmosphere of the scene."

36. "masterpiece, ASCII, 8k.absurdes, intricate, maximum resolution, hyper detailed, Mirage, DonMn1ghtm4reXL, glow, fog, obsidian armor with red ruby, details, hellgate london themed, demoniac armor, force huge demoniac wide wings, glowing wings, energy wings"

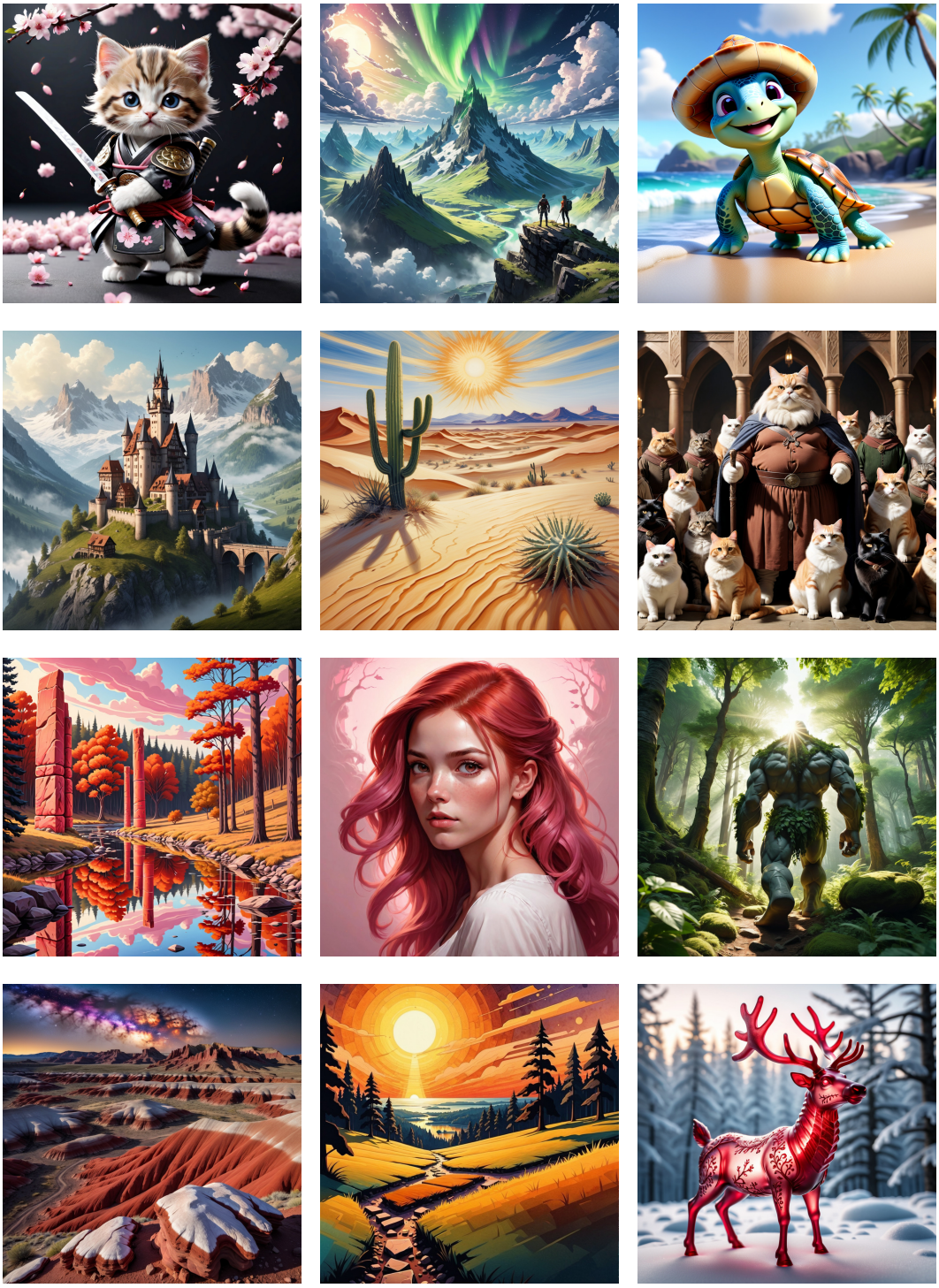

Figure 6: More qualitative generations using FlowRL

