Synthetic Curriculum Reinforces Compositional Text-to-Image Generation

Shijian Wang* [1,2,3] Runhao Fu* [3,5] Siyi Zhao[4] Qingqin Zhan[6] Xingjian Wang[6]
Jiarui Jin† [2] Yuan Lu† [2] Hanqian Wu† [1] Cunjian Chen[3]
[1]Southeast University [2]Xiaohongshu Inc. [3]Monash University
[4]Shanghai Jiao Tong University [5]Anhui University [6]Independent Researcher

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 named CompGen that addresses compositional weakness in existing T2I models. Specifically, we leverage scene graphs to establish a novel difficulty criterion for compositional ability and develop a corresponding adaptive Markov Chain Monte Carlo graph sampling algorithm. This difficulty-aware approach enables the synthesis of training curriculum data that progressively optimize T2I models through reinforcement learning. We integrate our curriculum learning approach into Group Relative Policy Optimization (GRPO) and investigate different curriculum scheduling strategies. Our experiments reveal that CompGen exhibits distinct scaling curves under different curriculum scheduling strategies, with easy-to-hard and Gaussian sampling strategies yielding superior scaling performance compared to random sampling. Extensive experiments demonstrate that CompGen significantly enhances compositional generation capabilities for both diffusion-based and auto-regressive T2I models, highlighting its effectiveness in improving the compositional T2I generation systems.

1. Introduction

Text-to-Image (T2I) generation has achieved remarkable progress in synthesizing visually compelling content from textual descriptions [2, 19, 49, 52, 55]. 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 [40, 46]. Thus, compositional T2I generation with complex instructions [26, 28] is still an open problem. To cope with this challenge, plenty of literature focuses on developing new network architectures such as attention models [8, 30, 43, 53], or introducing intermediate structures like object layout [9, 13, 64].

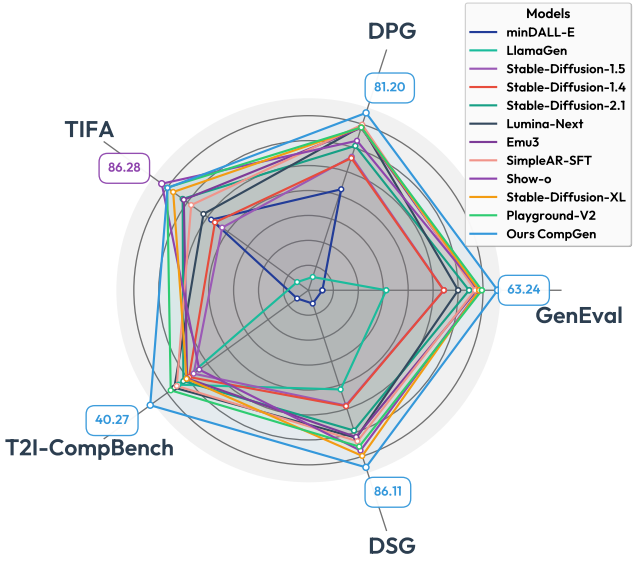

Figure 1. Overall performance of our CompGen, indicating that CompGen achieves state-of-the-art performance among models of the same scale.

Unlike recent methods that require synthesized ground-truth images [22, 44, 58] or intermediate skeletons [46] for supervised fine-tuning, our approach adopts a data-centric perspective that enhances compositional generalization through synthetic curriculum. Crucially, our method requires only textual prompts and applies reinforcement learning (RL) without the need for ground-truth image outputs. However, large-scale RL training for compositional T2I generation faces significant stability challenges due to the heterogeneous nature of compositional capabilities required [35, 45], which encompass object existence, attribute binding, relational understanding, and numerical counting.

*Equal contribution. Work done when Shijian internship at Xiaohongshu
†Corresponding authors

