ITERCOMP: ITERATIVE COMPOSITION-AWARE FEEDBACK LEARNING FROM MODEL GALLERY FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Advanced diffusion models like Stable Diffusion 3, Omost, and FLUX have made notable strides in compositional text-to-image generation. However, these methods typically exhibit distinct strengths for compositional generation, with some excelling in handling attribute binding and others in spatial relationships. This disparity highlights the need for an approach that can leverage the complementary strengths of various models to comprehensively improve the composition capability. To this end, we introduce IterComp, a novel framework that aggregates composition-aware model preferences from multiple models and employs an iterative feedback learning approach to enhance compositional generation. Specifically, we curate a gallery of six powerful open-source diffusion models and evaluate their three key compositional metrics: attribute binding, spatial relationships, and non-spatial relationships. Based on these metrics, we develop a composition-aware model preference dataset comprising numerous image-rank pairs to train composition-aware reward models. Then, we propose an iterative feedback learning method to enhance compositionality in a closed-loop manner, enabling the progressive self-refinement of both the base diffusion model and reward models over multiple iterations. Detailed theoretical proof demonstrates the effectiveness of this method. Extensive experiments demonstrate our significant superiority over previous methods, particularly in multi-category object composition and complex semantic alignment. IterComp opens new research avenues in reward feedback learning for diffusion models and compositional generation.

1 INTRODUCTION

The rapid advancement of diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020; Peebles & Xie, 2023) has recently brought unprecedented progress to the field of text-to-image generation, with powerful models like DALL-E 3 (Betker et al., 2023), Stable Diffusion 3, (Esser et al., 2024) and FLUX (BlackForest, 2024) demonstrating remarkable capabilities in generating aesthetic and diverse images. However, these models often struggle to follow complex prompts to achieve precise compositional generation (Omost-Team, 2024; Yang et al., 2024b; Zhang et al., 2024b), which requires the model to possess robust, comprehensive capabilities in various aspects, such as attribute binding, spatial relationships, and non-spatial relationships (Huang et al., 2023).

To enhance compositional generation, some works introduce additional conditions such as layouts/boxes (Li et al., 2023; Zhou et al., 2024; Wang et al., 2024a; Zhang et al., 2024b). InstanceDiffusion (Wang et al., 2024a) controls the generation process using layouts, masks, or other conditions through trainable instance masked attention layers. Although these layout-based methods demonstrate strong spatial awareness, they struggle with image realism, especially in generating non-spatial relationships and preserving aesthetic quality (Zhang et al., 2024b). Another potential solution leverages the impressive reasoning abilities of Large Language Models (LLMs) to decompose complex generation tasks into simpler subtasks (Yang et al., 2024b; Omost-Team, 2024; Wang et al., 2024b). RPG (Yang et al., 2024b) employs MLLMs as the global planner to transform the process of generating complex images into multiple simpler generation tasks within subregions. However, it requires designing complex prompts for LLMs, and it is challenging to achieve precise generation results due to their intricate outputs (Yang et al., 2024b).

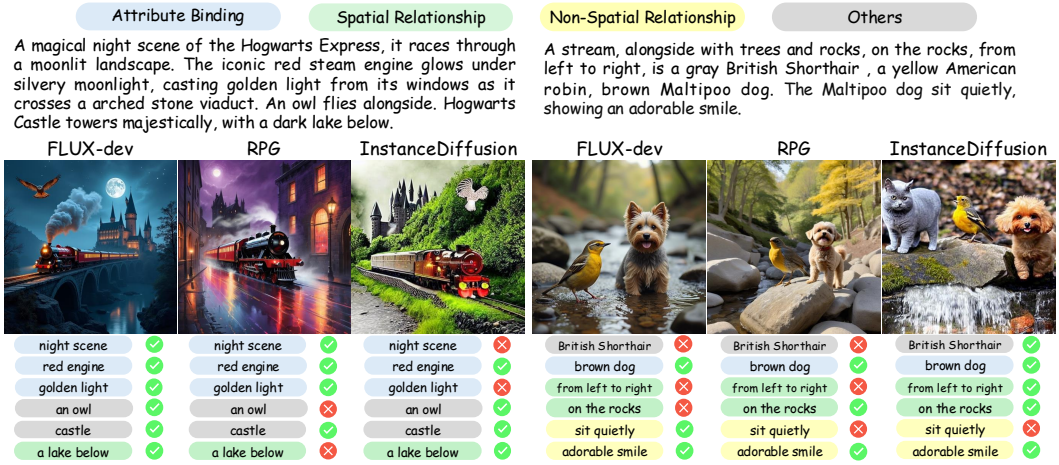


Figure 1: **Motivation of IterComp.** We select three types of compositional generation methods. The results show that different models exhibit distinct strengths across various aspects of compositional generation. fig. 3 further demonstrated these distinct strengths quantitatively.

We conducted extensive experiments to explore the unique strengths of different models in compositional generation. As shown in the left example in fig. 1, text-to-image model FLUX (BlackForest, 2024) demonstrates impressive performance in attribute binding and aesthetic quality due to its advanced training techniques and model architecture. In contrast, layout-to-image model InstanceDiffusion (Wang et al., 2024a) struggles to capture fine-grained visual details, such as 'night scene' or 'golden light.' In the right example of fig. 1, where the text prompt involves complex spatial relationships between multiple objects, FLUX (BlackForest, 2024) exhibits limitations in spatial awareness. In contrast, InstanceDiffusion (Wang et al., 2024a) excels in handling spatial relationships through layout guidance. This demonstrates that different models exhibit distinct strengths across various aspects of compositional generation. Moreover, fig. 3 further demonstrated these distinct strengths quantitatively. Naturally, a pertinent question arises: *Is there a method capable of excelling in all aspects of compositional generation?*

In order to enable the diffusion model to improve compositional generation comprehensively, we present a new framework, ***IterComp***, which collects *composition-aware model preferences* from various models, and then employs a novel yet simple *iterative feedback learning* framework to achieve comprehensive improvements in compositional generation. Firstly, we select six open-sourced models excelling in different aspects of compositionality to form our model gallery. We focus on three essential compositional metrics: attribute binding, spatial relationships, and non-spatial relationships to curate a new composition-aware model preference dataset, which consists of a large number of image-rank pairs. Next, to comprehensively capture diverse composition-aware model preferences, we train reward models to provide fine-grained compositional guidance during the finetuning of the base diffusion model. Finally, given that compositional generation is difficult to optimize, we propose iterative feedback learning. This approach enhances compositionality in a closed-loop manner, allowing for the progressive self-refinement of both the base diffusion model and reward models in multiple iterations. We theoretically and experimentally demonstrate the effectiveness of our method and its significant improvement in compositional generation.

Our contributions are summarized as follows:

- We propose the first iterative composition-aware reward-controlled framework *IterComp*, to comprehensively enhance the compositionality of the base diffusion model.
- We curate a model gallery and develop a high-quality composition-aware model preference dataset comprising numerous image-rank pairs.
- We utilize a new *iterative feedback learning* framework to progressively enhance both the reward models and the base diffusion model.
- Extensive qualitative and quantitative comparisons with previous SOTA methods demonstrate the superior compositional generation capabilities of our approach.

2 RELATED WORK

Compositional Text-to-Image Generation Compositional text-to-image generation is a complex and challenging task that requires a model with comprehensive capabilities, including the understanding of complex prompts and spatial awareness (Yang et al., 2024b; Zhang et al., 2024b). Some methods enhance prompt comprehension by using more powerful text encoders or architectures (Esser et al., 2024; Betker et al., 2023; Hu et al., 2024; Dai et al., 2023). Stable Diffusion 3 (Esser et al., 2024) utilizes three different-sized text encoders to enhance prompt comprehension. DALL-E 3 (Betker et al., 2023) enhances the understanding of rich textual details by expanding image captions through recaptioning. However, compositional capability such as spatial awareness remains a limitation of these models (Li et al., 2023; Chen et al., 2024a). Other methods attempt to enhance spatial awareness by the control of additional conditions (e.g., layouts) (Yang et al., 2023; Dahary et al., 2024). BoxDiff (Xie et al., 2023) and LMD (Lian et al., 2023b) guide the generated objects to strictly adhere to the layout by designing energy functions based on cross-attention maps. ControlNet (Zhang et al., 2023) and T2I-Adapter (Mou et al., 2024) specify high-level image features to control semantic structures. Although these methods enhance spatial awareness, they often compromise image realism (Zhang et al., 2024b). Additionally, some approaches leverage the powerful reasoning capabilities of LLMs to assist in the generation process (Yang et al., 2024b; Omost-Team, 2024; Wang et al., 2024b). RPG (Yang et al., 2024b) employs MLLM to decompose complex compositional generation tasks into simpler subtasks. However, these methods require designing complex prompts as inputs to the LLM, and the diffusion model struggles to produce precise results due to the LLM's intricate outputs (Yang et al., 2024b). In contrast, our method extracts these preferences from different models in model gallery and trains composition-aware reward models to refine the base diffusion model iteratively, achieving robust compositionality across multiple aspects.