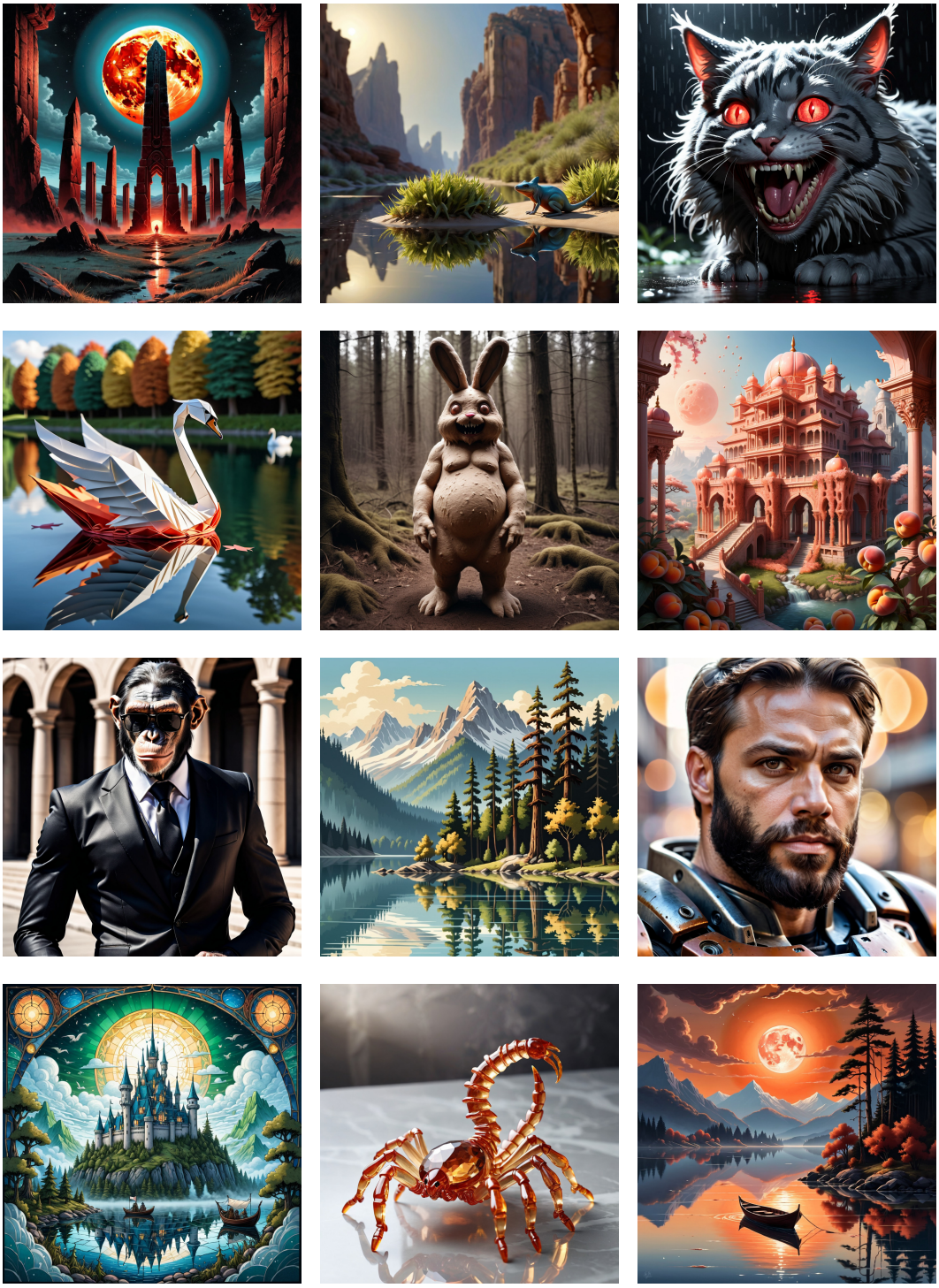

Figure 7: More qualitative generations using FlowRL

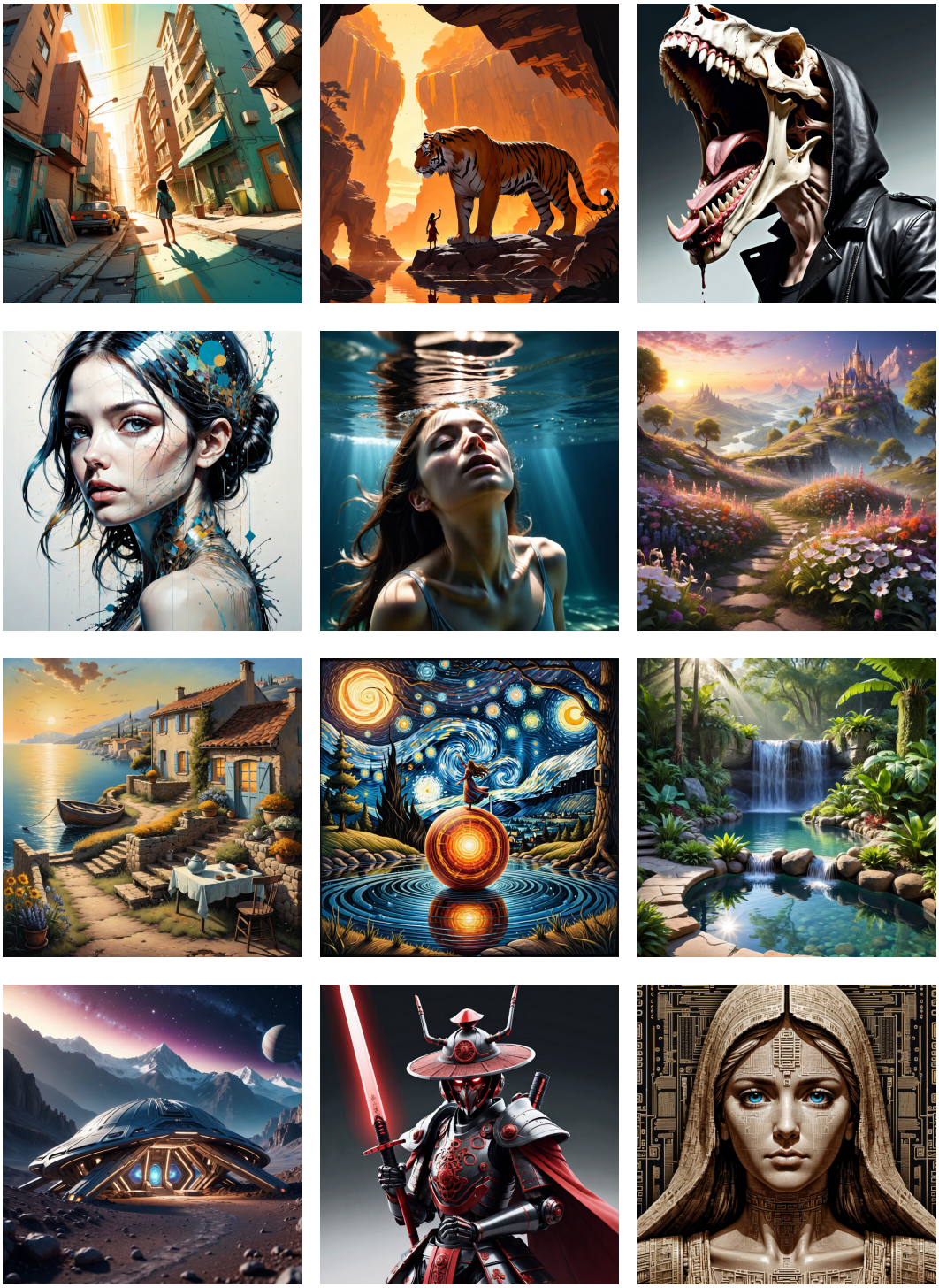