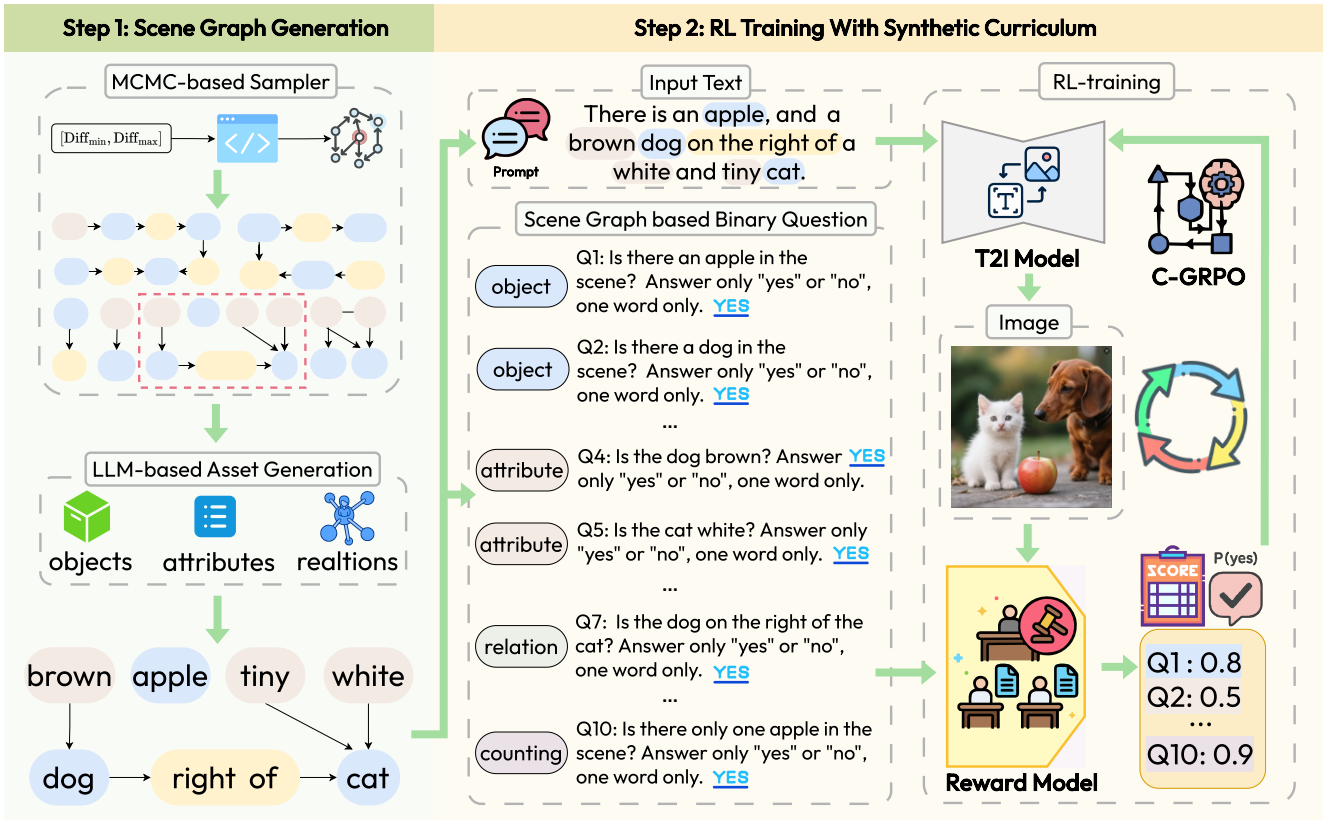

Figure 2. Overview of our CompGen framework, which is incentivized to construct a curriculum through end-to-end reinforcement learning without requiring ground-truth images.

To address this challenge, we propose **CompGen**, a novel compositional curriculum-based RL framework for 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 [31] as a compositional representation of visual scenes to systematically generate training data with controllable difficulty. We introduce a novel difficulty criterion that quantifies compositional complexity based on scene graph structural properties, including entity count, attribute diversity, and relational interconnectedness. Leveraging this difficulty metric, we then develop an adaptive Markov Chain Monte Carlo (MCMC) sampling algorithm [23] to systematically generate scene graphs at specific difficulty levels, thereby enabling precise curriculum control throughout the training process. For each sampled scene graph, we synthesize corresponding input text prompt for T2I 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 generation performance.

As presented in Figure 1, our extensive experiments demonstrate that CompGen significantly strengthens compositional generation capabilities across both diffusion and auto-regressive T2I architectures. On five established compositional generation benchmarks — GenEval [24], DPG [26], TIFA [27], T2I-CompBench [28], and DSG [12] — our approach consistently outperforms baseline models, achieving an average improvement of 11.72% when applied to Stable-Diffusion-1.5 and 7.61% when applied to SimpleAR. Moreover, our analysis reveals that curriculum scheduling critically impacts scaling behavior in curriculum-based GRPO training: easy-to-hard and Gaussian sampling strategies demonstrate superior performance and extended scaling potential 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 [2, 49, 52, 54] and auto-regressive models [7, 51, 71]. Beyond purely improving visual quality, recent investigations [61, 62] have focused on enhancing prompt-following capabilities, particularly for compositional prompts. How-

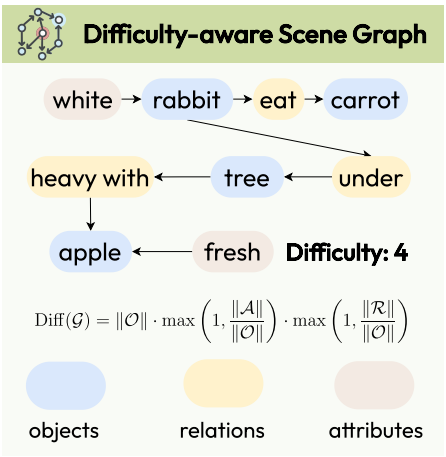

Figure 3. An illustrated example of scene graph corresponding to a specific difficulty level.

ever, existing T2I models often struggle with compositional understanding, leading to issues such as object omission and incorrect attribute binding [28, 47]. 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 [8, 18, 30, 43, 53]. For instance, DenseDiffusion [30] adjusts attention scores in both cross-attention and self-attention layers to ensure object features align with specified image regions, while CONFORM [43] 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 [9, 13, 64] or generated by large language models [21, 34] — to guide image synthesis. Some approaches incorporate additional modules like visual question-answering models or captioning models for refinement [66, 67, 70]; 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 [16, 64, 65] or employing reinforcement learning techniques [3, 16]. Caption-guided optimization represents another promising direction in this category [17, 42].

