

EXPANDING SCALING BOUNDARY OF COMPOSITIONAL TEXT-TO-IMAGE GENERATION VIA COMPOSITION CURRICULUM

Anonymous authors

Paper under double-blind review

ABSTRACT

Text-to-Image (T2I) generation has long been an open problem, with compositional synthesis remaining particularly challenging. This task requires accurate rendering of complex scenes containing multiple objects that exhibit diverse attributes as well as intricate spatial and semantic relationships, demanding both precise object placement and coherent inter-object interactions. In this paper, we propose a novel compositional curriculum reinforcement learning framework for T2I generation, named CompGen, to address compositional weaknesses in T2I models. Specifically, we leverage scene graphs and introduce a novel difficulty criterion along with a corresponding adaptive Markov Chain Monte Carlo graph sampling algorithm. Using this difficulty-aware approach, we generate training datasets for Group Relative Policy Optimization (GRPO) comprising prompts and question-answer pairs with varying complexity levels. We demonstrate that different training schedulers yield distinct scaling curves for GRPO, with data distributions following easy-to-hard progression or Gaussian sampling strategies producing superior scaling performance than random. Our extensive experiments demonstrate that CompGen significantly strengthens compositional generation capabilities for both diffusion and auto-regressive T2I models, which highlights its effectiveness in enhancing the compositional understanding of T2I generation systems.

1 INTRODUCTION

Text-to-Image (T2I) generation has achieved remarkable progress in synthesizing visually compelling content from textual descriptions (Ramesh et al., 2022; Saharia et al., 2022; Podell et al., 2023; Balaji et al., 2022). Despite these advances, current T2I models face significant limitations in compositional synthesis, particularly in accurately rendering complex scenes containing multiple objects with diverse attributes and intricate spatial relationships (Liu et al., 2022; Nie et al., 2024). Thus, compositional T2I generation with complex numerical instructions (Huang et al., 2025; Hu et al., 2024) is still an open problem. To cope with this challenge, plenty of literature focuses on developing new network architectures such as attention models (Chefer et al., 2023; Rassin et al., 2023; Meral et al., 2024; Kim et al., 2023), or introducing intermediate structures like object layout (Chen et al., 2024; Dahary et al., 2024; Wang et al., 2024c). Unlike recent methods that require synthesized ground-truth images (Gao et al., 2024; Sun et al., 2023a) or intermediate skeletons (Nie et al., 2024) for supervised fine-tuning, our approach takes a data-centric perspective and enhances compositional generalization solely through textual prompts, applying reinforcement learning (RL) without requiring ground-truth image outputs. However, large-scale compositional RL training can be highly unstable due to the heterogeneous capability requirements of compositional T2I generation tasks, which encompass object existence, attribute existence, relational understanding, and numerical counting.

To address this challenge, we propose **CompGen**, a novel compositional curriculum reinforcement learning framework for text-to-image (T2I) generation. CompGen draws inspiration from human cognitive development, which follows a curriculum learning progression: first mastering the recognition and generation of individual objects and their attributes within simple relational contexts, then gradually learning to understand and create complex multi-object compositions involving multiple relations. Specifically, our approach leverages scene graphs (Krishna et al., 2017) as a compositional

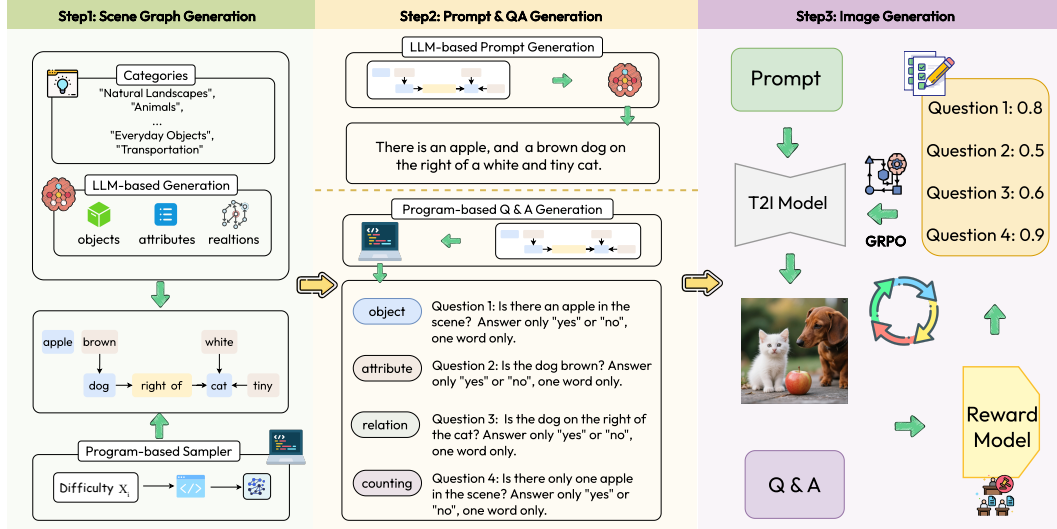


Figure 1: Overview of our CompGen framework, which is incentivized to construct a curriculum through end-to-end reinforcement learning without requiring ground-truth images.

representation of visual scenes to systematically generate training data with controllable complexity. First, we employ LLMs to generate diverse scene graph assets, where each scene graph encodes a structured representation of objects, their attributes, and relational dependencies. Second, we introduce a novel difficulty criterion that quantifies compositional complexity based on scene graph structural properties, including entity count, attribute diversity, and relational interconnectedness. Third, leveraging this difficulty metric, we develop an adaptive Markov Chain Monte Carlo (MCMC) sampling algorithm (Geyer, 1992) to systematically generate scene graphs at targeted complexity levels, thereby enabling precise curriculum control throughout the training process. Finally, for each sampled scene graph, we synthesize corresponding text prompts for image generation and construct comprehensive visual question-answer pairs for assessments, namely object existence, object counting, attribute recognition, and relational understanding. These question-answer pairs subsequently serve as reward metrics within our RL framework, guiding the model toward improved compositional reasoning performance.

