HiCoGen: Hierarchical Compositional Text-to-Image Generation in Diffusion Models via Reinforcement Learning

Hongji Yang[1], Yucheng Zhou[1], Wencheng Han[1],
Runzhou Tao[2], Zhongying Qiu[2], Jianfei Yang[2], Jianbing Shen[1†]
[1]SKL-IOTSC, CIS, University of Macau
[2]Zhejiang ZEEKR Automobile Research & Development Co., Ltd.
† Corresponding author

Abstract

*Recent advances in diffusion models have demonstrated impressive capability in generating high-quality images for simple prompts. However, when confronted with complex prompts involving multiple objects and hierarchical structures, existing models struggle to accurately follow instructions, leading to issues such as concept omission, confusion, and poor compositionality. To address these limitations, we propose a Hierarchical Compositional Generative framework (**HiCoGen**) built upon a novel **Chain of Synthesis (CoS)** paradigm. Instead of monolithic generation, HiCoGen first leverages a Large Language Model (LLM) to decompose complex prompts into minimal semantic units. It then synthesizes these units iteratively, where the image generated in each step provides crucial visual context for the next, ensuring all textual concepts are faithfully constructed into the final scene. To further optimize this process, we introduce a reinforcement learning (RL) framework. Crucially, we identify that the limited exploration of standard diffusion samplers hinders effective RL. We theoretically prove that sample diversity is maximized by concentrating stochasticity in the early generation stages and, based on this insight, propose a novel **Decaying Stochasticity Schedule** to enhance exploration. Our RL algorithm is then guided by a hierarchical reward mechanism that jointly evaluates the image at the global, subject, and relationship levels. We also construct **HiCoPrompt**, a new text-to-image benchmark with hierarchical prompts for rigorous evaluation. Experiments show our approach significantly outperforms existing methods in both concept coverage and compositional accuracy.*

1. Introduction

Diffusion models [1] have led to significant improvements in text-to-image (T2I) generation. Many T2I models [2–7] have been proven to be able to produce high-quality and diverse images. However, existing image diffusion models fail to generate images that cover all concepts within long and complex prompts. Despite being equipped with more powerful text encoders (CLIP [8], T5-Encoder [9]), the challenges of image generation with complex prompts remain unresolved. Specifically, complex prompts contain extensive information, resulting in a large semantic gap between text and image domains, as shown in Fig. 1. This leads to several issues: ① **concept missing**, where the model overlooks parts of the prompt due to overly long textual descriptions; ② **concept confusion**, as the text encoder fails to capture hierarchical relationships, causing the model to rely on priors and generate incorrect interactions between objects; ③ **overall quality decline**, where the visual fidelity of the generated image degrades.

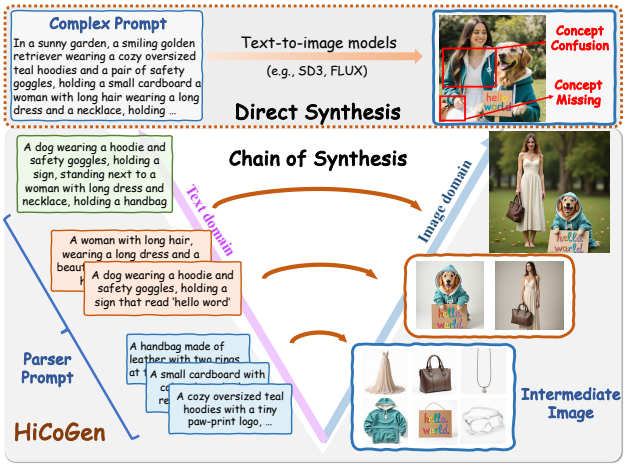


Figure 1. **The motivation of HiCoGen.** The semantic gap between text and images widens as the complexity of the text increases, particularly involving the prompts with a hierarchical relationship. While a single T2I model performs well in generating individual objects, it suffers from concept missing and confusion when processing complex prompts. HiCoGen employs a **Chain of Synthesis** for complex text to preserve the semantic content.

In stark contrast, Large Language Models (LLMs)

demonstrate powerful inference-time reasoning by scaling not only the amount of computation, but the *reasoning process* itself: techniques such as Chain-of-Thought [10] and reinforcement-style inference frameworks (e.g., Deepseek-R1 [11]) explicitly decompose a complex problem into a sequence of intermediate steps. This form of *content-level* inference-time scaling stands apart from recent work on diffusion inference-time scaling [12], which allocates additional compute to the *decoding process* (e.g., exploring more noises or sampling trajectories for a fixed holistic prompt) while keeping the textual input and its semantics unchanged. Inspired by the former line of work, we ask a different question: "*Instead of only scaling the sampling procedure of diffusion models, can we introduce a similar step-by-step, content-aware processing paradigm into visual synthesis itself to overcome the limitations of holistic generation?*"

To realize this principle in visual synthesis, we propose **Chain of Synthesis (CoS)**, a new paradigm that replaces monolithic generation with a progressive, part-by-part construction process. Building on this paradigm, our framework, **HiCoGen**, is the first system to realize this paradigm, engineered specifically for prompts with complex hierarchical and relational structures (e.g., "*one person holding an object standing next to another person who is also holding an object*"). Within HiCoGen, we first leverage an LLM to decompose the complex prompt into a set of minimal semantic units and rewrite the prompt for generation. The framework then synthesizes these units iteratively: the image generated in each step becomes the critical visual context for the next. This process constructs a visual compositional pathway, which progressively replaces entangled textual descriptions with concrete visual information, ensuring a faithful construction of the final complex scene.