Unlike previous approaches that require additional inputs or architectural modifications, our method adopts a data-centric strategy to enhance compositional generalization through synthetic curriculum-based RL, without increasing inference costs or altering model architecture.

3. 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. [31], scene graphs (composed of three types of nodes: objects, attributes, relationships, and directed edges) 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:*

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

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.

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. We provide an illustrated case of our scene graph of a specific difficulty level in Figure 3. We also empirically compare the CompGen performance of using Eq. (1) with other formulations in Section 5.4.

4. RL Training with Synthetic Curriculum

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

Definition 2 (Synthetic Curriculum-based RL)**.** *Given a target difficulty range $[\mathrm{Diff}_{\min}, \mathrm{Diff}_{\max}]$ where $0 < \mathrm{Diff}_{\min} \leq \mathrm{Diff}_{\max}$, a synthetic curriculum for T2I can be formulated as a T2I task with an input text T and a reward r corresponding to the output image I. Therefore, our methodology follows a structured pipeline: we first construct a scene graph $\mathcal{G}$ satisfying the difficulty constraint $\mathrm{Diff}_{\min} \leq \mathrm{Diff}(\mathcal{G}) \leq \mathrm{Diff}_{\max}$, then derive the corresponding input text T from this graph.*

These text prompts are processed by a T2I model to generate images Is, 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 aforementioned problem formulation, our CompGen framework operates through a two-stage methodology, as demonstrated in Figure 2. Specifically, the framework first generates a scene graph with a specific range of difficulty levels to construct a synthetic curriculum, and subsequently trains T2I models through curriculum-based RL.

4.1. Scene Graph Generation via Adaptive Markov Chain Monte Carlo Sampling

Given the target range of difficulty levels Diff_{min} and Diff_{max}, we define the constrained generation problem as the task of sampling scene graphs $\mathcal{G}$ such that $\text{Diff}_{\text{min}} \leq \text{Diff}(\mathcal{G}) \leq \text{Diff}_{\text{max}}$. A brute-force enumeration of all possible scene graphs to identify those satisfying the difficulty constraints is computationally intractable due to the exponentially large and high-dimensional nature of the graph space. To efficiently navigate this combinatorial landscape, we reformulate the task as an iterative sampling problem and employ a targeted sampling strategy based on Markov Chain Monte Carlo (MCMC) [5].

Our approach begins with an initial scene graph $\mathcal{G}_0$ and iteratively refines it through systematic modifications to discover graph structures that satisfy the specific difficulty constraints. In practice, we initialize $\mathcal{G}_0$ as a minimally complex baseline graph, typically comprising a randomly selected small number of object nodes, thereby establishing a neutral starting point for the MCMC sampling process. To enable this graph sampling process, we define a set of two kinds of reversible graph transformation operations, namely $\mathcal{T}_{\text{add}}$ and $\mathcal{T}_{\text{delete}}$. The addition operation, $\mathcal{T}_{\text{add}}$, introduces a new element (i.e., an attribute/relation node and its associated edges), while its inverse, the delete operation $\mathcal{T}_{\text{delete}}$, eliminates an existing element (i.e., a node and all its incident edges). At each MCMC sampling step, a candidate graph $\mathcal{G}'$ is proposed from the current state $\mathcal{G}$ by randomly selecting and applying one of these transformations, thereby defining our proposal distribution $q(\mathcal{G}'|\mathcal{G})$. Crucially, these operations are designed to maintain detailed balance through symmetric transition probabilities; specifically, the probability of proposing to add a particular element to graph $\mathcal{G}$ equals the probability of proposing to delete that identical element from the resulting graph $\mathcal{G}'$.

To systematically guide the MCMC sampling process toward the target difficulty range, we define an energy function $\text{Energy}(\mathcal{G})$ that quantifies the extent to which a graph's difficulty deviates from the specified constraints:

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

The energy function evaluates to zero when a graph's difficulty falls within the target range and assumes positive values otherwise, thereby establishing the optimization objective of identifying graphs with minimal (ideally zero) energy. We subsequently employ the Metropolis-Hastings algorithm [11] to make stochastic acceptance decisions for proposed transformations. Transitions to lower-energy states (i.e., configurations closer to the target difficulty range) are unconditionally accepted, while transitions to higher-energy states are accepted with probability determined by the Metropolis criterion to facilitate exploration and prevent convergence to suboptimal local minima. The acceptance probability is formally defined as:

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

where $\pi(\mathcal{G}) \propto \exp(-\text{Energy}(\mathcal{G})/\tau)$ is the target distribution that favors low-energy graphs. Due to the symmetric design of our proposal distribution $q(\mathcal{G}'|\mathcal{G})$, the proposal ratio $\frac{q(\mathcal{G}|\mathcal{G}')}{q(\mathcal{G}'|\mathcal{G})}$ is unity, simplifying the acceptance probability to depend only on the change in energy $\text{Acc}(\mathcal{G}'|\mathcal{G}) = \min\left(1, \exp\left(\frac{\text{Energy}(\mathcal{G})-\text{Energy}(\mathcal{G}')}{\tau}\right)\right)$. Here, temperature parameter τ governs the trade-off between exploration and exploitation, thereby rendering our sampling algorithm adaptive to the current graph state and energy landscape.