Our extensive experiments demonstrate that CompGen significantly strengthens compositional generation capabilities across both diffusion and autoregressive T2I architectures. On five established compositional generation benchmarks — GenEval (Ghosh et al., 2023), DPG (Hu et al., 2024), TIFA (Hu et al., 2023), T2I-Bench (Huang et al., 2025), and DSG (Cho et al., 2023) — our approach consistently outperforms baseline models, achieving an average improvement of 12.52% when applied to Stable-Diffusion-1.5 and 3.44% when applied to LlamaGen. Furthermore, our analysis reveals that different training schedulers produce distinct scaling behaviors in GRPO training, with easy-to-hard progression and Gaussian sampling strategies demonstrating superior compositional performance and extending the scaling boundaries compared to random sampling approaches.

2 RELATED WORK

Large text-to-image (T2I) generative models have attracted considerable attention in recent years and can be broadly categorized into two main families: diffusion-based models (Ramesh et al., 2022; Rombach et al., 2022; Podell et al., 2023; Balaji et al., 2022) and auto-regressive models (Ramesh et al., 2021; Yu et al., 2022; Chang et al., 2023). Beyond purely improving visual quality, recent investigations (Wang et al., 2024a;b) have focused on enhancing prompt-following capabilities, particularly for compositional prompts. However, existing T2I models often struggle with compositional understanding, leading to issues such as object omission and incorrect attribute binding (Okawa et al., 2023; Huang et al., 2025). To address these limitations, recent efforts to improve compositional alignment can be grouped into three main categories. **Attention-based methods** modify attention maps within the UNet architecture to enforce object presence and spatial separation (Chefer et al., 2023; Rassin et al., 2023; Feng et al., 2022; Meral et al., 2024; Kim et al., 2023). For instance, DenseDiffusion (Kim et al., 2023) adjusts attention scores in both cross-attention and self-attention

layers to ensure object features align with specified image regions, while CONFORM (Meral et al., 2024) strengthens associations between relevant objects and attributes through contrastive objectives. However, these approaches face scalability and computational efficiency limitations as they operate only during inference. **Planning-based methods** utilize intermediate structures such as object layouts — either manually defined (Chen et al., 2024; Dahary et al., 2024; Wang et al., 2024c) or generated by large language models (Gani et al., 2023; Lian et al., 2023) — to guide image synthesis. Some approaches incorporate additional modules like visual question-answering models or captioning models for refinement (Yang et al., 2024; Wu et al., 2024; Wen et al., 2023); however, these additions increase inference costs and may still suffer from incorrect attribute bindings due to inherent model limitations. **Learning-based methods** focus on training-time improvements, including fine-tuning diffusion models with vision-language supervision (Wang et al., 2024c; 2022; Fan et al., 2023) or employing reinforcement learning techniques (Fan et al., 2023; Black et al., 2023). Caption-guided optimization represents another promising direction in this category (Fang et al., 2023; Ma et al., 2024). Unlike previous approaches that require additional inputs or architectural modifications, our method adopts a data-centric strategy to enhance compositional generalization through textual prompts and targeted training alone, without increasing inference costs or altering model architecture.

3 PRELIMINARY: SCENE-GRAPH AS A DIFFICULTY MEASURER

The core principle of CompGen is to progressively develop compositional generation capabilities through curriculum learning, advancing from simple to complex samples. The central challenge in this approach lies in establishing a principled framework for defining and quantifying the difficulty of compositional samples. Following the formalized image representation framework introduced by Krishna et al. (2017), scene graphs provide a structured approach to capturing compositional complexity. We therefore propose to measure sample difficulty through the compositional complexity inherent in scene graph structures. Formally, we provide the definition of difficulty based on scene graphs as follows:

Definition 1 (Scene Graph Formulated Difficulty). *Given a scene graph $\mathcal{G} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$ where $\mathcal{O}$ is the set of objects, $\mathcal{A}$ is the set of attributes associated with the objects, and $\mathcal{R}$ is the set of relations, we measure the difficulty of $\mathcal{G}$ as:*

$$\text{Diff}(\mathcal{G}) = \|\mathcal{O}\| \cdot \max\left(1, \frac{\|\mathcal{A}\|}{\|\mathcal{O}\|}\right) \cdot \max\left(1, \frac{\|\mathcal{R}\|}{\|\mathcal{O}\|}\right), \quad (1)$$

where, the total number of objects is $\|\mathcal{O}\|$, the average attribute density is $\|\mathcal{A}\|/\|\mathcal{O}\|$, and the average relational connectivity is $\|\mathcal{R}\|/\|\mathcal{O}\|$ per object, where relations are treated as directed edges in the scene graph representation.

Eq. (1) demonstrates that the computational difficulty of $\mathcal{G}$ is determined by three key factors identified by the total number of objects, the average attribute density and the average relational connectivity. Consequently, the overall difficulty exhibits a positive correlation with the intrinsic complexity of the scene graph, reflecting both its structural density and semantic richness.

4 COMPGEN: RL TRAINING WITH COMPOSITION CURRICULUM

Given the difficulty measure introduced in Definition 1, we can construct a mapping function $\{\mathcal{G}\} \rightarrow \mathbb{R}^+$ over the space of scene graphs $\mathcal{G}$. Building upon this foundation, we formalized our problem as follows:

Definition 2 (Compositional Text-to-Image Generation with Difficulty Constraint RL). *Given a target difficulty range $[\text{Diff}_{\min}, \text{Diff}_{\max}]$ where $0 < \text{Diff}_{\min} \leq \text{Diff}_{\max}$, our objective is to generate a pair (T, R) consisting of: (i) An input text prompt T that describes a visual scene with compositional complexity and (ii) A reward function $R : \mathcal{I} \rightarrow \mathbb{R}$ that evaluates generated images $I \in \mathcal{I}$.*

Therefore, our methodology follows a structured pipeline: we first construct a scene graph $\mathcal{G}$ satisfying the difficulty constraint $\text{Diff}_{\min} \leq \text{Diff}(\mathcal{G}) \leq \text{Diff}_{\max}$, then derive the corresponding input text T from this graph. These text prompts are processed by a T2I model to generate images I , which are subsequently evaluated by the reward function R . The resulting rewards enable optimization of the T2I model within a RL framework, ensuring adherence to the specified difficulty constraints.