Diffusion Model Alignment Building on the success of reinforcement learning from human feedback (RLHF) in Large Language Models (LLMs) (Ouyang et al., 2022; Bai et al., 2022), numerous methods in diffusion models have attempted to use similar approaches for model alignment (Lee et al., 2023; Fan et al., 2024; Sun et al., 2023). Some methods use a pretrained reward model or train a new one to guide the generation process(Zhang et al., 2024a; Black et al., 2023; Deng et al., 2024; Clark et al., 2023; Prabhudesai et al., 2023). For instance, ImageReward (Xu et al., 2024) manually annotated a large dataset of human-preferred images and trained a reward model to assess the alignment between images and human preferences. Reward Feedback Learning (ReFL) is proposed for tuning diffusion models with the ImageReward model. RAHF (Liang et al., 2024a) is trained on RichHF-18K, a high-quality dataset rich in human feedback, and is capable of predicting the unreasonable parts in generated images. Some methods bypass the training of a reward model and directly finetune diffusion models on human preference datasets (Yang et al., 2024a; Liang et al., 2024b; Yang et al., 2024c). Diffusion-DPO (Wallace et al., 2024) reformulates Direct Preference Optimization (DPO) to account for a diffusion model's notion of likelihood, utilizing the evidence lower bound to derive a differentiable objective. The potential for alignment in diffusion models goes beyond this. We iteratively align the base model with composition-aware model preferences from the model gallery, effectively enhancing its performance on compositional generation.

3 METHOD

In this section, we present our method, IterComp, which collects composition-aware model preferences from the model gallery and utilizes iterative feedback learning to enhance the comprehensive capability of the base diffusion model in compositional generation. An overview of IterComp is illustrated in fig. 2. In section 3.1, we introduce the method for collecting the composition-aware model preference dataset from the model gallery. In section 3.2, we describe the training process for the composition-aware reward models and multi-reward feedback learning. In section 3.2, we propose the iterative feedback learning framework to enable the self-refinement of both the base diffusion model and reward models, progressively enhancing compositional generation.

3.1 COLLECTING HUMAN PREFERENCES OF COMPOSITIONALITY

Compositional Metric and Model Gallery We focus on three key aspects of compositionality: attribute binding, spatial relationships, and non-spatial relationships (Huang et al., 2023), to collect

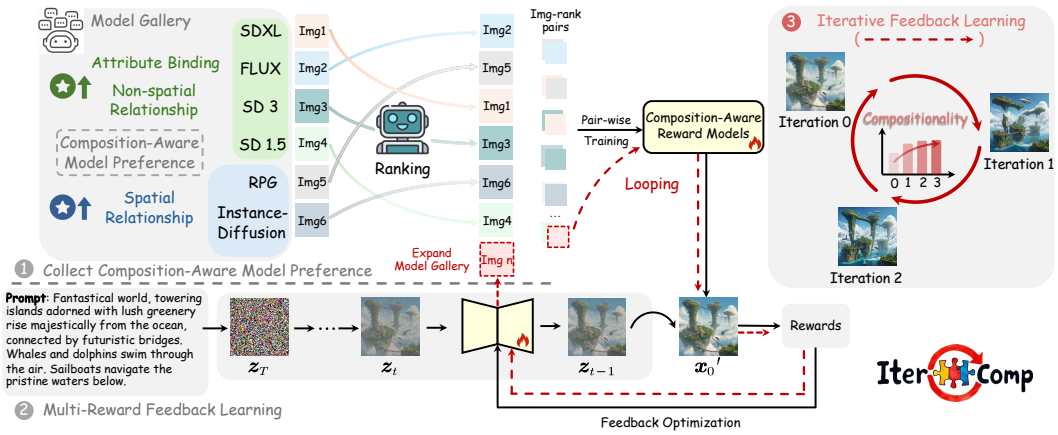


Figure 2: Overview of IterComp. We collect composition-aware model preferences from multiple models and employ an iterative feedback learning approach to enable the progressive self-refinement of both the base diffusion model and reward models.

composition-aware model preferences. We initially select six open-sourced models excel in different aspects of compositional generation as our model gallery: FLUX-dev (BlackForest, 2024), Stable Diffusion 3 (Esser et al., 2024), SDXL (Podell et al., 2023), Stable Diffusion 1.5 (Rombach et al., 2022), RPG (Yang et al., 2024b), and InstanceDiffusion (Wang et al., 2024a).

Human Ranking on Attribute Binding For attribute binding, we randomly select 500 prompts from each of the following categories: color, shape, and texture in the T2I-CompBench (Huang et al., 2023), resulting in a total of 1,500 prompts. Three professional experts ranked the images generated by the six models for each prompt, and their rankings were weighted to determine the final result. The primary criterion is whether the attributes mentioned in the prompt were accurately reflected in the generated images, especially the correct representation and binding of attributes to the corresponding objects.

Human Ranking on Complex Relationships For spatial and non-spatial relationships, we select 1,000 prompts for each category from the T2I-CompBench (Huang et al., 2023) and apply the same manual annotation method to obtain the rankings. For spatial relationships, the primary ranking criterion is whether the objects are correctly generated and whether their spatial positioning matches the prompt. For non-spatial relationships, the focus is on whether the objects display natural and realistic actions.

Analysis of Composition-aware Model Preference Dataset For each prompt, we obtain 6 images and $\binom{6}{2} = 15$ image-rank pairs. As shown in table 1, in total, we collected a dataset with 22,500 image-rank pairs for model preference in attribute binding, 15,000 for spatial relationships, and 15,000 for non-spatial relationships. We visualize the proportion of generated images ranked first for each model in fig. 3. The results demonstrate that different models exhibit distinct strengths across various aspects of compositional generation, and this dataset effectively captures a diverse range of composition-aware model preferences.

3.2 Composition-aware Multi-Reward Feedback Learning

Composition-aware Reward Model Training To achieve comprehensive improvements in compositional generation, we utilize three types of composition-aware datasets described in section 3.1, decomposing compositionality into three subtasks and training a specific reward model for each. Specifically, the reward model $\mathcal{R}_{\theta_i}(\boldsymbol{c}, \boldsymbol{x}_0)$ is trained using the input format $\boldsymbol{x}_0^w \succ \boldsymbol{x}_0^l \mid \boldsymbol{c}$, where $\boldsymbol{x}_0^w$ and $\boldsymbol{x}_0^l$ denoting the "winning" and "losing" images, $\boldsymbol{c}$ denoting the text prompt. We select two images corresponding to the same prompt from the composition-aware model preference datasets to form an input image-rank pair, and trained the reward model using the following loss function:

$$\mathcal{L}(\theta_i) = -\mathbb{E}_{(\boldsymbol{c},\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim\mathcal{D}_i}\left[\log\left(\sigma\left(\mathcal{R}_{\theta_i}\left(\boldsymbol{c},\boldsymbol{x}_0^w\right) - \mathcal{R}_{\theta_i}\left(\boldsymbol{c},\boldsymbol{x}_0^l\right)\right)\right)\right] \tag{1}$$

Table 1: Statistics on the composition-aware model preference dataset. The dataset consists of 3,500 text prompts, 27,500 images, and 52,500 image-rank pairs.

Category	Counts		
	Texts	Images	Image-rank pairs
Attribute Binding	1,500	9,000	22,500
Spatial Relationship	1,000	6,000	15,000
Non-spatial Relationship	1,000	6,000	15,000
Total	3,500	21,000	52,500

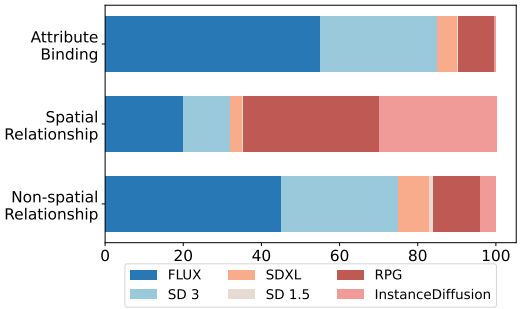


Figure 3: The proportion of each model ranked first.

where $\mathcal{D}$ denotes the composition-aware model preference dataset, $\sigma(\cdot)$ is the sigmoid function.

The three composition-aware reward models apply BLIP (Li et al., 2022; Xu et al., 2024) as feature extractors. We combine the extracted image and text features with cross attention mechanism, and use a learnable MLP to generate a score scalar for preference comparison.