Concretely, we employ an adaptive temperature schema analogous to simulated annealing. We initialize with a high temperature τ to facilitate extensive exploration of the graph space and progressively reduce it throughout the sampling process. This annealing schedule enables the sampler to initially identify promising regions of the solution space before progressively refining the search to achieve precise convergence on graphs that satisfy the specified difficulty constraints. Our empirical evaluation demonstrates that this approach outperforms baseline methods in both sampling efficiency and graph diversity, as detailed in Appendix H.1. We additionally investigate the effects of various initialization strategies on sampling performance and node type diversity in Appendix H.2. The complete algorithmic procedure is presented in Algorithm 2 in Appendix A.

4.2. Curriculum-based Group Relative Policy Optimization

Having obtained a scene graph $\mathcal{G}$ through the constrained sampling process described above, we instantiate the abstract

graph structure by randomly sampling concrete semantic values from our comprehensive library of scene graph assets. This library comprises diverse objects, attributes, and relationships constructed using an LLM with implementation details provided in Appendix B and specific prompts detailed in Appendix I.1.

Upon instantiation of the scene graph, we generate the corresponding input text T for the T2I model and formulate an appropriate reward function for the subsequent training phase (as exemplified in Figure 2). The reward function R takes as input the binary question-answer pairs and the generated images, then outputs a scalar reward value $r \in \mathcal{R}$ to guide the compositional learning process.

Generation of Input Text T. 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 $\text{Diff}(\mathcal{D}) \in [\text{Diff}_{\text{min}}, \text{Diff}_{\text{max}}]$, we employ constrained LLM based generation with strict constraints to produce descriptions that exactly match the graph's specifications. The generation process integrates the linguistic capabilities of LLMs with systematic validation mechanisms to maintain strict fidelity to the source graph. This is achieved through three core components: (i) mandatory inclusion constraints that ensure exact matching of all objects and attributes, (ii) structural validation that preserves relational dependencies, and (iii) multi-stage content verification that maintains semantic integrity. This approach guarantees that the generated text prompt comprehensively reflects all elements of the original scene graph while preserving its compositional difficulty. In our implementation, we utilize Deepseek-V3 model [37] as the underlying LLM, with detailed prompts provided in Appendix I.2.

Generation of Curriculum. Given input text prompts T, any T2I model generates corresponding images Is. We evaluate the compositional accuracy of these generated images through structured question-answer pairs that are systematically constructed from the sampled scene graphs. Specifically, we adopt a programmatic approach [22] to generate precise and comprehensive binary questions that exhaustively cover all structural elements of the scene graphs. Since each scene graph $\mathcal{G} = (\mathcal{O}, \mathcal{A}, \mathcal{R})$ encodes objects, attributes, and relations, with object cardinality representing an additional critical dimension, we design four complementary question categories to achieve comprehensive coverage: (i) object verification questions (Q_{object}) to assess the presence of each object $o \in \mathcal{O}$; (ii) count verification questions (Q_{count}) to verify the correct number of occurrences for repeated objects in $\mathcal{O}$; (iii) attribute validation questions ($Q_{\text{attribute}}$) to confirm the presence of each attribute $a \in \mathcal{A}$ for its corresponding objects; (iv) relation confirmation questions (Q_{relation}) to verify the existence of each relation $r \in \mathcal{R}$ between specific object pairs.

Generation of Reward $r \in \mathcal{R}$. Building upon the binary question-answer pairs constructed above and inspired by VQAScore [36], we develop an automated reward mechanism for T2I model training. Specifically, we leverage a multimodal LLM (MLLM) to evaluate image-text alignment by computing the predicted probability of answering "yes" to each binary question as a 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. In our implementation, we adopt LLaVA-v1.6-13B [38] as the reward model due to its robust performance on visual question answering tasks. To validate the generalizability of our approach, we conduct comprehensive ablation studies with alternative MLLMs in Section 5.3.

Model Training. During training, we integrate Curriculum Learning with Group Relative Policy Optimization (GRPO) to progressively enhance the T2I model's compositional generation ability. The core idea of Curriculum-based GRPO (C-GRPO) is to dynamically adjust the emphasis on different compositional difficulty levels throughout training, enabling the model to first master simpler concepts before tackling more complex compositions. For each generated image $I^{(i)}$, we evaluate its compositional quality using binary question-answer pairs through our reward model as $r_j^{(i)} = p_{\text{reward}}(\text{answer}_j | I^{(i)}, \text{question}_j)$, where $r_j^{(i)} \in [0, 1]$ represents the reward model's confidence that image $I^{(i)}$ correctly answers the j-th binary question corresponding to a specific compositional difficulty level.

Following Parashar et al. [48], we employ curriculum scheduling strategies to weight these rewards based on training progress. We explore two scheduling approaches (as detailed in Appendix C), namely Easy-to-Hard scheduling and Gaussian scheduling. The curriculum-weighted reward at training step t is computed as $\widehat{r}_j^{(i)}(t) = \sum_{j'=1}^{||\text{Diff}||} \widehat{p}(t, j') \cdot r_j^{(i)}$, where $||\text{Diff}||$ denotes the number of compositional difficulty levels, and $\widehat{p}(t, j')$ represents the curriculum sampling probability for difficulty level j' at step t. We derive the detailed formulation for $\widehat{p}(t, j')$ in Appendix C. The overall reward for image $I^{(i)}$ is computed as the average score across all the sampled questions: $\widehat{r}^{(i)}(t) = \frac{1}{M} \sum_{j=1}^{M} \widehat{r}_j^{(i)}(t)$, where M is the number of samples. This reward design ensures that at each training stage, the model focuses on the appropriate difficulty level while gradually progressing toward more complex compositions.