Based on the problem formulation above, our CompGen framework comprises three key components: (i) scene graph generation under difficulty constraints, (ii) input text derivation from scene graphs, and (iii) reward function design for reinforcement learning optimization. Each component is thoroughly examined in the following subsections.

4.1 SCENE GRAPH GENERATION VIA ADAPTIVE MARKOV CHAIN MONTE CARLO

Given the target difficulty bounds $\text{Diff}_{\min}$ and $\text{Diff}_{\max}$, we define the constrained sampling problem as the task of generating scene graphs $\mathcal{G}$ such that $\text{Diff}_{\min} \leq \text{Diff}(\mathcal{G}) \leq \text{Diff}_{\max}$. This problem presents significant computational challenges, requiring efficient exploration of the high-dimensional graph space while simultaneously maintaining detailed balance for unbiased sampling, preserving structural validity of scene graphs, and enabling adaptive exploration under the imposed difficulty constraints.

To overcome these challenges, we develop a Markov Chain Monte Carlo (MCMC) sampling strategy (Brooks, 1998) that efficiently targets specific difficulty ranges. This approach enables sampling from a distribution concentrated on graphs with desired difficulty levels while avoiding computationally prohibitive explicit enumeration. Our method employs a set of four reversible graph transformation operations $\mathcal{T} = \{t_i\}_{i=1}^4$, which correspond to node addition, node removal, edge addition, and edge removal operations. These transformations are designed to ensure both ergodicity and detailed balance—fundamental requirements for unbiased MCMC sampling. We further define $\mathcal{T}_{\text{valid}} \subseteq \mathcal{T}$ as the subset of transformations that preserve essential graph properties, including connectivity and structural validity. The sampling procedure is directed by an energy function (i.e., $\text{Energy}(\mathcal{G})$) that quantifies the deviation from the target difficulty range:

$$\text{Energy}(\mathcal{G}) := \Delta(\mathcal{G}) = \text{Dist}(\text{Diff}(\mathcal{G}), [\text{Diff}_{\min}, \text{Diff}_{\max}]) = \begin{cases} \text{Diff}_{\min} - \text{Diff}(\mathcal{G}), & \text{if } \text{Diff}(\mathcal{G}) < \text{Diff}_{\min} \\ \text{Diff}(\mathcal{G}) - \text{Diff}_{\max}, & \text{if } \text{Diff}(\mathcal{G}) > \text{Diff}_{\max} \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where $\text{Dist}(\text{Diff}(\mathcal{G}), [\text{Diff}_{\min}, \text{Diff}_{\max}])$ is the distance function that quantifies the deviation of a scalar $\text{Diff}(\mathcal{G})$ from the target interval $[\text{Diff}_{\min}, \text{Diff}_{\max}]$, with the specific computation given above. We then adopt the Metropolis-Hastings algorithm (Chib & Greenberg, 1995) with acceptance probability defined as:

$$\text{Acc}(\mathcal{G}'|\mathcal{G}) = \min \left(1, \frac{\pi(\mathcal{G}')q(\mathcal{G}|\mathcal{G}')}{\pi(\mathcal{G})q(\mathcal{G}'|\mathcal{G})} \right), \quad (3)$$

where $\pi(\mathcal{G}) \propto \exp(-\text{Energy}(\mathcal{G})/\tau)$ is the target distribution and $q(\mathcal{G}'|\mathcal{G})$ is the proposal distribution. Since all our graph operations are symmetric, the proposal ratio cancels, yielding:

$$\text{Acc}(\mathcal{G}'|\mathcal{G}) = \min \left(1, \exp \left(\frac{\Delta(\mathcal{G}) - \Delta(\mathcal{G}')}{\tau} \right) \right). \quad (4)$$

To balance exploration and exploitation, we employ a logarithmic cooling schedule $\tau = 1/\log(t+2)$, which allows broad early exploration and progressively tighter focus near the target range. To avoid stagnation in local minima, the sampler reinitializes after a fixed number of stalled iterations, reshuffling graph structures while preserving key properties such as node counts and edge density.

4.2 REINFORCEMENT LEARNING STRATEGIES WITH COMPOSITION CURRICULUM

Having obtained a scene graph $\mathcal{G}$ through the constrained sampling process according to the above subsection, we turn to the generation of input text prompts T and the design of appropriate reward functions R .

Input Text Generation. CompGen transforms structured scene graphs into natural language prompts while preserving their original compositional difficulty. Given a scene graph $\mathcal{G} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$ with associated difficulty $\text{Diff}(\mathcal{D}) \in [\text{Diff}_{\min}, \text{Diff}_{\max}]$, we employ constrained Large Language Model (LLM)-based generation with strict constraints to produce descriptions that exactly match the graph’s specifications. The generation process combines precise constraints with LLM capabilities to maintain strict fidelity to the source graph. This is achieved through three core mechanisms: mandatory inclusion of all objects and attributes ensures exact term matching, while structural and multi-stage

content validation preserve relationships and integrity. By combining the linguistic flexibility of LLMs with systematic safeguards such as controlled temperature scheduling and iterative validation, the process guarantees that the generated prompt fully reflects all elements of the original graph and preserves its difficulty. While implementing, we use Deepseek-V3 (Liu et al., 2024) as the LLM.

Reward Generation. For any text-to-image (T2I) model, generated images can be obtained from input text prompts. We evaluate the compositional accuracy of these images through structured question-answering pairs, which are systematically constructed from the sampled scene graphs. Specifically, we choose a programmatic approach (Gao et al., 2024) to generate precise and comprehensive binary questions that fully cover the scene graphs. Questions are generated directly from scene graphs $\mathcal{G} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$. Since the scene graph encodes three types of information—objects, attributes, and relations—and the number of repeated objects is also a critical factor, we design four types of questions to cover all aspects: object verification questions (Q_{obj}) assessing presence of $o \in \mathcal{O}$, count questions (Q_{count}) verifying the number of occurrences of repeated objects in $\mathcal{O}$, attribute validation questions (Q_{attr}) checking $a \in \mathcal{A}$ for corresponding objects, and relation confirmation questions (Q_{rel}) verifying $r \in \mathcal{R}$ between object pairs.