Figure 8: More qualitative generations using FlowRL

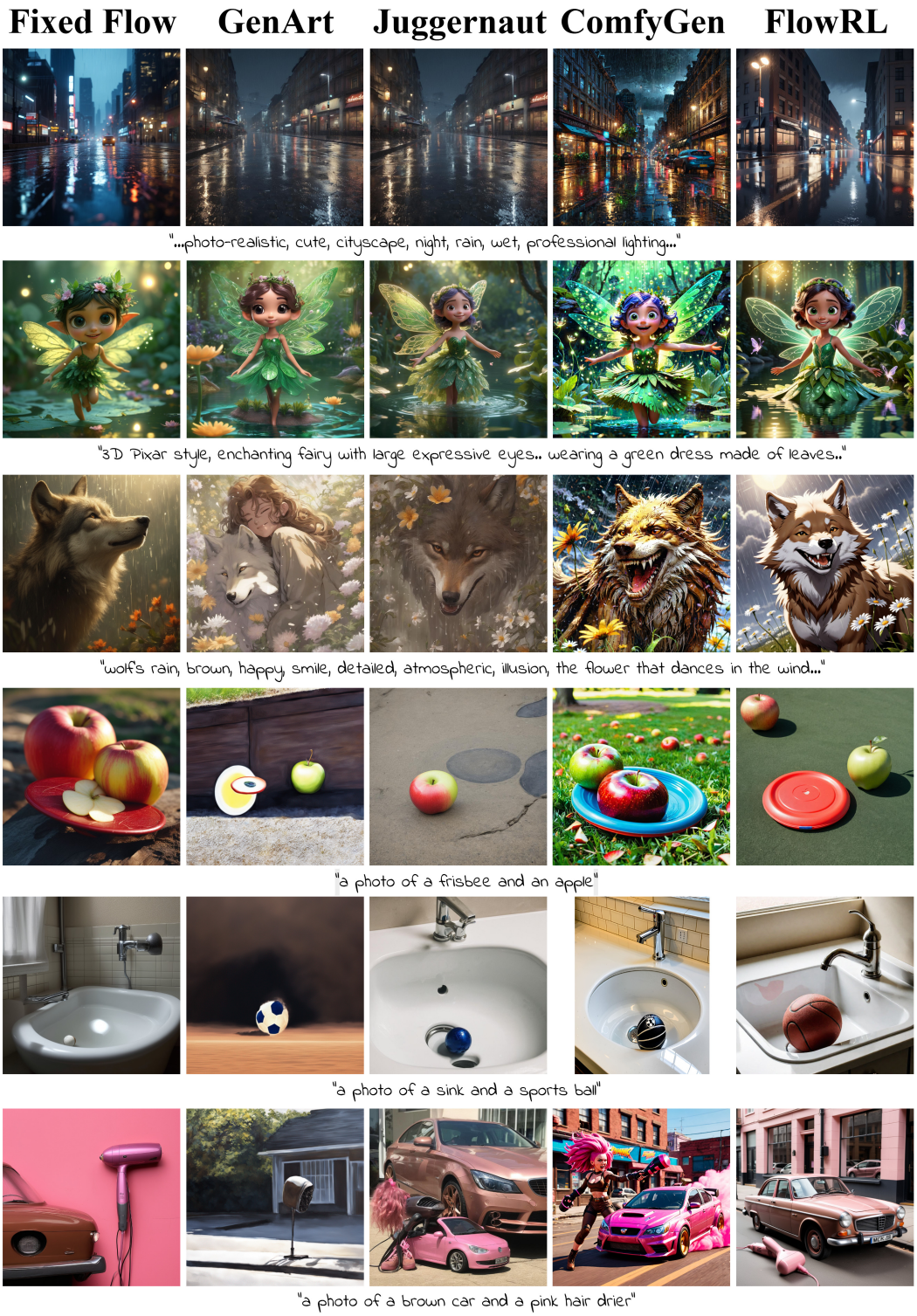

Figure 9: Additional qualitative comparisons on CivitAI prompts (top 3) and GenEval prompts (bottom 3)