The curriculum-aware advantages at step t are normalized within each group of G images as:

$$A_i(t) = \frac{\widehat{r}^{(i)}(t) - \text{Mean}(\{\widehat{r}^{(k)}(t)\}_{k=1}^{G})}{\text{Std}(\{\widehat{r}^{(k)}(t)\}_{k=1}^{G})}, \tag{4}$$

where $\text{Mean}(\{\widehat{r}^{(k)}(t)\}_{k=1}^{G})$ and $\text{Std}(\{\widehat{r}^{(k)}(t)\}_{k=1}^{G})$ denote

Table 1. Comparison of different T2I models on compositional generation benchmarks. Models with gray background are the baseline models that our method builds upon, while those with yellow background are our trained models. The best performance among all models are marked in **red** and the performance improvement of our trained models over the baseline models are marked in **↑green**.

Model	# Params	GenEval	DPG	TIFA	T2I-CompBench	DSG	Avg.
Diffusion							
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%
Lumina-Next	1.7B	46.00%	75.66%	79.98%	34.57%	70.61%	61.36%
AutoRegressive							
LlamaGen	0.8B	31.28%	42.92%	75.03%	33.26%	58.30%	48.16%
Show-o	1.3B	56.00%	67.27%	**86.28%**	29.00%	77.00%	63.11%
SimpleAR-SFT	0.5B	53.00%	78.48%	81.06%	33.76%	71.98%	63.66%
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%
Ours							
Stable-Diffusion-1.5 w/ ours	0.9B	53.88% (↑**11.80%**)	78.67% (↑**16.43%**)	85.71% (↑**7.04%**)	37.68% (↑**7.74%**)	77.16% (↑**15.59%**)	66.62% (↑**11.72%**)
SimpleAR w/ ours	0.5B	**63.24%** (↑**10.24%**)	**81.20%** (↑**2.72%**)	85.53% (↑**4.47%**)	**40.27%** (↑**6.51%**)	**86.11%** (↑**14.13%**)	**71.27%** (↑**7.61%**)

the mean and standard deviation of rewards within the group.

For each sampled input text prompt T, C-GRPO generates G distinct images $\{I^{(1)}, I^{(2)}, \ldots, I^{(G)}\}$ using the current policy $p_{\theta_{\text{old}}}$. The policy at step t is optimized by maximizing:

$$\mathcal{J}_{\text{C-GRPO}}(\theta) = \mathbb{E}_T \Bigg[\frac{1}{G} \sum_{i=1}^{G} \Bigg(\min \Bigg(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}} A_i(t), \text{clip}\Big(\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}, 1-\epsilon, 1+\epsilon \Big) A_i(t) \Bigg) - \beta \, \text{KL}\Big(p_\theta(\cdot|T) \Big\| p_{\text{ref}}(\cdot|T) \Big) \Bigg) \Bigg] \tag{5}$$

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.

5. Experiment

5.1. Experimental Setup

Benchmark Datasets and Metrics. To thoroughly assess the compositional generation capability of models trained with our CompGen framework, we evaluate on the following five compositional T2I benchmarks: (i) Geneval [24], (ii) T2I-CompBench [28], (iii) TIFA [27], (iii) DPG-Bench [26], (iv) DSG [12]. Specifically, for T2I-CompBench, we report the performance on its complex compositions task, while for the remaining benchmarks, we report the average performance across their respective subtasks. Further details on the datasets and metrics are available in Appendix D.

Baseline Models. To validate the effectiveness of our proposed CompGen method, we conduct comprehensive comparisons against several state-of-the-art T2I generation models. These include diffusion-based T2I models such as Stable Diffusion 1.4/1.5/2.1 [54], Playground v2 [32], Stable Diffusion XL [49] and LUMINA-NEXT [72]; as well as auto-regressive T2I models such as Show-o [68], Emu3 [60], SimpeAR [63], LLamaGen [59] and minDALL-E [29]. We provide detailed descriptions for each baseline in Appendix E.

We provide the implementation details of CompGen in Appendix F.