Given the binary question-answer pairs we constructed, and inspired by VQAScore (Lin et al., 2024), we adopt a multimodal large language model (MLLM)’s predicted probability of answering “yes” as the initial fine-grained reward score. We take the average of the VQA scores for all questions corresponding to each image as the reward signal for reinforcement training, enabling the T2I model to perform policy updates. For ease of notation, we use $p_{\text{reward}}(\cdot)$ to denote the MLLM in use. While practical, we apply Llava-v1.6-13B (Liu et al., 2023), and we also evaluate different MLLMs in Section 5.3

RL Training. During training, we employ Group Relative Policy Optimization (GRPO) to optimize T2I model’s compositional generation ability. For each text prompt T sampled from dataset $\mathcal{D}$, GRPO generates G distinct images $\{I^{(1)}, I^{(2)}, \dots, I^{(G)}\}$ using the current policy $p_{\theta_{\text{old}}}$. The policy is optimized by maximizing:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{T \sim \mathcal{D}} \left[\frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}}{\pi_{\theta_{\text{old}}}} A_i, \quad \text{clip} \left(\frac{\pi_{\theta}}{\pi_{\theta_{\text{old}}}}, 1 - \epsilon, 1 + \epsilon \right) A_i \right) - \beta \text{KL} \left(p_{\theta}(\cdot|T) \parallel p_{\text{ref}}(\cdot|T) \right) \right) \right], \quad (5)$$

where $\pi_{\theta} = p_{\theta}(I^{(i)}|T)$ denotes the probability of generating image $I^{(i)}$ given text prompt T under the current policy, $\pi_{\theta_{\text{old}}} = p_{\theta_{\text{old}}}(I^{(i)}|T)$ is the probability under the old policy, ϵ and β are hyperparameters for clipping and KL regularization respectively, and p_{ref} is the reference policy.

For each generated image $I^{(i)}$, we evaluate its quality using binary question-answer pairs through our reward model as:

$$r_j^{(i)} = p_{\text{reward}}(\text{answer}_j | I^{(i)}, \text{question}_j), \quad (6)$$

where $r_j^{(i)} \in [0, 1]$ represents the reward model’s confidence that image $I^{(i)}$ correctly answers binary question. The overall reward for image $I^{(i)}$ is computed as the average score across all the sampled questions: $r^{(i)} = \frac{1}{M} \sum_{j=1}^M r_j^{(i)}$, where M is the number of samples. The advantages are normalized within each group of G images:

$$A_i = \frac{r^{(i)} - \text{mean}(\{r^{(k)}\}_{k=1}^G)}{\text{std}(\{r^{(k)}\}_{k=1}^G)}, \quad (7)$$

where $\text{mean}(\{r^{(k)}\}_{k=1}^G)$ and $\text{std}(\{r^{(k)}\}_{k=1}^G)$ denote the mean and standard deviation of rewards within the group, ensuring zero-mean unit-variance normalization of advantages.

Table 1: Comparison of different T2I models on compositional generation benchmarks. Upward arrows indicate improvement direction; red numbers show improvement magnitude over baseline.

Model	# Params	GenEval	DPG	TIFA	T2I-Bench	DSG	Avg.
<i>Diffusion</i>							
Stable-Diffusion-1.4	0.9B	42.04%	61.89%	79.14%	30.80%	61.71%	55.12%
Stable-Diffusion-1.5	0.9B	42.08%	62.24%	78.67%	29.94%	61.57%	54.90%
Stable-Diffusion-2.1	0.9B	50.00%	65.47%	82.00%	32.01%	68.09%	59.51%
Playground-V2	2.6B	59.00%	74.54%	86.20%	36.13%	74.54%	66.08%
Stable-Diffusion-XL	2.6B	55.87%	74.65%	83.50%	31.30%	83.40%	65.74%
<i>Diffusion Transformer</i>							
PixArt- α	0.6B	48.00%	71.11%	82.90%	41.17%	71.11%	62.85%
Lumina-Next	1.7B	46.00%	75.66%	79.98%	34.57%	70.61%	61.36%
<i>AutoRegressive</i>							
Show-o	1.3B	56.00%	67.27%	86.28%	29.00%	77.00%	63.11%
LlamaGen	775M	31.28%	42.92%	75.03%	33.26%	58.30%	48.16%
Emu3	14B	54.00%	74.19%	81.86%	31.20%	70.31%	62.31%
minDALL-E	1.3B	23.00%	55.23%	79.40%	18.98%	45.63%	44.45%
<i>Ours</i>							
Stable-Diffusion-1.5 w/o ours	0.9B	53.88% ($\uparrow 11.80$)	78.67% ($\uparrow 16.43$)	85.71% ($\uparrow 7.04$)	37.68% ($\uparrow 7.74$)	77.16% ($\uparrow 15.59$)	66.62% ($\uparrow 11.72$)
LlamaGen w/o ours	775M	35.71% ($\uparrow 4.43$)	48.67% ($\uparrow 5.75$)	77.92% ($\uparrow 2.89$)	36.14% ($\uparrow 2.88$)	59.57% ($\uparrow 1.27$)	51.60% ($\uparrow 3.44$)

5 EXPERIMENT

5.1 EXPERIMENTAL SETUP

Datasets and Benchmarks. To thoroughly assess the model’s compositional capability, we evaluate it on the following five compositional benchmarks: (i) **Geneval** (Ghosh et al., 2023), (ii) **T2I-CompBench** (Huang et al., 2025), (iii) **TIFA** (Hu et al., 2023), (iii) **DPG-Bench** (Hu et al., 2024), (iv) **DSG** (Cho et al., 2023). We provide a detailed description for each dataset in Appendix A.

Baseline Models. To validate the effectiveness of our proposed method, we conduct comprehensive comparisons against several state-of-the-art text-to-image generation models. These include diffusion models such as Stable Diffusion 1.4/2.1 (Rombach et al., 2022), Playground v2 (Li et al.), and Stable Diffusion XL (Podell et al., 2023); Diffusion Transformer-based models including PixArt-alpha (Chen et al., 2023) and LUMINA-NEXT (Zhuo et al., 2024); as well as auto-regressive T2I models such as Show-o (Xie et al., 2024), Emu3 (Sun et al., 2023b) and minDALL-E (Kim et al., 2021).