Moreover, to optimize the fidelity of the CoS pipeline, we introduce a reinforcement learning (RL) framework based on GRPO [11]. However, effectively applying RL to diffusion models is a major challenge, as standard sampling methods offer a limited exploration space, hindering the discovery of optimal generation policies. To overcome this fundamental barrier, we first conduct a rigorous theoretical analysis of the stochastic sampling dynamics. We prove that sample diversity is critically dependent on early-stage stochasticity. Based on this insight, we introduce a **Decaying Stochasticity Schedule**, a novel sampling mechanism that provably maximizes exploration and is thus essential for enabling effective RL. Building on this enhanced sampler, our RL algorithm is guided by a novel **hierarchical reward mechanism**, which supplies global-, subject-, and relationship-level rewards for step-by-step supervision.

Finally, to facilitate rigorous and standardized evaluation of compositional reasoning in T2I models, a capability central to our work, we construct **HiCoPrompt**. It is a new, systematically generated benchmark featuring prompts with clearly defined hierarchical and compositional structures. Extensive experiments on HiCoPrompt validate that our HiCoGen significantly outperforms existing models.

Our contribution can be summarized as follows:

- We introduce a novel Chain of Synthesis (CoS) paradigm for compositional generation, and propose HiCoGen, the first framework to realize this step-by-step process for T2I models, overcoming limitations of monolithic methods.
- Combined with our enhanced sampler and a GRPO-based algorithm, we design a novel hierarchical reward mechanism that provides fine-grained supervision on compositional aspects.
- We identify the limited exploration of standard diffusion samplers as a key barrier to RL-based optimization. We prove that optimal sample diversity is achieved by concentrating stochasticity in the early generation stages and introduce a Decaying Stochasticity Schedule to realize it.
- We construct the HiCoPrompt benchmark with explicit hierarchical structures to facilitate robust evaluation in compositional text-to-image generation.

2. Related Work

Text-to-Image Diffusion Models. From UNet-based diffusion models (e.g., LDM [4], SDXL [13]) to the more powerful Transformer-based diffusion models (e.g., DiT [3], SD3 [7], FLUX [5]), Text-to-image diffusion models have been proven to produce high-fidelity images. However, they often fail to generate images under complex prompts involving multiple entities or subjects. To introduce more controllability into the diffusion model, recent work [14–19] has included additional adapters or adjusted the prompt for a more accurate generation.

Subject-driven Diffusion Models. While T2I models can generate diverse images of generic concepts, they still struggle to synthesize images of a specific subject provided by users. DreamBooth [20] and textual inversion [21] apply LoRA [22] to insert a certain concept and perform a subject-driven generation. However, these methods only focus on the trained subject and are limited in their generalization. To improve generalization, recent works [23–26] have introduced the in-context generation framework for subject-driven generation or customized generation. Nevertheless, In-Context Generation still suffers from missing the given concepts or confusing different concepts.

Diffusion Reinforcement Learning. Inspired by the great success of LLMs finetuning using RL from Human Feedback (RLHF) [27], DDPO [28], Diffusion-DPO [29] and DPOK [30] introduced Direct Preference Optimization (DPO) [31] into diffusion models to align with human preferences. Recent works [32, 33, 33, 34] validate the improvement of the diffusion with reinforcement learning. Motivated by the success of Deepseek-R1 [11] in training the