5.2. Performance Comparisons and Analysis

CompGen significantly improves T2I compositional generation capabilities across both diffusion and auto-regressive architectures. As shown in Table 1, CompGen delivers substantial improvements on compositional generation benchmarks. Applied to Stable-Diffusion-1.5 (0.9B parameters), it achieves +11.72 percentage points average improvement, surpassing both the stronger Stable-Diffusion-2.1 (59.51%) and the larger Playground-V2 (66.08%, 2.6B parameters). On auto-regressive models, CompGen demonstrates similar versatility. Applied to SimpleAR, it improves performance from 63.66% to 71.27% (+7.61 points), establishing a new state-of-the-art that outperforms all evaluated models, including the 14B-parameter Emu3 (62.31%), with

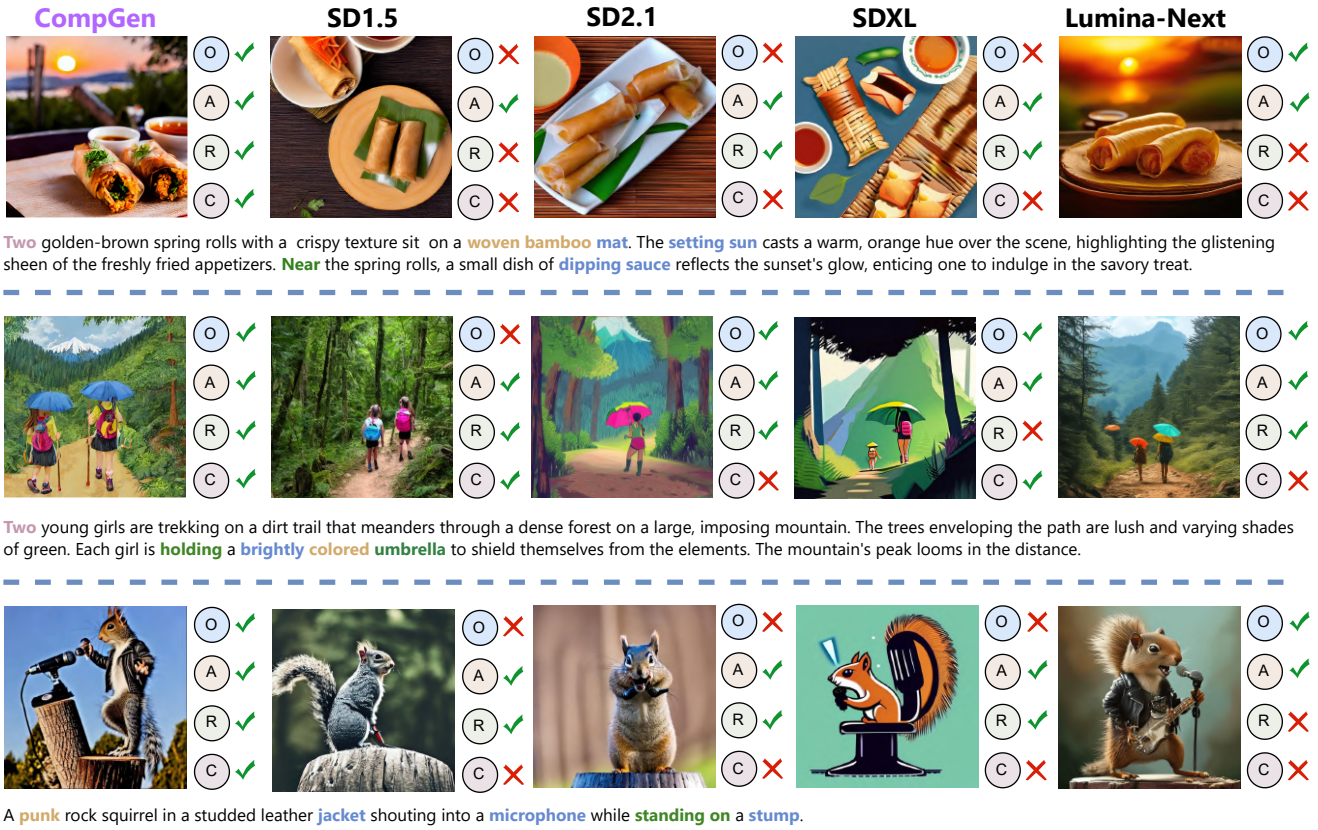

Figure 4. Qualitative comparison of our CompGen with other strong text-to-image generation models (SD1.5, SD2.1, SDXL, and Lumina-Next). Within each prompt, we color the elements for which at least one model makes an error: the object in **blue**, the attribute in **brown**, the relationship in **green**, and the count in **purple**. Ⓞ, Ⓐ, Ⓡ, Ⓒ denote Object, Attribute, Relationship, and Count, respectively. A ✓ indicates correct generation, while a ✗ indicates an error. Additional examples appear in Appendix J.

Table 2. Investigation on reward models using Stable-Diffusion-1.5 as backbone. Our adopted model is highlighted in yellow. The best performance is marked in **red**.

Model	Reward Model	GenEval	DPG	TIFA	T2I-CompBench	DSG	Avg.
Stable-Diffusion-1.5	–	42.08%	62.24%	78.67%	29.94%	61.57%	54.90%
Stable-Diffusion-1.5 w/ VQAScore [36]	LLaVA-v1.6-13B	44.02%	73.41%	80.19%	36.36%	73.23%	61.44%
CompGen (Stable-Diffusion-1.5)	CLIP-FlanT5-XXL	45.11%	71.26%	81.42%	31.60%	73.74%	60.63%
	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%
	LLaVA-v1.6-13B	**53.88%**	**78.67%**	**85.71%**	37.68%	**77.16%**	**66.62%**

best performance across most individual benchmarks.