To generate high-quality training data for compositional RL, we employ the CompGen framework to construct 10K samples evenly distributed across difficulty levels ranging from 1 to 10. Each sample includes text prompts and corresponding question-answer pairs, with answers set to “yes” to enable a VQA model to evaluate whether the generated images contain the described content. To demonstrate the effectiveness and generalizability of our CompGen for enhancing compositional generation capabilities of T2I models, we select Stable-Diffusion v1.5 (Rombach et al., 2022) and LlamaGen (Sun et al., 2024), which cover two prominent T2I model architectures.

5.2 PERFORMANCE COMPARISONS AND ANALYSIS

CompGen significantly enhances T2I compositional generation capabilities. Our experimental results demonstrate that CompGen consistently delivers substantial improvements across all compositional generation benchmarks. Specifically, when applied to Stable Diffusion 1.5, CompGen achieves an average performance gain of 11.72%, elevating the baseline from 54.90% to 66.62%. This enhanced performance not only surpasses the original baseline but also outperforms several stronger contemporary models, including Stable Diffusion 2.1 (59.51%), PixArt- α (62.85%), and even competitive autoregressive models like Emu3 (62.31%). Notably, our enhanced SD-1.5 model with only 0.9B parameters achieves comparable or superior compositional generation performance to much larger models such as Playground-V2 (2.6B, 66.08%), demonstrating the efficiency and effectiveness of our approach in bridging the compositional generation gap without requiring substantial architectural modifications or parameter scaling.

Table 2: Performance changes of CompGen with various reward models.

Model	Reward Model	GenEval	DPG	TIFA	T2I-Bench	DSG	Avg.
<i>Diffusion</i>							
Stable-Diffusion-1.5	–	42.08%	62.24%	78.67%	29.94%	61.57%	54.10%
Stable-Diffusion-1.5 w/ VQAScore	LLaVA-v1.6-13B	44.02%	73.41%	80.19%	36.36%	73.23%	61.42%
	CLIP-FlanT5-XXL	45.11%	71.26%	81.42%	31.60%	73.74%	60.62%
	InstructBLIP	42.04%	64.00%	76.29%	39.21%	64.58%	57.22%
	LLaVA-v1.5-13B	49.23%	74.54%	84.84%	37.43%	75.97%	64.40%
CompGen (Stable-Diffusion-1.5)	LLaVA-v1.6-13B	53.88%	78.67%	85.71%	37.68%	77.16%	66.62%
<i>AutoRegressive</i>							
LlamaGen	–	31.28%	42.92%	75.03%	33.26%	58.30%	48.16%
LlamaGen w/ VQAScore	LLaVA-v1.6-13B	33.15%	44.51%	75.94%	34.28%	59.46%	49.47%
	CLIP-FlanT5-XXL	32.47%	44.57%	75.98%	33.71%	58.60%	49.07%
	InstructBLIP	32.03%	44.75%	75.54%	33.52%	57.98%	48.76%
	LLaVA-v1.5-13B	31.09%	45.72%	75.96%	31.28%	57.54%	48.32%
CompGen (LlamaGen)	LLaVA-v1.6-13B	35.71%	48.67%	77.92%	36.14%	59.57%	51.60%

Stronger base models exhibit greater improvements with CompGen. Our analysis reveals a compelling correlation between base model capability and the magnitude of improvement achieved through CompGen training. The more capable Stable Diffusion 1.5 baseline (54.90%) experiences a substantial 11.72% improvement, with particularly pronounced gains in DPG (+16.43%) and DSG (+15.59%) benchmarks that specifically evaluate complex compositional reasoning. In contrast, the weaker LlamaGen baseline (48.16%) shows more modest improvements of 3.44% on average, with relatively smaller gains across individual benchmarks. This pattern suggests that CompGen’s effectiveness is amplified when applied to models with stronger foundational generation capabilities.

CompGen generalizes beyond compositionality without overfitting to training objectives. As illustrated in Figure 3, our enhanced models maintain high-quality image generation across diverse scenarios while significantly improving compositional accuracy. The qualitative results demonstrate that CompGen effectively improves complex multi-object compositions, spatial relationships, and attribute bindings while preserving overall image fidelity and avoiding the introduction of artifacts.

5.3 ABLATION STUDY

Ablation on Reward Design. We conduct an ablation study comparing our fine-grained reward design against VQAScore (Lin et al., 2024) to validate the effectiveness of our compositional reward formulation. While VQAScore employs a single binary question asking whether the prompt matches the image, our approach decomposes the evaluation into multiple fine-grained aspects including object existence, attribute presence, relational understanding, and counting accuracy. As shown in Table 2, our CompGen framework consistently outperforms VQAScore across both diffusion and autoregressive models. Specifically, when applied to Stable-Diffusion-1.5 with LLaVA-v1.6-13B as the reward model, CompGen achieves 66.62% average performance compared to 61.42% for VQAScore, representing a substantial 5.2 percentage point improvement. The gains are particularly pronounced on complex compositional benchmarks like GenEval (+9.86%) and DPG (+5.26%), demonstrating that fine-grained reward signals better capture the nuanced requirements of compositional generation. Interestingly, while autoregressive models show more modest improvements (51.60% vs 49.47% for LlamaGen), the consistent gains across different architectures validate the generalizability of our reward design.

Ablation on Reward Model. To investigate the impact of reward model capability on CompGen performance, we conduct ablation experiments across four different vision-language models: LLaVA-v1.6-13B (Liu et al., 2023), LLaVA-v1.5-13B (Liu et al., 2023), CLIP-FlanT5-XXL (Radford et al., 2021), and InstructBLIP-FlanT5-XXL (Dai et al., 2023). All experiments utilize our compositional fine-grained reward design with uniformly sampled 10K training data on both Stable-Diffusion-1.5 and LlamaGen base models. As shown in Table 2, the choice of reward model significantly influences