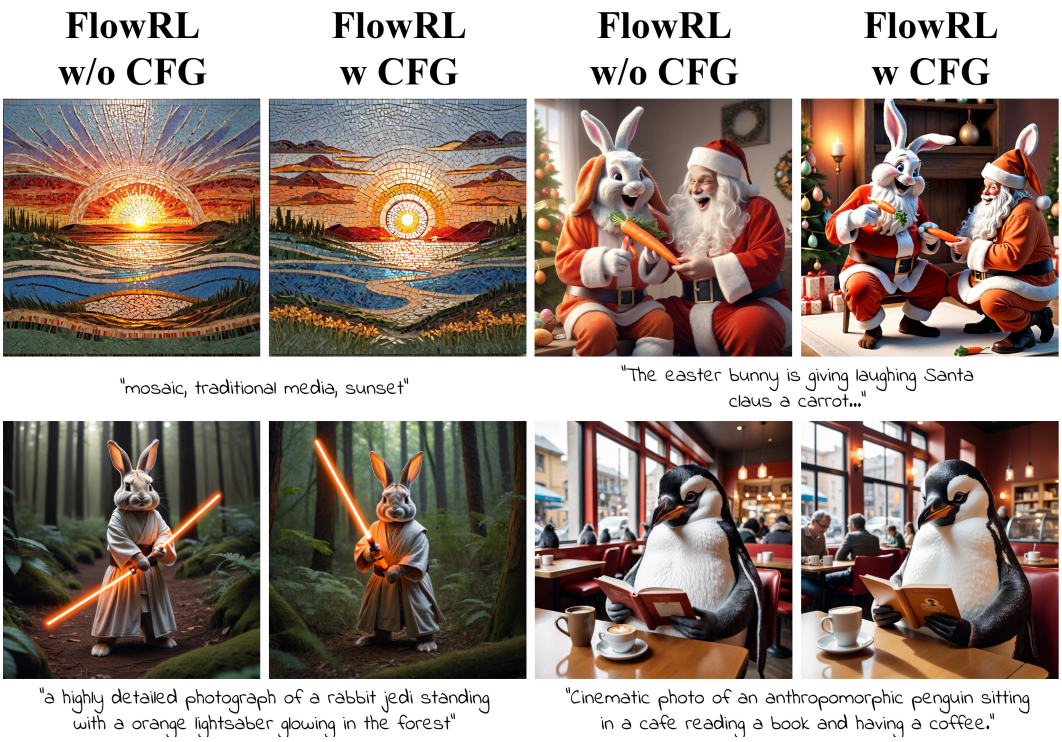

Figure 10: Qualitative example on influence of CFG on the output image

A.4 Tokenization Encoding Method

We developed a systematic procedure to transform JSON-based workflow representations into a compact, encoded format. This process utilizes schema learning to ensure both accuracy and efficiency in data transformation.

Methodology First, we infer a schema from a collection of workflow JSON files by iterating through each file and extracting the class types and field information for all utilized nodes. The resulting schema is stored for future encoding tasks. In the encoded representation, nodes are sorted and formatted to include their class types and input values, with explicit references to connected nodes. Each line in the encoded output corresponds to a node from the original JSON structure, providing a clear and organized mapping.