CompGen improves compositionality without sacrificing visual quality or overfitting. Figure 4 demonstrates CompGen's superior compositional capabilities. While strong baselines (SDXL, Lumina-Next) frequently misinterpret object counts (e.g., "two spring rolls"), fail to bind attributes correctly (e.g., a squirrel "in a studded leather jacket"), or miss complex relationships (e.g., "shouting into a microphone"), CompGen accurately renders these details while maintaining visual quality without artifacts.

Additional examples appear in Appendix J.

5.3. Impact of Reward Model

Here, we investigate the impact of reward model capability on CompGen's performance by comparing four vision-language models as reward models: LLaVA-v1.6-13B [38], LLaVA-v1.5-13B [38], CLIP-FlanT5-XXL [50], and InstructBLIP-FlanT5-XXL [14]. Table 2 shows a strong positive correlation between reward model capability and CompGen performance. The strongest model (LLaVA-v1.6-13B) achieves 66.62% average score, outperforming the weakest (InstructBLIP) by 9.4 percentage points (57.22%). This scaling behavior indicates that CompGen's effective-

Table 3. Investigation on different *difficulty measures* using Stable-Diffusion-1.5 with 10K training data. Our proposed difficulty measure is marked with yellow background. The best performance for each benchmark is marked in **red**. $\|\mathcal{O}\|$ is the number of objects in the scene graph, $\|\mathcal{A}\|$ is the number of attributes and $\|\mathcal{R}\|$ is the number of relations.

Difficulty Measure	GenEval	DPG	TIFA	T2I-CompBench	DSG
$(\|\|\mathcal{O}\|\| + \|\|\mathcal{A}\|\| + \|\|\mathcal{R}\|\|)$ [22]	50.12%	72.59%	79.40%	37.00%	71.21%
$(\|\|\mathcal{O}\|\| + \|\|\mathcal{R}\|\|)/2$ [10]	48.83%	72.91%	75.24%	35.59%	74.33%
Ours	**53.88%**	**78.67%**	**85.71%**	**37.68%**	**77.16%**

ness improves directly with vision-language model advancements, pointing to a clear path for future improvements via better reward models. We offer a visualization in Figure 7.

We also provide additional reward function analysis in Appendix G.1. Our findings indicate that a fine-grained, multi-aspect reward function provides more effective guidance for complex compositional generation than a coarse-grained reward function.

5.4. Impact of Difficulty Measures

We evaluate the effectiveness of different difficulty measures in Table 3 by training Stable-Diffusion-1.5 on 10K prompts uniformly sampled across difficulty levels 1-10. Existing methods [10, 22] typically employ additive metrics that sum or average compositional components. In contrast, our proposed difficulty measure uses a multiplicative formulation to better capture the combinatorial explosion in compositional complexity as the number of components increases. Results demonstrate that curriculum learning guided by our difficulty measure achieves superior performance across all benchmarks, yielding an average score of 66.62% — a 4.56 percentage point improvement over the strongest additive baseline. This substantial gain underscores the importance of accurately modeling the exponential nature of compositional T2I difficulty.

We further investigate the impact of data difficulty distribution in Appendix G.2. Our results show that a balanced training curriculum covering a wide range of data difficulties is crucial for robust RL training, as training exclusively on skew-easy or skew-hard samples [57] leads to poor compositional generalization.

5.5. Analysis of Curriculum Scheduling Strategy

We investigate how curriculum scheduling strategies impact model performance and scaling behavior. Using Stable-Diffusion-1.5 on the GenEval [24] benchmark, we compare three scheduling strategies [48]: (i) **Random Scheduling**, uniform sampling across all difficulty levels as a baseline; (ii) **Easy-to-Hard Scheduling**, a deterministic sequential progression from easiest to hardest samples; and (iii) **Gaussian Scheduling**, a smooth transition where sampling follows a

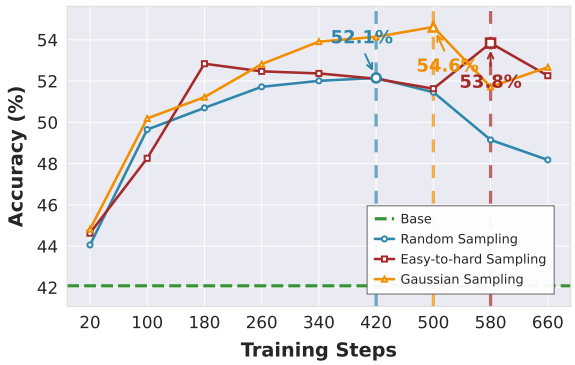

Figure 5. Scaling trend of CompGen with different curriculum scheduling strategies.

bell-shaped curve that gradually shifts from easy to hard data. Models are trained with batch size 32 for up to 660 steps. Further implementation details appear in Appendix C. Figure 5 demonstrates that curriculum strategies substantially improve CompGen performance. While the baseline model without curriculum learning stagnates at 42% accuracy, all curriculum approaches achieve significant gains. Random sampling provides strong improvement, peaking at 52.1% accuracy. The structured curriculum strategies yield even larger gains: Gaussian scheduling achieves the highest accuracy of 54.6% at 500 steps — a 30% relative improvement (12.6 percentage points) over the baseline. Beyond absolute performance gains, curriculum learning extends the scaling regime compared to conventional training. While the baseline plateaus early, curriculum strategies continue improving beyond 500 training steps before gradually declining.