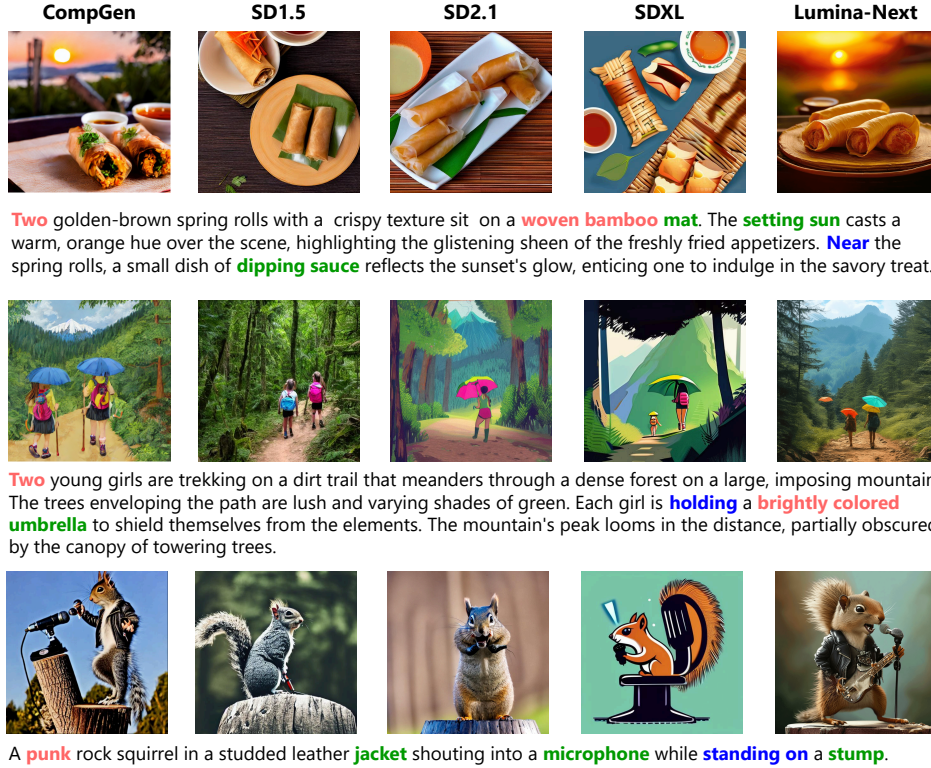


Figure 2: Qualitative comparison of compositional generation across CompGen and other strong T2I models.

the effectiveness of our approach, with LLaVA-v1.6-13B consistently achieving the best performance across all benchmarks. Specifically, LLaVA-v1.6-13B outperforms the second-best LLaVA-v1.5-13B by 2.22% and 3.28% average improvement for diffusion and autoregressive models respectively, highlighting the importance of advanced vision-language understanding capabilities. Notably, the performance gap between different reward models is more pronounced in diffusion models (66.62% vs 57.22% for best vs worst) compared to autoregressive models (51.60% vs 48.32%), suggesting that diffusion models are more sensitive to reward signal quality. InstructBLIP shows surprisingly weaker performance despite being instruction-tuned, potentially due to its limited compositional reasoning capabilities.

5.4 CURRICULUM LEARNING EXTENDS PERFORMANCE SCALING BOUNDARY.

We investigate three curriculum learning strategies (Parashar et al., 2025) that control the presentation of training samples according to their difficulty as follows. (i) The easy-to-hard sampling strategy follows a deterministic schedule where samples are organized from easy to hard, enabling the model to gradually acquire increasingly complex patterns. (ii) The balance sampling strategy instead performs uniform random sampling across all difficulty levels, ensuring unbiased exposure throughout training. (iii) The Gaussian sampling strategy models the sampling distribution with a bell-shaped curve whose center progressively moves from easier to harder samples, with tunable parameters controlling the spread and transition speed. Here, we investigate whether curriculum learning can effectively extend the training boundary using Stable Diffusion 1.5 as the base model and test on the GenEval Benchmark. We provide a detailed description in Appendix B.

Our experimental results demonstrate that curriculum learning strategies significantly enhance CompGen performance and successfully extend the scaling boundary beyond conventional training approaches. As illustrated in Figure 4, all three curriculum strategies substantially outperform the baseline model, which maintains a static accuracy of 42% regardless of training duration. The balance sampling strategy achieves a peak performance of 52.1% at 420 training steps, representing a 24% relative improvement over the baseline. More remarkably, the easy-to-hard sampling strategy reaches

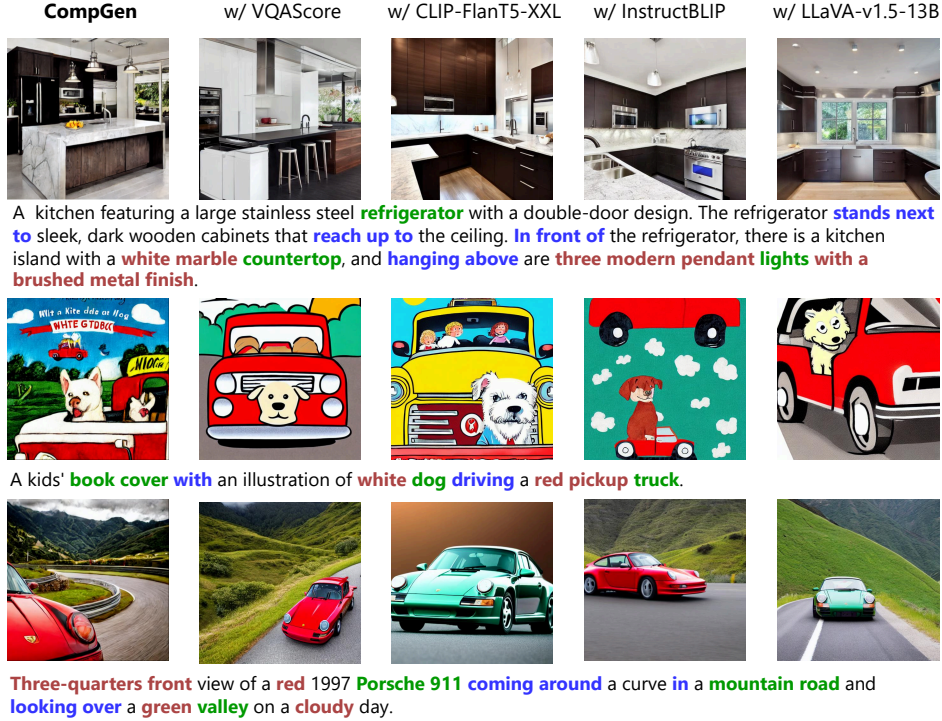


Figure 3: Qualitative comparison of compositional generation across CompGen and ablation models.