Multi-Reward Feedback Learning Due to the multi-step denoising process in diffusion models, yielding likelihoods for their generations is impossible, making the RLHF approach used in language models unsuitable for diffusion models. Some existing methods (Xu et al., 2024; Zhang et al., 2024a) finetune diffusion models directly by treating the scores of the reward model as the human preference loss. To optimize the base diffusion model using multiple composition-aware reward models, we design the loss function as follows:

$$\mathcal{L}(\theta) = \lambda \mathbb{E}_{\boldsymbol{c}_j \sim \mathcal{C}} \sum_i \left(\phi\left(\mathcal{R}_i\left(\boldsymbol{c}_j, p_\theta\left(\boldsymbol{c}_j\right)\right)\right)\right) \tag{2}$$

where $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n\}$ denotes the prompt set, $p_\theta(\boldsymbol{c})$ denotes the generate image of diffusion model with parameter θ under the condition of prompt $\boldsymbol{c}$. We calculate the loss for each reward model $\mathcal{R}_i(\cdot)$ and sum them to obtain the multi-reward feedback loss.

3.3 Iterative Optimization of Composition-aware Feedback Learning

Compositional generation is challenging to optimize due to its inherent complexity and multifaceted nature, requiring both our reward models and base diffusion model to excel in aspects such as complex text comprehension and the generation of complex relationships. To ensure more thorough optimization, we propose an iterative feedback learning framework that progressively refines both the reward models and the base diffusion model over multiple iterations.

At the $(k+1)$-th iteration of the optimization described in section 3.2, we denote the reward models and the base diffusion model from the previous iteration as $\mathcal{R}^k(\cdot)$ and $p_\theta^k(\cdot)$, respectively. For each prompt $\boldsymbol{c}$ in the datasets $\mathcal{D}^k$, we sample an image $\boldsymbol{x}_0^* = p_\theta^k(\boldsymbol{c})$ and expand the composition-aware model preference dataset $\mathcal{D}^k$ with the sampled image. The image rankings for each prompt are updated using the trained reward model $\mathcal{R}_\theta^k(\cdot)$, while preserving the relative ranks of the initial six images. Following this process, we update the composition-aware model preference dataset to a more comprehensive version, denoted as $\mathcal{D}^{k+1}$. Using this dataset, we finetune both the reward models and the base diffusion model to get $\mathcal{R}^{k+1}(\cdot)$ and $p_\theta^{k+1}(\cdot)$. The detailed process of iterative feedback learning can be found in algorithm 1.

Effectiveness of Iterative Feedback Learning Through this iterative feedback learning framework, the reward models become more effective at understanding complex compositional prompts, providing more comprehensive guidance to the base diffusion model for compositional generation. The optimization objective of the iterative feedback learning process is formalized in the following lemma (proof provided in the appendix A.2):

Lemma 1. *The unified optimization framework of iterative feedback learning can be formulated as:*

$$\max_\theta \; J(\theta) = \mathbb{E}_{\left[\boldsymbol{c} \sim \mathcal{C}, (\boldsymbol{x}_0^w, \boldsymbol{x}_0^l) \sim p_\theta^*(\cdot|\boldsymbol{c})\right]} \left[\log \sigma \left(\beta \log \frac{p_\theta^*\left(\boldsymbol{x}_{0:T}^w \mid \boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{0:T}^w \mid \boldsymbol{c}\right)} - \beta \log \frac{p_\theta^*\left(\boldsymbol{x}_{0:T}^l \mid \boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}_{0:T}^l \mid \boldsymbol{c}\right)}\right)\right] \tag{3}$$

Algorithm 1 Iterative Composition-aware Feedback Learning

Dataset: Composition-aware model preference dataset $\mathcal{D}_0 = \{((\boldsymbol{c}_1, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l), \ldots, (\boldsymbol{c}_n, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)\}$
Prompt set $\mathcal{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_n\}$
Input: Base model with pretrained parameters p_θ, reward model $\mathcal{R}$, reward-to-loss map function ϕ, reward re-weight scale λ, iterative optimization iterations $iter$
Initialization: Number of noise scheduler time steps T, time step range for finetuning $[T_1, T_2]$

for $k = 0, \ldots, iter$ **do**
  for $(\boldsymbol{c}_i, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l) \in \mathcal{D}_k$ **do**
    $\mathcal{L} \leftarrow \log\left(\sigma\left(\mathcal{R}_{\theta_i}^k\left(\boldsymbol{c}, \boldsymbol{x}_0^w\right) - \mathcal{R}_{\theta_i}^k\left(\boldsymbol{c}, \boldsymbol{x}_0^l\right)\right)\right)$ // Reward model loss
    $\mathcal{R}_{\theta_{i+1}}^k \leftarrow \mathcal{R}_{\theta_i}^k(\boldsymbol{c}_i, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)$ // Update the reward models
  end for // Get $\mathcal{R}^{k+1}$ after training
  for $\boldsymbol{c}_i \in \mathcal{C}$ **do**
    $t \leftarrow rand(T_1, T_2)$ // Pick a random timestep $t \in [T_1, T_2]$
    $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    for $j = T, \ldots, t+1$ **do**
      no grad: $\boldsymbol{z}_{j-1} \leftarrow p_{\theta_i}^k(\boldsymbol{z}_j)$
    end for
    with grad: $\boldsymbol{z}_{t-1} \leftarrow p_{\theta_i}^k(\boldsymbol{z}_t)$
    $\boldsymbol{x}_0 \leftarrow \text{VaeDec}(\boldsymbol{z}_0) \leftarrow \boldsymbol{z}_{t-1}$ // Predict image from the original latent
    $\mathcal{L} \leftarrow \lambda\phi(\sum_\theta \mathcal{R}_\theta^{k+1}(\boldsymbol{c}_i, \boldsymbol{x}_0))$ // Multi-reward feedback learning loss
    $p_{\theta_{i+1}}^k \leftarrow p_{\theta_i}^k$ // Update the base diffusion model
  end for // Get p^{k+1} after training
  for $(\boldsymbol{c}_i, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l) \in \mathcal{D}_k$ **do**
    $\boldsymbol{x}_0^* \leftarrow p^{k+1}(\boldsymbol{c}_i)$ // Sample images from the optimized base diffusion model
  end for
  $\mathcal{D}_{k+1} \leftarrow rank(\mathcal{D}_k \cup \boldsymbol{x}_0^*)$ // Expand the dataset and update ranking
end for

where $p^*(\cdot)$ denotes the optimized base diffusion model. We simplify the bilevel problem of iterative feedback learning into a single-level objective. Based on this, we present the following theorem regarding the gradient of this objective:

Theorem 1. *Assume that* $F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l) = \log\sigma\left(\beta\log\frac{p_\theta^*(\boldsymbol{x}_{0:T}^w|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_{0:T}^w|\boldsymbol{c})} - \beta\log\frac{p_\theta^*(\boldsymbol{x}_{0:T}^l|\boldsymbol{c})}{p_{\text{ref}}(\boldsymbol{x}_{0:T}^l|\boldsymbol{c})}\right)$*, the gradient of optimization object can be written as the sum of two terms:* $\nabla_\theta J(\theta) = T_1 + T_2$*, where:*

$$T_1 = \mathbb{E}\left[\left(\nabla_\theta \log p_\theta\left(\boldsymbol{x}_{0:T}^w \mid \boldsymbol{c}\right) + \nabla_\theta \log p_\theta\left(\boldsymbol{x}_{0:T}^l \mid \boldsymbol{c}\right)\right) F_\theta\left(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l\right)\right] \tag{4}$$

$$T_2 = \mathbb{E}_{\left[\boldsymbol{c}\sim\mathcal{C}, (\boldsymbol{x}_0^w, \boldsymbol{x}_0^l)\sim p_\theta^*(\cdot|\boldsymbol{c})\right]}\left[\nabla_\theta[F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)]\right] \tag{5}$$

It is evident that T_2 represents the gradient form of direct preference optimization. In addition, we have another term T_1, which guides the gradient of optimization objective. As shown in eq. (4), the gradient directs the generation of $\boldsymbol{x}_0^w$ and $\boldsymbol{x}_0^w$ to optimize the implicit reward function $F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)$. The gradient term T_1 helps the model better distinguish between winning and losing samples, increasing the probability of generating high-quality images while reducing the probability of generating low-quality images. This improves the model's alignment with the reward model's preferences during generation, thereby enhancing the comprehensive capabilities of compositional generation.

Superiority over Diffusion-DPO and ImageReward Here we clarify some superiorities of IterComp over Diffusion-DPO (Wallace et al., 2024) and ImageReward (Xu et al., 2024). Our IterComp first focuses on composition-aware rewards to optimize T2I models for realistic complex generation scenarios, and constructs a powerful model gallery to collect multiple composition-aware model preferences. Then our novel iterative feedback learning framework can effectively achieve progressive self-refinement of both base diffusion model and reward models over multiple iterations.