Incorporating Workflow-Specific Tokens To more effectively capture the structure of ComfyUI workflows, we enhanced the base tokenizer by introducing custom tokens that represent key workflow elements such as node types, connections, and parameters. This enriched tokenization scheme helps the model better understand relationships between workflow components. Below, we provide examples of some of the custom tokens added to the tokenizer:

- ng Everywhere3
- AspectSize
- Automatic CFG
- BNK_AddCLIPSDXLRParams
- BNK_CLIPTextEncodeAdvanced
- BasicPipeToDetailerPipe
- Image Levels Adjustment
- Image Remove Background (rembg)

- CLIP Positive-Negative XL w/Text (WLSH)
- CLIP=
- CLIPLoader
- CLIPMergeSimple
- CLIPSetLastLayer
- CLIPTextEncode
- CLIPTextEncodeSDXL
- CLIPTextEncodeSDXLRefiner
- CLIP_NEGATIVE
- CONDITIONING=
- CR Apply LoRA Stack
- CR Apply Model Merge
- SDXL 1.0/animagineXLV31_v30.safetensors
- SDXL 1.0/crystalClearXL_ccxl.safetensors
- SDXL 1.0/dreamshaperXL_turboDpmppSDEKarras.safetensors
- SDXL 1.0/envyhyperdrivexl_v10.safetensors
- SDXL 1.0/faces_v1.safetensors
- SDXL 1.0/jibMixRealisticXL_v90BetterBodies.safetensors
- SDXL 1.0/juggernautXL_v9Rdphoto2Lightning.safetensors

A.5 User study

To evaluate our method against baselines, we conducted a user study using a structured survey. For the study, we randomly sampled 50 prompts and generated corresponding images with each baseline. From these, we filtered out results which contained unsafe content (e.g., nudity, violence), resulting in 7–11 comparison questions per baseline. These comparisons were aggregated into a survey where participants were shown a prompt and the outputs from FlowRL and one baseline, and asked to select their preferred image.

We collected approximately 200 responses per baseline. Figure 11, provides an example of a question from our survey.

A.6 Implementation details

A.6.1 SFT stage

We implement our model based on a pre-trained Meta Llama3.1- 8B [19]. We used the unsloth [10] library to fine-tune the model using LoRA [22]. The SFT stage was trained on a single NVIDIA H100 80GB HBM3 GPU for 10 hours.

LoRA Configuration: To enable parameter-efficient fine-tuning, we applied LoRA (Low-Rank Adaptation) to the model's attention and feed-forward layers. The LoRA rank was set to $r = 16$, with an alpha value of $\alpha = 16$, and a dropout rate of 0.0. Target modules included "q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj", as well as "lm_head" and "embed_tokens" since we added new tokens to our vocabulary.

Prompt structure During the supervised fine-tuning (SFT) stage, the LLM is provided with both the prompt and one of the encoded flows:

```
">>> Prompt:
    {p_i}
>>> Flow:
    {f_$}"
```