53.8% accuracy at 580 steps, while Gaussian sampling attains the highest peak performance of 54.6% at 500 steps — a 30% relative improvement.

Crucially, these results reveal that curriculum learning not only improves absolute performance but also extends the effective training horizon: while traditional approaches plateau early, curriculum strategies continue scaling performance gains well beyond 500 training steps before exhibiting gradual decline. The Gaussian sampling strategy demonstrates the most efficient scaling trajectory, achieving optimal performance with fewer training steps, while easy-to-hard sampling exhibits the most extended scaling boundary, maintaining competitive performance the longest during extended training.

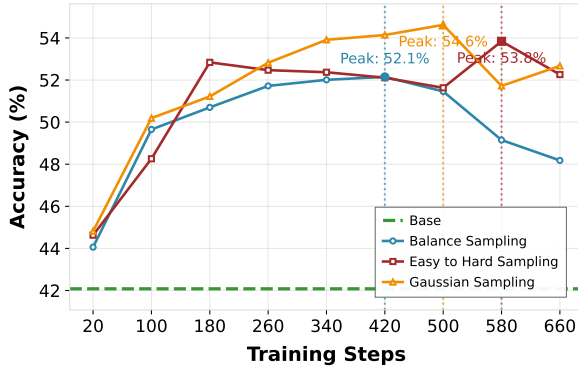


Figure 4: Scaling trend of CompGen with different curriculum learning strategies.

6 CONCLUSION

In this work, we introduced CompGen, a novel compositional curriculum reinforcement learning framework that leverages scene graphs and incorporates a principled difficulty criterion with adaptive MCMC sampling, our approach, free from the requirements of need of ground-truth images, systematically improves the ability of T2I models to generate complex scenes with multiple objects, diverse attributes, and intricate spatial-semantic relationships. In the future, it would be interesting to explore more sophisticated metrics that incorporate semantic complexity, visual realism requirements, and cross-modal alignment challenges. Additionally, developing adaptive curriculum strategies that dynamically adjust difficulty based on model performance might further optimize training efficiency.

ETHICS STATEMENT

This work focuses on the study of text-to-image generation and synthesis. All datasets used in our experiments are publicly available and commonly adopted in prior research. We followed the respective dataset licenses and usage terms. No personally identifiable information (PII) or sensitive private data was collected, generated, or annotated by the authors. Our study does not raise direct ethical concerns such as misuse of personal data, harmful content generation, or bias amplification beyond what is already inherent in the benchmark datasets. All generated images in our experiments are for research purposes only and do not involve the creation of deceptive or malicious content.

REPRODUCIBILITY STATEMENT

In order to ensure reproducibility, we provide a comprehensive description of datasets, model implementations, and experimental settings in the main paper and the appendix. The benchmarks and evaluation metrics we used are standard and publicly available. All baselines are either taken from released model checkpoints or trained/evaluated with publicly accessible open-source implementations. To further promote reproducibility, hyperparameters, training details, and evaluation protocol are clearly documented. We release our code at <https://anonymous.4open.science/status/T2I-Scaling-DICC> to enable the community to fully reproduce our results. We commit to following ICLR guidelines for transparency and reproducibility in scientific reporting.

REFERENCES

- Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 1, 2
- Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3
- Stephen Brooks. Markov chain monte carlo method and its application. *Journal of the royal statistical society: series D (the Statistician)*, 47(1):69–100, 1998. 4
- Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2
- Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 2
- Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 6
- Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5343–5353, 2024. 1, 3
- Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995. 4
- Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *arXiv preprint arXiv:2310.18235*, 2023. 2, 6, 14
- Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. In *European Conference on Computer Vision*, pp. 432–448. Springer, 2024. 1, 3

- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 7
- Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36:79858–79885, 2023. 3
- Guian Fang, Zutao Jiang, Jianhua Han, Guangsong Lu, Hang Xu, and Xiaodan Liang. Boosting text-to-image diffusion models with fine-grained semantic rewards. *arXiv preprint arXiv:2305.19599*, 5, 2023. 3
- Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 2
- Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. *arXiv preprint arXiv:2310.10640*, 2023. 3
- Ziqi Gao, Weikai Huang, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. Generate any scene: Evaluating and improving text-to-vision generation with scene graph programming. *arXiv preprint arXiv:2412.08221*, 2024. 1, 5
- Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pp. 473–483, 1992. 2
- 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: 52132–52152, 2023. 2, 6, 14
- Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 1, 2, 6, 14
- Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20406–20417, 2023. 2, 6, 14
- Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 2, 6, 14
- Chiheon Kim Doyup Lee Saehoon Kim, Sanghun Cho, and Woonhyuk Baek. mindall-e on conceptual captions, 2021. 6
- Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7701–7711, 2023. 1, 2
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1, 3
- Daiqing Li, Aleks Kamko, Ali Sabet, Ehsan Akhgari, Linmiao Xu, and Suhail Doshi. Playground v2. URL [<https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic>] (<https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic>). 6
- Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655*, 2023. 3

- Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024. 5, 7
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 5
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5, 7
- Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pp. 423–439. Springer, 2022. 1
- Bingqi Ma, Zhuofan Zong, Guanglu Song, Hongsheng Li, and Yu Liu. Exploring the role of large language models in prompt encoding for diffusion models. *arXiv preprint arXiv:2406.11831*, 2024. 3
- Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9005–9014, 2024. 1, 2, 3
- Weili Nie, Sifei Liu, Morteza Mardani, Chao Liu, Benjamin Eckart, and Arash Vahdat. Compositional text-to-image generation with dense blob representations. *arXiv preprint arXiv:2405.08246*, 2024. 1
- Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36:50173–50195, 2023. 2
- Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, et al. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*, 2025. 8
- 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. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 6
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. Pmlr, 2021. 7
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021. 2
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1, 2
- Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023. 1, 2
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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. 2, 6
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

- Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd Van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. *arXiv preprint arXiv:2311.17946*, 2023a. 1
- Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 6
- Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023b. 6
- Fu-Yun Wang, Zhaoyang Huang, Alexander Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024a. 2
- Fu-Yun Wang, Xiaoshi Wu, Zhaoyang Huang, Xiaoyu Shi, Dazhong Shen, Guanglu Song, Yu Liu, and Hongsheng Li. Be-your-outpainter: Mastering video outpainting through input-specific adaptation. In *European Conference on Computer Vision*, pp. 153–168. Springer, 2024b. 2
- Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5544–5552, 2024c. 1, 3
- Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 3
- Song Wen, Guian Fang, Renrui Zhang, Peng Gao, Hao Dong, and Dimitris Metaxas. Improving compositional text-to-image generation with large vision-language models. *arXiv preprint arXiv:2310.06311*, 2023. 3
- Tsung-Han Wu, Long Lian, Joseph E Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting llm-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6327–6336, 2024. 3
- Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 6
- 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 *Forty-first International Conference on Machine Learning*, 2024. 3
- Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2
- Le Zhuo, Ruoyi Du, Han Xiao, Yangguang Li, Dongyang Liu, Rongjie Huang, Wenze Liu, Xiangyang Zhu, Fu-Yun Wang, Zhanyu Ma, et al. Lumina-next: Making lumina-t2x stronger and faster with next-dit. *Advances in Neural Information Processing Systems*, 37:131278–131315, 2024. 6

A DATASET

To thoroughly assess the model’s compositional capability, we evaluate it on the following five compositional benchmarks: (i) **Geneval** (Ghosh et al., 2023) consists of 553 highly structured prompts spanning six key evaluation dimensions: single object, dual objects, color, count, spatial positioning, and attribute binding. (ii) **T2I-CompBench** (Huang et al., 2025) comprises 6,000 compositional text prompts categorized into three main areas—attribute binding, object relationships, and complex compositions—and further divided into six subcategories: color binding, shape binding, texture binding, spatial relationships, non-spatial relationships, and complex compositions. (iii) **TIFA** (Hu et al., 2023) evaluates generation quality through 4,000 diverse text prompts and 25,000 automatically generated questions using a VQA (Visual Question Answering) model. It spans 12 question categories, including existence verification, object count, color identification, and spatial reasoning. (iii) **DPG-Bench** (Hu et al., 2024) includes 1,065 densely annotated prompts with an average token length of 83.91. These prompts describe complex visual scenarios involving multiple objects and modifiers. (iv) **DSG** (Cho et al., 2023) is an evaluation framework for text-to-image models that generates precise, semantically grounded questions in dependency graphs. DSG ensures reliable evaluations and is supported by the open-sourced DSG-1k benchmark, containing 1,060 diverse prompts.

B COMPOSITION CURRICULUM

- **Balance Scheduling.** As a straightforward method to prevent forgetting, the balance scheduling strategy allows for sampling from all tasks with equal probability throughout the entire training process. This can be viewed as a basic form of curriculum learning, or more generally, as the default behavior of most policy optimization algorithms when task difficulty is not considered. For M tasks, the sampling probability $\mathcal{P}_{\text{balanced}}(i, j)$ for task j at any training iteration i is set to:

$$\mathcal{P}_{\text{balanced}}(i, j) = \frac{1}{M}$$

This means that at each training step, all tasks have an equal chance of being selected. While this method effectively prevents the model from forgetting previously learned tasks, it might introduce more difficult tasks too early, leading to reward sparsity issues and potentially hindering the curriculum learning strategy from achieving optimal results.

- **Easy-to-Hard Scheduling.** Timestep curriculum learning strategies adopt a phased approach, dividing the training process into a series of incrementally difficult stages. In each training phase, the model focuses solely on tasks of a specific difficulty level. We can define the sampling indicator function for task j at training iteration i as $\mathcal{P}_{\text{timestep}}(i, j)$. For a total of M tasks ($j = 1, \dots, M$), this function takes a value of 1 if the current training iteration i falls within the predefined stage $[\tau_j, \tau_{j+1}]$ for task j ; otherwise, it is 0. Here, τ_j denotes the starting training step for the j -th stage, with $\tau_1 = 0$ and $\tau_{M+1} = N_T$ being the total number of training steps. This implies that at a given iteration i , only tasks corresponding to the current stage are sampled for training.

$$\mathcal{P}_{\text{timestep}}(i, j) = \begin{cases} 1, & \text{if } \tau_j \leq i < \tau_{j+1} \\ 0, & \text{otherwise} \end{cases}$$

Therefore, at training step i , the sampling distribution will be $[\mathcal{P}_{\text{timestep}}(i, 1), \dots, \mathcal{P}_{\text{timestep}}(i, M)]$.

- **Gaussian Scheduling.** Gaussian scheduling models task sampling as a mixture of Gaussian distributions to provide flexible and fine-grained control over the training process. Each task k ($k = 1, \dots, K$) is assumed to follow a one-dimensional Gaussian distribution with the same variance σ^2 but different means

$$\mu_k = k - 1.$$

A latent curriculum position x_t moves from easier to harder tasks as training step t increases:

$$x_t = \left(\frac{t}{T} \right)^\beta (K - 1),$$

where T is the total number of training steps, $\beta > 0$ controls the moving speed of x_t , and K is the number of tasks.

The unnormalized sampling score for task k at step t is

$$\mathbf{S}_{\text{Gaussian}}(t, k) = \exp\left(-\frac{(x_t - \mu_k)^2}{2\sigma^2}\right),$$

where σ determines the sampling concentration.

The normalized sampling probability is

$$\mathcal{P}_{\text{Gaussian}}(t, k) = \frac{\mathbf{S}_{\text{Gaussian}}(t, k)}{\sum_{m=1}^K \mathbf{S}_{\text{Gaussian}}(t, m)}.$$

Smaller σ produces sharper, stage-like transitions, while larger σ smooths the task shifts. A lower β slows the move toward harder tasks, allowing more training on easy tasks in the early phase.

C USE OF LLMs

During the preparation of this manuscript, we made limited use of publicly available large language models (LLMs) to assist with English writing. All technical content, including the formulation of ideas, design of methodologies, implementation of experiments, and interpretation of results, was entirely conceived and written by the authors without the involvement of LLMs. The role of LLMs was strictly confined to stylistic and linguistic improvements, in a manner comparable to grammar- or spell-checking software. We ensured that no novel research insights, data, or analyses were generated by LLMs, and all scientific claims and results presented in this work remain the sole responsibility of the authors.