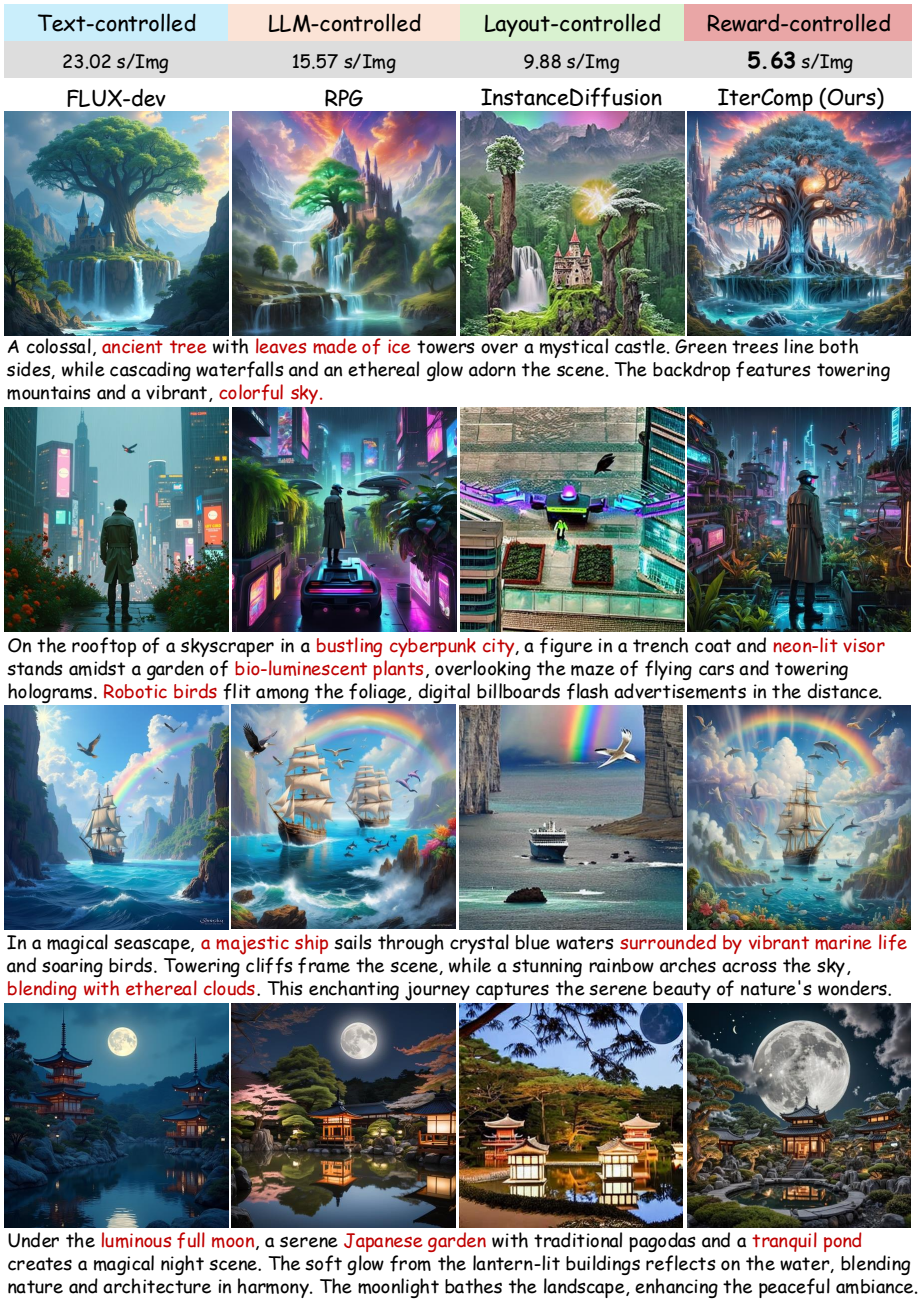


Figure 4: Qualitative comparison between our IterComp and three types of compositional generation methods: text-controlled, LLM-controlled, and layout-controlled approaches. IterComp is the first reward-controlled method for compositional generation, utilizing an iterative feedback learning framework to enhance the compositionality of generated images. Colored text denotes the advantages of IterComp in generated images.

4 Experiments

4.1 Experimental Setup

Datasets and Training Setting The reward models are trained on the composition-aware model preference dataset, consisting of 3,500 prompts and 52,500 image-rank pairs. For training the three reward models, we finetune BLIP and the learnable MLP with a learning rate of $1e-5$ and a batch size of 64. During the iterative feedback learning process, we randomly select 10,000 prompts from DiffusionDB (Wang et al., 2022) and use SDXL (Betker et al., 2023) as the base diffusion model,

Table 2: Evaluation results about compositionality on T2I-CompBench (Huang et al., 2023). IterComp consistently demonstrates the best performance regarding attribute binding, object relationships, and complex compositions. We denote the best score in blue and the second-best score in green. The baseline data is quoted from GenTron (Chen et al., 2024b).

Model	Attribute Binding			Object Relationship		Complex↑
	Color ↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	
Stable Diffusion 1.4 (Rombach et al., 2022)	0.3765	0.3576	0.4156	0.1246	0.3079	0.3080
Stable Diffusion 2 (Rombach et al., 2022)	0.5065	0.4221	0.4922	0.1342	0.3096	0.3386
Attn-Exct v2 (Chefer et al., 2023)	0.6400	0.4517	0.5963	0.1455	0.3109	0.3401
Stable Diffusion XL (Betker et al., 2023)	0.6369	0.5408	0.5637	0.2032	0.3110	0.4091
PixArt-α (Chen et al., 2023)	0.6886	0.5582	0.7044	0.2082	0.3179	0.4117
ECLIPSE (Patel et al., 2024)	0.6119	0.5429	0.6165	0.1903	0.3139	-
Dimba-G (Fei et al., 2024)	0.6921	0.5707	0.6821	0.2105	0.3298	0.4312
GenTron (Chen et al., 2024b)	0.7674	0.5700	0.7150	0.2098	0.3202	0.4167
GLIGEN (Li et al., 2023)	0.4288	0.3998	0.3904	0.2632	0.3036	0.3420
LMD+ (Lian et al., 2023a)	0.4814	0.4865	0.5699	0.2537	0.2828	0.3323
InstanceDiffusion (Wang et al., 2024a)	0.5433	0.4472	0.5293	0.2791	0.2947	0.3602
IterComp (Ours)	0.7982	0.6217	0.7683	0.3196	0.3371	0.4873

finetuning it with a learning rate of $1e-5$ and a batch size of 4. We set $T = 40$, $[T_1, T_2] = [1, 10]$, $\phi =$ ReLU, and $\lambda = 1e-3$. All experiments are conducted on 4 NVIDIA A100 GPUs.

Baseline Models We curate a model gallery of six open-source models, each excelling in different aspects of compositional generation: FLUX (BlackForest, 2024), Stable Diffusion 3 (Esser et al., 2024), SDXL (Betker et al., 2023), Stable Diffusion 1.5 (Rombach et al., 2022), RPG (Yang et al., 2024b), and InstanceDiffusion (Wang et al., 2024a). To ensure the base diffusion model thoroughly and comprehensively learns composition-aware model preferences, we progressively expand the model gallery by incorporating new models (e.g., Omost (Omost-Team, 2024), Stable Cascade (Pernias et al., 2023), PixArt-α (Chen et al., 2023)) at each iteration. For performance comparison in compositional generation, we select several state-of-the-art methods, including FLUX (BlackForest, 2024), SDXL (Betker et al., 2023), and RPG (Yang et al., 2024b) to compare with our approach. We use GPT-4o (OpenAI, 2024) for the LLM-controlled methods. Additionally, GPT-4o is also employed to infer the layout from the prompt for the layout-controlled methods.

4.2 Main Results

Qualitative Comparison As shown in fig. 4, IterComp achieves superior compositional generation results compared to the three main types of compositional generation methods: text-controlled, LLM-controlled, and layout-controlled approaches. In comparison to text-controlled methods FLUX (BlackForest, 2024), IterComp excels in handling spatial relationships, significantly reducing errors such as object omissions and inaccuracies in numeracy and positioning. When compared to LLM-controlled methods like RPG (Yang et al., 2024b), IterComp produces more reasonable object placements, avoiding the unrealistic positioning caused by LLM hallucinations. Compared to layout-controlled methods like InstanceDiffusion (Wang et al., 2024a), IterComp demonstrates a clear advantage in both semantic aesthetics and compositionality, particularly when generating under complex prompts.

Quantitative Comparison We compare IterComp with previous outstanding compositional text/layout-to-image models on the T2I-CompBench (Huang et al., 2023) in six key compositional scenarios. As shown in table 2, IterComp demonstrates a remarkable preference across all evaluation tasks. Layout-controlled methods such as LMD+ (Lian et al., 2023a) and InstanceDiffusion (Wang et al., 2024a) excel in generating accurate spatial relationships, while text-to-image models like SDXL (Betker et al., 2023) and GenTron (Chen et al., 2024b) exhibit particular strengths in attribute binding and non-spatial relationships. In contrast, IterComp achieves comprehensive improvement in compositional generation. It obtains the strengths of various models by collecting composition-aware model preferences, and employs a novel iterative feedback learning to enable self-refinement of both the base diffusion model and reward models in a closed-loop manner.