Notably, Gaussian scheduling demonstrates the most efficient scaling trajectory, reaching optimal performance with fewer steps, while easy-to-hard scheduling exhibits the most extended scaling range, maintaining competitive performance the longest during prolonged training.

6. Conclusion

In this work, we introduced CompGen, a novel compositional curriculum reinforce learning framework that leverages scene graphs and incorporates a principled difficulty criterion with an adaptive Markov Chain Monte Carlo sampling algorithm. Our approach free from the requirements of need of ground-truth images, systematically improves the ability of T2I models to generate complex compositional 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.

References

[1] Jingkun An, Yinghao Zhu, Zongjian Li, Enshen Zhou, Haoran Feng, Xijie Huang, Bohua Chen, Yemin Shi, and Chengwei Pan. Agfsync: Leveraging ai-generated feedback for preference optimization in text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1746–1754, 2025. 1

[2] 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

[3] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 3

[4] Antonio Blanca, Sarah Cannon, and Will Perkins. Fast and perfect sampling of subgraphs and polymer systems. *ACM Transactions on Algorithms*, 20(1):1–30, 2024. 7

[5] 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

[6] Luiz FO Chamon and Alejandro Ribeiro. Greedy sampling of graph signals. *IEEE Transactions on Signal Processing*, 66(1):34–47, 2017. 7

[7] 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

[8] 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, 3

[9] 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*, pages 5343–5353, 2024. 1, 3

[10] Zuyao Chen, Jinlin Wu, Zhen Lei, and Chang Wen Chen. What makes a scene? scene graph-based evaluation and feedback for controllable generation. *arXiv preprint arXiv:2411.15435*, 2024. 8

[11] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49 (4):327–335, 1995. 4

[12] 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, 3

[13] 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*, pages 432–448. Springer, 2024. 1, 3

[14] 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

[15] DeepSeek-AI. Deepseek-v3 technical report, 2024. 1

[16] 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

[17] 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

[18] 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. 3

[19] Ao Fu, Ziqi Ni, and Yi Zhou. Dual audio-centric modality coupling for talking head generation. *arXiv preprint arXiv:2503.22728*, 2025. 1

[20] Yu-Hsiang Fu, Chung-Yuan Huang, and Chuen-Tsai Sun. Using global diversity and local topology features to identify influential network spreaders. *Physica A: Statistical Mechanics and its Applications*, 433:344–355, 2015. 7

[21] 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

[22] 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, 8

[23] Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992. 2

[24] 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, 8, 3

[25] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528, 2021. 3

[26] 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, 3

[27] 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*, pages 20406–20417, 2023. 2, 6, 3

[28] 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, 3, 6

[29] Chiheon Kim Doyup Lee Saehoon Kim, Sanghun Cho, and Woonhyuk Baek. mindall-e on conceptual captions, 2021. 6, 5

[30] 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*, pages 7701–7711, 2023. 1, 3

[31] 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. 2, 3

[32] Daiqing Li, Aleks Kamko, Ali Sabet, Ehsan Akhgari, Linmiao Xu, and Suhail Doshi. Playground v2. 6, 4

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[34] 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

[35] Yuanzhi Liang, Yijie Fang, Rui Li, Ziqi Ni, Ruijie Su, and Chi Zhang. Integrating reinforcement learning with visual generative models: foundations and advances. *Vicinagearth*, 3(1):2, 2026. 1

[36] 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*, pages 366–384. Springer, 2024. 5, 7, 6

[37] 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

[38] 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

[39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 6

[40] 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*, pages 423–439. Springer, 2022. 1

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[42] 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

[43] 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*, pages 9005–9014, 2024. 1, 3

[44] Ziqi Ni, Ao Fu, and Yi Zhou. Freak: Frequency-modulated high-fidelity and real-time audio-driven talking portrait synthesis. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 1036–1044, 2025. 1

[45] Ziqi Ni, Yuanzhi Liang, Rui Li, Yi Zhou, Haibin Huang, Chi Zhang, and Xuelong Li. Seeing what matters: Visual preference policy optimization for visual generation. *arXiv preprint arXiv:2511.18719*, 2025. 1

[46] 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

[47] 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. 3

[48] 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. 5, 8

[49] 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, 4

[50] 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*, pages 8748–8763. PmLR, 2021. 7

[51] 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*, pages 8821–8831. Pmlr, 2021. 2

[52] 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

[53] 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, 3

[54] 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*, pages 10684–10695, 2022. 2, 6, 4

[55] 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

[56] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, and Y. Wu. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. 2024. 5

[57] Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. Efficient reinforcement finetuning via adaptive curriculum learning. *arXiv preprint arXiv:2504.05520*, 2025. 8

[58] 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*, 2023. 1

[59] 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, 5

[60] Quan Sun, Qiying 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*, 2023. 6, 5

[61] 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, 2024. 2

[62] 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*, pages 153–168. Springer, 2024. 2

[63] Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025. 6, 5

[64] 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*, pages 5544–5552, 2024. 1, 3

[65] 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

[66] 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

[67] 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*, pages 6327–6336, 2024. 3

[68] 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, 5

[69] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025. 6

[70] 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

[71] 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

[72] 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, 4