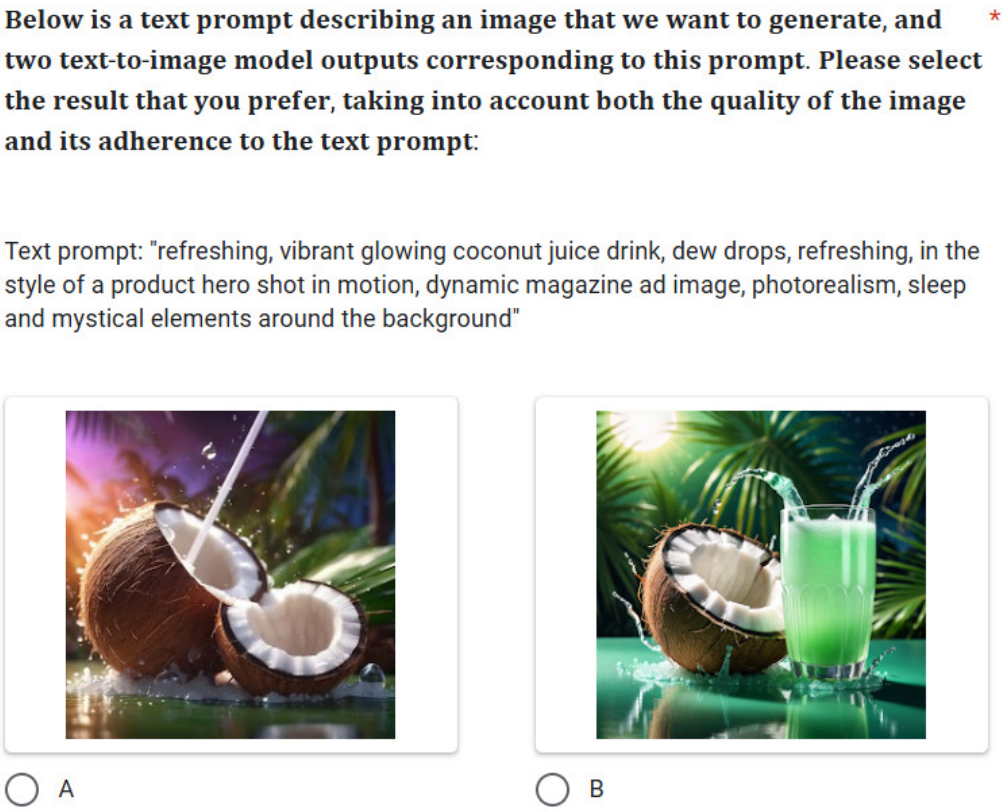

Figure 11: An example question from the user study.

In contrast, during the reinforcement learning (RL) fine-tuning stage, only the prompt is given to the LLM, and it is tasked with generating one or more candidate flows. This setup encourages the model to learn to produce the most appropriate flow for each prompt:

```
">>> Prompt:
    {p_i}
>>> Flow:"
```

A.6.2 Reward model training

For training the Reward BERT model, we utilized the "answerdotai/ModernBERT-base" [57] as the foundational architecture. Beyond its improved classification performance, we selected ModernBert because it was trained on sequence lengths that match our expected prompt and encoded-flow format. We used the Adam optimizer with the default parameters and a learning rate of $8e-5$. The maximum sequence length was set to 4096 tokens, with a batch size of 128 over 10 epochs. Fine-tuning was done on a single NVIDIA A100-SXM4-80GB for approximately 4 hours.

Dataset: Each data-point consisted of the triplet (f_i, p_i, s_i): flow, prompt and human-preference normalized score. and was inserted to the model in this format:

`"[PROMPT] {p_i} [FLOW] {f_i}".`

The model was tasked with prediction the output score s_i for each pair, using an MSE loss.

A.6.3 GRPO Fine-Tuning Hyperparameters

Below, we detail the key hyperparameters and configurations used in the GRPO (Group Relative Policy Optimization) fine-tuning stage:

LoRA Configuration: To enable parameter-efficient fine-tuning, we applied LoRA (Low-Rank Adaptation) to the model's attention and feed-forward layers. The LoRA rank was set to $r = 16$, with an alpha value of $\alpha = 16$, and a dropout rate of 0.0. Target modules included "q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj", following best practices for large language model adaptation. Note that this step does not optimize the "lm_head" or "embed_tokens" layers as this step aims to further tune the SFT model, which already knows the flow vocabulary.

Optimization Settings: We used the Adam optimizer with a learning rate of $5e-6$, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay of 0.1. Training was performed with a batch size of 16 per device.

GRPO-Specific Parameters: We used the group size of 4 (number of generations per prompt for group-based reward calculation). clipping coefficient of 0.2, max grad norm of 0.5 and KL-regularization coefficient of 0.2. We also used generation temperature of 0.9, and maximal output tokens of 500.

Training Procedure: Fine-tuning was conducted for 2 epochs over the CivitAI promp train set. We trained on a single NVIDIA A100-SXM4-80GB node (8 GPUs) for approximately 10 hours. We used an ensemble of 7 BERT reward models and used their mean as the surrogate reward. we set the uncertainty threshold to 0.08 and set the "uncertain reward value" to 0. For the prefix-reward mechanism, we used 5 different Bert models.