IterComp achieves a high level of compositionality while simultaneously enhancing the realism and aesthetics of the generated images. As shown in table 3, we evaluate the improvement in image realism by calculating the CLIP Score, Aesthetic Score, and ImageReward. IterComp significantly

Table 3: Evaluation on image realism.

Model	CLIP Score↑	Aesthetic Score↑	ImageReward↑
Stable Diffusion 1.4 (Rombach et al., 2022)	0.307	5.326	-0.065
Stable Diffusion 2.1 (Rombach et al., 2022)	0.321	5.458	0.216
Stable Diffusion XL (Betker et al., 2023)	0.322	5.531	0.780
GLIGEN (Li et al., 2023)	0.301	4.892	-0.077
LMD+ (Lian et al., 2023a)	0.298	4.964	-0.072
InstanceDiffusion (Wang et al., 2024a)	0.302	5.042	-0.035
IterComp (Ours)	0.337	5.936	1.437

Table 4: Evaluation on inference time.

Model	Inference Time↓
FLUX-dev	23.02 s/Img
Stable Diffusion XL (Betker et al., 2023)	5.63 s/Img
Omost (Omost-Team, 2024)	21.08 s/Img
RPG (Yang et al., 2024b)	15.57 s/Img
InstanceDiffusion (Wang et al., 2024a)	9.88 s/Img
IterComp (Ours)	5.63 s/Img

Table 5: Comparison between IterComp and other diffusion alignment methods.

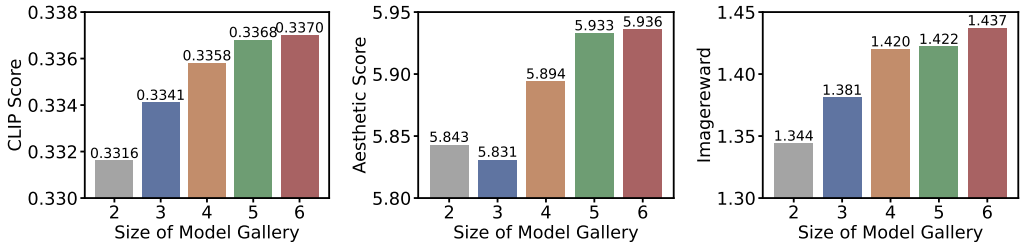
Model	Average Result on T2I-CB↑	CLIP Score↑	Aesthetic Score↑
Stable Diffusion XL (Betker et al., 2023)	0.4441	0.322	5.531
Diffusion-DPO (Wallace et al., 2024)	0.4417	0.326	5.572
ImageReward (Xu et al., 2024)	0.4639	0.323	5.613
IterComp (Ours)	0.5554	0.337	5.936

outperforms previous models across all three scenarios, demonstrating remarkable fidelity and precision in alignment with the complex text prompt. These promising results highlight the versatility of IterComp in both compositionality and fidelity. We provide more quantitative comparison results between IterComp and other diffusion alignment methods in appendix A.7.

IterComp requires less time to generate high-quality images. In table 4, we compare the inference time of IterComp with other outstanding models, such as FLUX (BlackForest, 2024), RPG (Yang et al., 2024b) in generating a single image. Using the same text prompts and fixing the denoising steps to 40, IterComp demonstrates faster generation, because it avoids the complex attention computations in RPG and Omost. Our method can incorporate composition-aware knowledge from different models without adding any computational overhead. This efficiency highlights its potential for various applications and offers a new perspective on handling complex generation tasks.

We compare IterComp with state-of-the-art diffusion alignment methods, Diffusion-DPO (Wallace et al., 2024) and ImageReward (Xu et al., 2024). As demonstrated in table 5, IterComp significantly outperforms previous diffusion alignment methods across all three scenarios. Iterative feedback learning allows models to achieve self-refinement over multiple iterations, resulting in comprehensive improvements in compositionality and realism.

4.3 Ablation Study



(a) Impact on CLIP Score. (b) Impact on Aesthetic Score. (c) Impact on ImageReward.

Figure 5: Ablation study on the model gallery size.

Effect of Model Gallery Size In the ablation study on model gallery size, as shown in fig. 5, we observe that increasing the size of the model gallery leads to improved performance for IterComp across various evaluation tasks. To leverage this finding and provide more fine-grained reward guidance, we progressively expand the model gallery over multiple iterations by incorporating the optimized base diffusion model and new models such as Omost (Omost-Team, 2024).

Effect of composition-aware iterative feedback learning We conducted an ablation study (see fig. 6) to evaluate the impact of composition-aware iterative feedback learning. The results show that this approach significantly improves both the accuracy of compositional generation and the aesthetic quality of the generated images. As the number of iterations increases, the model's preferences gradually converge. Based on this observation, we set the number of iterations to 3 in IterComp.

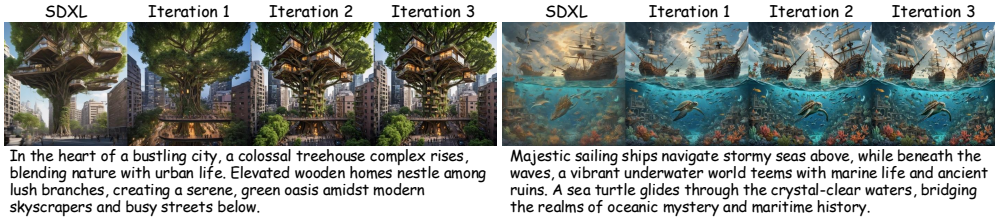


Figure 6: Ablation study on the iterations of feedback learning.

Figure 7: The generation performance of integrating IterComp into RPG and Omost.

4.4 Generalization Study

IterComp can serve as a powerful backbone for various compositional generation tasks, leveraging its strengths in spatial awareness, complex prompt comprehension, and faster inference. As shown in fig. 7, we integrate IterComp into Omost (Omost-Team, 2024) and RPG (Yang et al., 2024b). The results demonstrate that equipped with the more powerful IterComp backbone, both Omost and RPG achieve excellent compositional generation performance, highlighting IterComp's strong generalization ability and potential for broader applications.

5 Conclusion

In this paper, we propose a novel framework, IterComp, to address the challenges of complex and compositional text-to-image generation. IterComp aggregates composition-aware model preferences from a model gallery and employs an iterative feedback learning approach to progressively refine both the reward models and the base diffusion models over multiple iterations. For future work, we plan to further enhance this framework by incorporating more complex modalities as input conditions and extending it to more practical applications.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

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

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

BlackForest. Black forest labs; frontier ai lab, 2024. URL `https://blackforestlabs.ai/`.

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.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

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, 2024a.

Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, and Juan-Manuel Perez-Rua. Gentron: Diffusion transformers for image and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6441–6451, 2024b.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023.

Omer Dahary, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be yourself: Bounded attention for multi-subject text-to-image generation. *arXiv preprint arXiv:2403.16990*, 2024.

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

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

Mucong Ding, Souradip Chakraborty, Vibhu Agrawal, Zora Che, Alec Koppel, Mengdi Wang, Amrit Bedi, and Furong Huang. Sail: Self-improving efficient online alignment of large language models. *arXiv preprint arXiv:2406.15567*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

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

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, Youqiang Zhang, and Junshi Huang. Dimba: Transformer-mamba diffusion models. *arXiv preprint arXiv:2406.01159*, 2024.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

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

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*, pp. 12888–12900. PMLR, 2022.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023.

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*, 2023a.

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*, 2023b.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, et al. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024a.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024b.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4296–4304, 2024.

Omost-Team. Omost github page, 2024.

OpenAI. Hello gpt-4o, 2024. URL `https://openai.com/index/hello-gpt-4o/`.

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

Maitreya Patel, Changhoon Kim, Sheng Cheng, Chitta Baral, and Yezhou Yang. Eclipse: A resource-efficient text-to-image prior for image generations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9069–9078, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023.

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.

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

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.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023.

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

Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6232–6242, 2024a.

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

Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

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

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024a.

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

Shentao Yang, Tianqi Chen, and Mingyuan Zhou. A dense reward view on aligning text-to-image diffusion with preference. *arXiv preprint arXiv:2402.08265*, 2024c.

Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14246–14255, 2023.

Jiacheng Zhang, Jie Wu, Yuxi Ren, Xin Xia, Huafeng Kuang, Pan Xie, Jiashi Li, Xuefeng Xiao, Weilin Huang, Min Zheng, et al. Unifl: Improve stable diffusion via unified feedback learning. *arXiv preprint arXiv:2404.05595*, 2024a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

Xinchen Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Dynamic equilibrium between realism and compositionality improves text-to-image diffusion models. *arXiv preprint arXiv:2402.12908*, 2024b.

Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6818–6828, 2024.

A APPENDIX

This supplementary material is structured into several sections that provide additional details and analysis related to IterComp. Specifically, it will cover the following topics:

- In appendix A.1, we provide a preliminary about Stable Diffusion (SD) and Reward Feedback Learning (ReFL).
- In appendix A.2, we provide detailed theoretical proof of the effectiveness of iterative feedback learning.
- In appendix A.3, we conduct an experimental analysis to assess model stability.
- In appendix A.4, we provide the results of user study.
- In appendix A.5, we present the quantitative comparison between IterComp and RPG.
- In appendix A.6, we present the quantitative comparison between IterComp and two layout-based models: InstanceDiffusion and MIGC.
- In appendix A.7, we provide more visualization results for IterComp.

A.1 PRELIMINARY

Stable Diffusion Stable Diffusion (SD) (Rombach et al., 2022) performs multi-step denoising on random noise $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to generate a clear latent $\boldsymbol{z}_0$ in the latent space under the guidance of text prompt $\boldsymbol{c}$. During the training, an input image $\boldsymbol{x}_0$ is processed by a pretrained autoencoder to obtain its latent representation $\boldsymbol{z}_0$. A random noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is injected into $\boldsymbol{z}_0$ in the forward process as follow:

$$\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \tag{6}$$

where α_t is the noise schedule. The UNet ϵ_θ is trained to predict the added noise with the optimization objective:

$$\min_\theta \ \mathcal{L}(\theta) = \mathbb{E}_{[\boldsymbol{z}_0 \sim \mathcal{E}(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}), t]} \left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, \tau(\boldsymbol{c}))\|_2^2 \right] \tag{7}$$

where $\mathcal{E}(\cdot)$ denote the preteained encoder of VAE, $\tau(\cdot)$ denotes the pretrained text encoder.

Reward Feedback Learning Reward Feedback Learning (ReFL) (Xu et al., 2024) is proposed to align diffusion models with human preferences. The reward model serves as the preference guidance during the finetuning of the diffusion model. ReFL begins with an input prompt $\boldsymbol{c}$ and a random noise $\boldsymbol{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noise $\boldsymbol{z}_T$ is progressively denoised until it reaches a randomly selected timestep t. The latent $\boldsymbol{z}_0$ is directly predicted from $\boldsymbol{z}_t$, and the decoder from a pretrained VAE is used to generate the predicted image $\boldsymbol{x}_0$. The pretrained reward model $\mathcal{R}(\cdot)$ provides a reward score as feedback, which is used to finetune the diffusion model as follows:

$$\min_\theta \ \mathcal{L}(\theta) = -\mathbb{E}_{\boldsymbol{c}\sim\mathcal{C}} \left(\mathcal{R}\left(\boldsymbol{c}, \boldsymbol{x}_0\right) \right) \tag{8}$$

where the prompt $\boldsymbol{c}$ is randomly selected from the prompt dataset $\mathcal{C}$.

A.2 THEORETICAL PROOF OF THE EFFECTIVENESS OF ITERATIVE FEEDBACK LEARNING

A.2.1 PROOF OF LEMMA 1

Proof of Lemma 1. Considering the general form of RLHF, we change the optimization problem of iterative feedback learning to a bilevel optimization (Wallace et al., 2024; Ding et al., 2024):

$$\begin{aligned} &\min_{\mathcal{R}} \quad -\mathbb{E}_{\left[\boldsymbol{c}\sim\mathcal{C}, (\boldsymbol{x}_0^w, \boldsymbol{x}_0^l)\sim p_{\mathcal{R}}^*(\cdot|\boldsymbol{c})\right]} \left[\log \sigma \left(\mathcal{R}\left(\boldsymbol{c}, \boldsymbol{x}_0^w\right) - \mathcal{R}\left(\boldsymbol{c}, \boldsymbol{x}_0^l\right) \right) \right] \\ &\text{s.t. } p_{\mathcal{R}}^* := \arg\max_p \mathbb{E}_{\boldsymbol{c}\sim\mathcal{C}} \left[\mathbb{E}_{\boldsymbol{x}_0\sim p(\cdot|\boldsymbol{c})} \mathcal{R}(\boldsymbol{c}, \boldsymbol{x}_0) \right] - \beta \mathbb{D}_{\mathrm{KL}}[p\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right) || p_{\mathrm{ref}}\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right)] \end{aligned} \tag{9}$$

where $p_{\mathcal{R}}^*$ denotes the optimized base models under the guidance of reward model $\mathcal{R}$. We have the reparameterization of the reward model (also shown in previous works by (Wallace et al., 2024)):

$$\mathcal{R}(\boldsymbol{c}, \boldsymbol{x}_0) = \beta \mathbb{E}_{p_{\mathcal{R}}(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0, \boldsymbol{c})} \left[\log \frac{p_{\mathcal{R}}^*\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right)}{p_{\mathrm{ref}}\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right)} \right] + \beta \log Z(\boldsymbol{c}) \tag{10}$$

$$Z(\boldsymbol{c}) = \sum_{\boldsymbol{x}} p_{\text{ref}}\left(\boldsymbol{x}_{0:T} \mid \boldsymbol{c}\right) \exp\left(\mathcal{R}(\boldsymbol{c}, \boldsymbol{x}_0)/\beta\right) \tag{11}$$

Substituting this reward reparameterization into eq. (9), we get the new optimization objective as:

$$\min_{p^*_{\mathcal{R}}} -\mathbb{E}_{\left[\boldsymbol{c}\sim\mathcal{C},(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim p^*_{\mathcal{R}}(\cdot|\boldsymbol{c})\right]}\left[\log\sigma\left(\beta\log\frac{p^*_{\mathcal{R}}\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)} - \beta\log\frac{p^*_{\mathcal{R}}\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}\right)\right] \tag{12}$$

This new optimization objective is denoted as $J(p^*_{\mathcal{R}})$, we get:

$$\max_{p^*_{\mathcal{R}}} J(p^*_{\mathcal{R}}) = \mathbb{E}_{\left[\boldsymbol{c}\sim\mathcal{C},(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim p^*_{\mathcal{R}}(\cdot|\boldsymbol{c})\right]}\left[\log\sigma\left(\beta\log\frac{p^*_{\mathcal{R}}\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)} - \beta\log\frac{p^*_{\mathcal{R}}\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}\right)\right] \tag{13}$$

We use p_θ to parameterize the policy and formulate the final optimization objective as:

$$\max_{\theta} J(\theta) = \mathbb{E}_{\left[\boldsymbol{c}\sim\mathcal{C},(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim p^*_{\theta}(\cdot|\boldsymbol{c})\right]}\left[\log\sigma\left(\beta\log\frac{p^*_{\theta}\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)} - \beta\log\frac{p^*_{\theta}\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}\right)\right] \tag{14}$$

□

A.2.2 Proof of Theorem 1

Proof of Theorem 1. The gradient of the optimization objective in eq. (14) can be written as:

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_{\boldsymbol{c},\boldsymbol{x}_0^w,\boldsymbol{x}_0^l} p_\theta(\boldsymbol{x}^w_{0:T}\,|\,\boldsymbol{c})\, p_\theta(\boldsymbol{x}^l_{0:T}\,|\,\boldsymbol{c})\left[\log\sigma\left(\beta\log\frac{p^*_\theta\left(\boldsymbol{x}^w_{0:T}\,|\,\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^w_{0:T}\,|\,\boldsymbol{c}\right)} - \beta\log\frac{p^*_\theta\left(\boldsymbol{x}^l_{0:T}\,|\,\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^l_{0:T}\,|\,\boldsymbol{c}\right)}\right)\right] \tag{15}$$

Assume that:

$$F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l) = \log\sigma\left(\beta\log\frac{p^*_\theta\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right)} - \beta\log\frac{p^*_\theta\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}{p_{\text{ref}}\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)}\right) \tag{16}$$

$$\hat{p}_\theta\left(\boldsymbol{x}^w_{0:T}, \boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right) = p_\theta\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right) p_\theta(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}) \tag{17}$$

The gradient can be decomposed into two terms:

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta \sum_{\boldsymbol{c},\boldsymbol{x}_0^w,\boldsymbol{x}_0^l} \hat{p}_\theta\left(\boldsymbol{x}^w_{0:T}, \boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right) F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l) \\ &= \underbrace{\sum_{\boldsymbol{c},\boldsymbol{x}_0^w,\boldsymbol{x}_0^l} \nabla_\theta \hat{p}_\theta\left(\boldsymbol{x}^w_{0:T},\boldsymbol{x}^l_{0:T}\,|\,\boldsymbol{c}\right) F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)}_{T_1} + \underbrace{\mathbb{E}_{\left[\boldsymbol{c}\sim\mathcal{C},(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim p^*_\theta(\cdot|\boldsymbol{c})\right]}\left[\nabla_\theta\left[F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l)\right]\right]}_{T_2}\end{aligned} \tag{18}$$

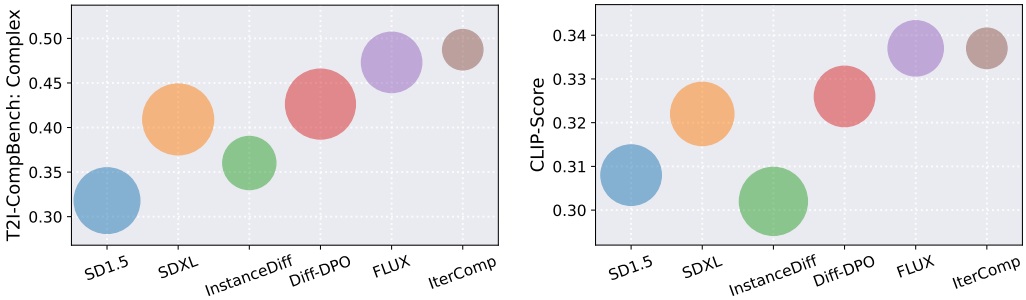
By expanding the distribution $\hat{p}_\theta$ in T_1, a more specific form is obtained:

$$\begin{aligned}T_1 &= \sum_{\boldsymbol{c},\boldsymbol{x}_0^w,\boldsymbol{x}_0^l} \nabla_\theta \hat{p}_\theta\left(\boldsymbol{x}^w_{0:T},\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right) F_\theta(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l) \\ &= \mathbb{E}\left[\left(\nabla_\theta \log p_\theta\left(\boldsymbol{x}^w_{0:T}\mid\boldsymbol{c}\right) + \nabla_\theta \log p_\theta\left(\boldsymbol{x}^l_{0:T}\mid\boldsymbol{c}\right)\right) F_\theta\left(\boldsymbol{c}, \boldsymbol{x}_0^w, \boldsymbol{x}_0^l\right)\right]\end{aligned} \tag{19}$$

□

A.3 Analysis on Model Stability

To evaluate the model stability, we selected five methods for comparison: SD1.5 (Rombach et al., 2022), SDXL (Podell et al., 2023), InstanceDiffusion (Wang et al., 2024a), Diffusion-DPO (Wallace et al., 2024), and FLUX (BlackForest, 2024), along with two evaluation metrics: Complex and CLIP-score. Using the same 50 seeds, we calculated the mean and variance of the models' performance for these metrics. To facilitate visualization, we used the variance of each method as the radius and scaled it uniformly by a common factor (10^4) for stability analysis.



(a) T2I-CompBench: Complex. (b) CLIP Score.

Figure 8: Analysis on model stability.

Regarding the stability of compositionality, as shown in fig. 8a, we found that IterComp not only achieved the best overall performance but also demonstrated superior stability. This can be attributed to the iterative feedback learning paradigm enable the model to analyze and refine its output at each optimization step, effectively self-correcting and self-improving. The iterative training approach enables the model to perform feedback training based on its own generated samples rather than solely relying on external data, this enables the model to steadily improve over multiple iterations based on its own foundation. This enables the model to steadily improve over multiple iterations, building on its existing foundation, which significantly enhances its stability.

For the stability of realism or generation quality, as shown in fig. 8b, our method also exhibited the highest stability. Therefore, the iterative training approach not only improves the model's performance but also substantially enhances its stability across different dimensions.

A.4 User Study

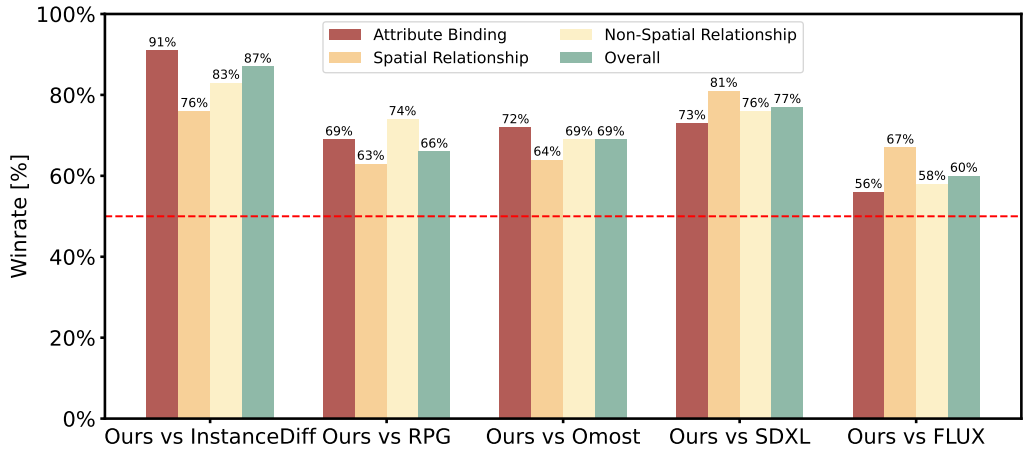


Figure 9: Results of user study.

We conducted a comprehensive user study to evaluate the effectiveness of IterComp in compositional generation. The study involved 41 randomly selected participants from diverse backgrounds. We compared IterComp with five other methods across four aspects: attribute binding, spatial relationships, non-spatial relationships and overall performance. Each comparison involved 25 prompts, culminating in a final survey of 125 prompts and generating 20,500 votes. From the win rate distribution of IterComp shown in the fig. 9, it is evident that IterComp demonstrates significant advantages across all three aspects of compositional generation.

Specifically, compared to the layout-based model InstanceDiffusion (Wang et al., 2024a), IterComp shows an absolute advantage in attribute binding. For text-based models SDXL (Podell et al., 2023) and FLUX (BlackForest, 2024), IterComp leads significantly in spatial relationships. This highlights that the model gallery design effectively collects composition-aware model preferences and enhances performance across different compositional aspects through iterative feedback learning.

Table 6: Comparison between IterComp and RPG on DPG-Bench

Model	Global	Entity	Attribute	Relation	Other	Average
IterComp	89.91	88.64	86.73	84.77	89.74	81.17
RPG	91.01	87.39	84.53	87.92	89.84	81.28
RPG+IterComp	92.74	91.33	89.10	92.38	90.13	84.72

Table 7: Comparison between IterComp and RPG on Genval.

Model	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall
IterComp	0.97	0.85	0.63	0.86	0.33	0.41	0.675
RPG	0.97	0.86	0.66	0.79	0.30	0.38	0.660
RPG+IterComp	0.99	0.90	0.72	0.90	0.35	0.48	0.723

A.5 Comparison between IterComp and RPG

We employed two up-to-date benchmarks: DPG-Bench (Hu et al., 2024) and GenEval (Ghosh et al., 2024) for testing to evaluate the capabilities of IterComp and RPG (Yang et al., 2024b) in compositional generation. As demonstrated in table 6 and table 7, IterComp outperforms RPG in metrics like attributes and colors. This is due to our training of a specific reward model for attribute binding, which iteratively enhances IterComp over multiple iterations. Leveraging the strong planning and reasoning capabilities of LLMs, RPG excels in areas such as relations, counting, and positioning. When IterComp is used as the backbone for RPG, the model exhibits remarkable performance across all aspects. This highlights IterComp's superiority in compositional generation. It's important to note that IterComp is a simple SDXL-like model that doesn't require complex computations during inference. As a result, under the same conditions such as prompts and inference steps, IterComp is nearly three times faster than RPG.

A.6 Comparison between IterComp and Layout-based Methods

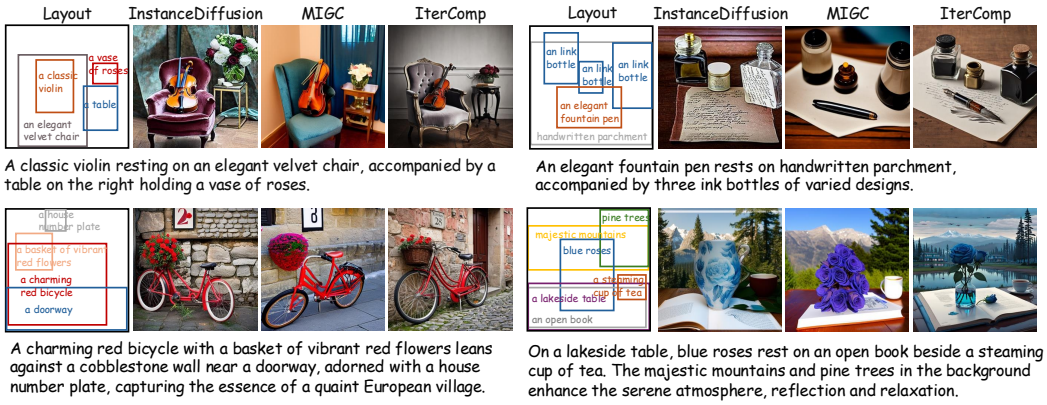


Figure 10: Qualitative comparison between IterComp and two layout-to-image methods: InstanceDiffusion and MIGC.

We provide additional experiments between IterComp, InstanceDiffusion (Wang et al., 2024a), and MIGC (Zhou et al., 2024). As shown in fig. 10, these examples clearly show that while MIGC and InstanceDiffusion can accurately generate objects in the specified positions of the layout, there is a notable gap in generation quality compared to IterComp, such as aesthetics and details. Moreover, the images generated by these two methods often appear visually unrealistic, with significant flaws such as incomplete violins or mismatches between bicycle and its basket. This highlights the clear superiority of our IterComp on compositional generaton.

A.7 More Visualization Results

Figure 11: Qualitative comparison between IterComp and three types of compositional generation methods: text-controlled, LLM-controlled, and layout-controlled approaches. We use GPT-4o to infer the layout from the prompt for InstanceDiffusion. Colored text denotes the advantages of IterComp in generated images.

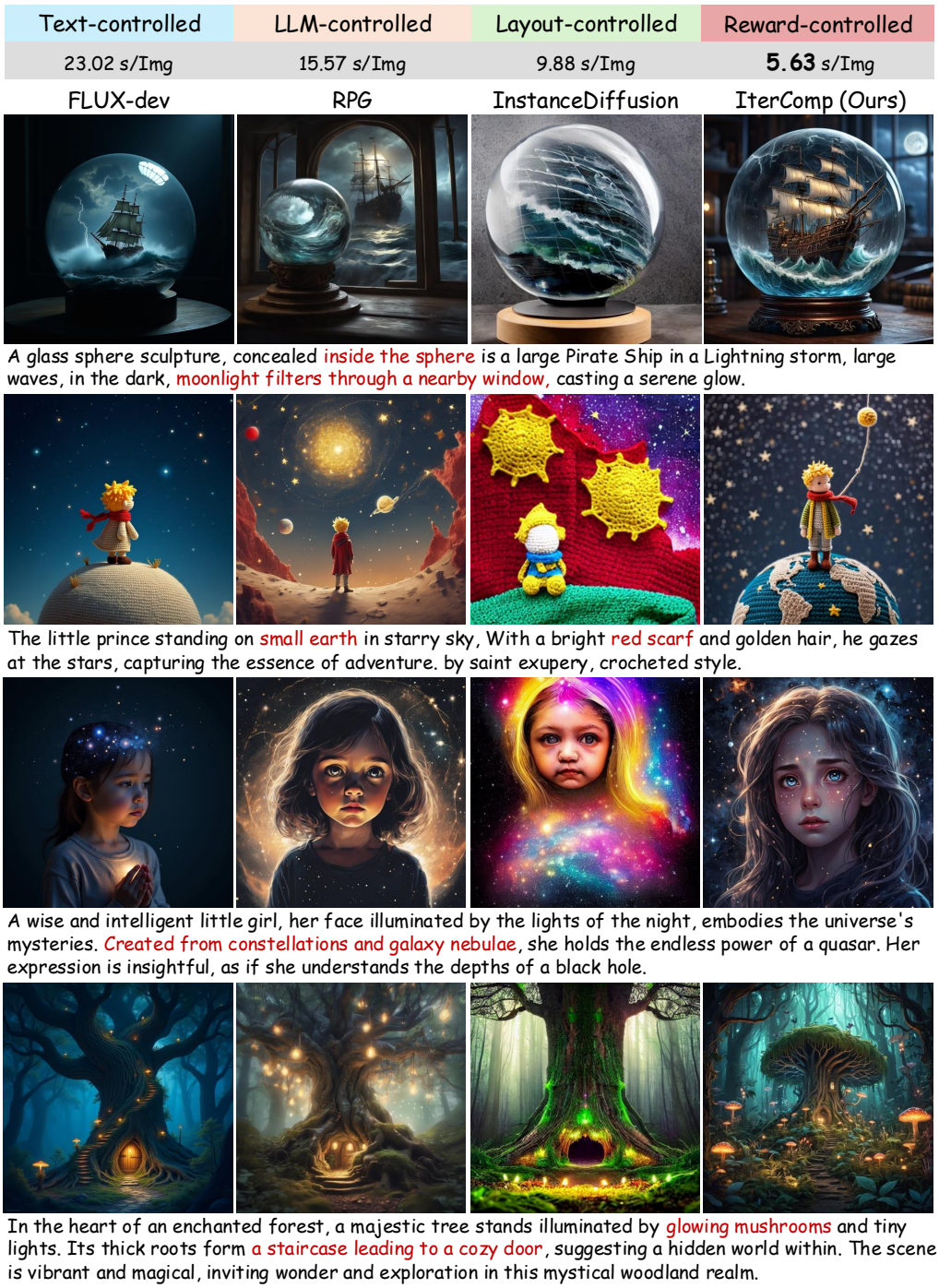


Figure 12: Qualitative comparison between IterComp and three types of compositional generation methods: text-controlled, LLM-controlled, and layout-controlled approaches. We use GPT-4o to infer the layout from the prompt for InstanceDiffusion. Colored text denotes the advantages of IterComp in generated images.

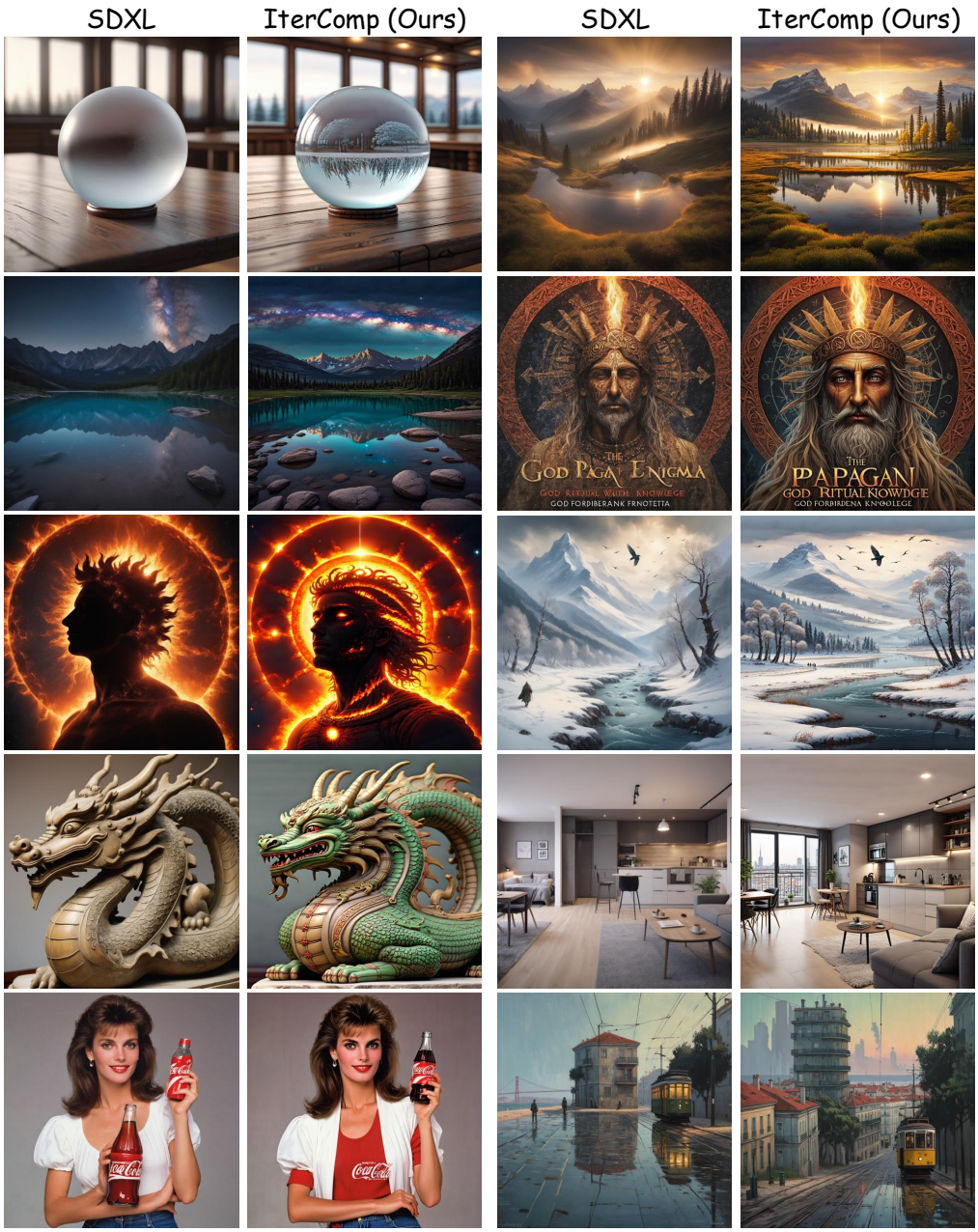


Figure 13: More visualization results for IterComp and its base diffusion model, SDXL.