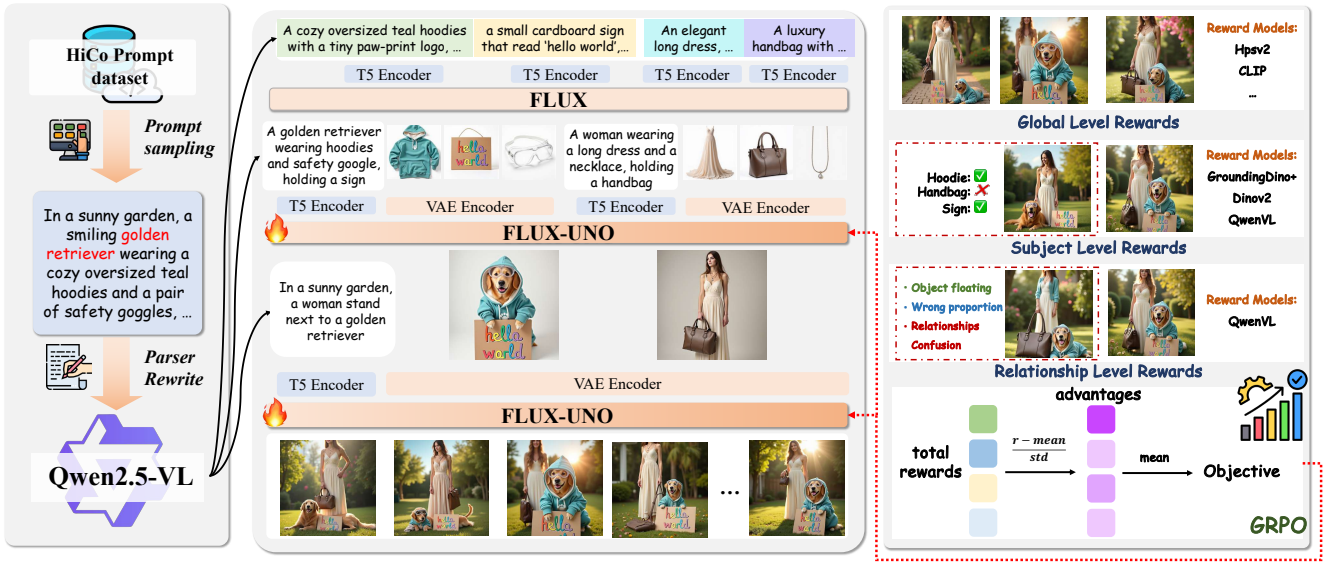


Figure 2. The overall pipeline of our proposed HiCoGen framework. When facing a complex hierarchical compositional prompt, HiCoGen applies the Chain of Synthesis to progressively construct the image part-by-part and employs in-context generative models to assemble the different components into the final image. This ensures all the concepts in the text domain are present in the image domain.

thinking process with RL, Flow-GRPO [35] and Dance-GRPO [36] employ GRPO [37] to further enhance the performance of visual generation.

3. Method

3.1. Preliminary

Diffusion Models. The forward process of diffusion is linearly adding a Gaussian noise item ϵ to the real data z_0:

$$\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\epsilon, \tag{1}$$

where the $\mathbf{u} = \epsilon - \mathbf{z}_0$ is defined as the "velocity" in restrict flow [7]. To reach a lower noise level $\mathbf{s}$, we have:

$$\mathbf{z}_s = \mathbf{z}_t + \hat{\mathbf{u}} \cdot (s-t) \tag{2}$$

where the $\hat{\mathbf{u}}$ is the denoising model output at timestep t. Given this output, the noisy latent can eventually be converted to a clear version.

GRPO. Given a condition c, the generative models will sample multiple outputs $\{y_1, y_2, \ldots, y_n\}$ from the model $\pi_{\theta_{old}}$ and compute a group of rewards $\{r_1, r_2, \ldots, r_n\}$ with reward models. The advantage function A_i for each output will be calculated as:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_n\})}{\text{std}(\{r_1, r_2, \ldots, r_n\})} \tag{3}$$

Then, the policy model π_θ can be optimized by maximizing the following objective function:

$$\mathcal{J}(\theta) = \mathbb{E}_{\{y_i\}_{i=1}^{n} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{c}), \mathbf{a}_{t,i} \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{s}_{t,i})} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{1}{T} \sum_{t=1}^{T} \min\left(\rho_{t,i} A_i, \text{clip}(\rho_{t,i}, 1-\epsilon, 1+\epsilon) A_i\right) \right], \tag{4}$$

3.2. HiCoGen Framework

Diffusion models often struggle to follow long or complex text conditions, particularly when a prompt involves multiple complex subjects. The core idea of the HiCoGen Framework is to transform the intractable monolithic generative task into a chain of tractable single-subject generation and in-context composition tasks. Instead of demanding a single diffusion model to resolve complex compositional bindings in a single step, HiCoGen first generates a high-fidelity instance of a semantic unit and then iteratively integrates these results into the new scene, one by one, using the previously generated image as the representation of certain textual information, as shown in Fig. 2.

Parse&Rewrite LLM. Given the complex input prompt $\mathcal{O}$, we first employ a semantic parser (e.g., LLMs) to decompose it into an ordered set of subject-centric components.

$$\begin{aligned} \mathcal{O} &= \{P^{(1)}, P^{(2)}, \cdots, P^{(n)}\}, \\ P^{(i)} &= \{c^{(1)}, c^{(2)}, \cdots, c^{(m)}\}, \end{aligned} \tag{5}$$

where P_i is the sub-prompt for the i-th subject, and could be decomposed into multiple fine-grained component (or attributes) c_j. Then, LLM also rewrites the prompt at each level to make it feasible to generate an image.

Each layer's prompt provides detailed descriptions only for the specific subject or attribute at that level, while the higher layer includes only general categorical information without such details.

Chain of Synthesis. To perform a CoS, a primary challenge is how to assemble the generated contents. For a text-to-image DiT model with multi-modal attention, the input z is the concatenated result of the text token c and the noisy

latent z_t, and can be expressed as follows:

$$z = \text{concat}(c, z_t), \tag{6}$$

After generating the specific contents $\{z_0^0, z_0^1, \cdots z_0^m\}$ in the latent domain, we leave these generated contents to continue engaging in text-to-image generation. This can be formatted as follows:

$$\hat{z} = \text{concat}(P^{(i)}, \hat{z}_t, z_0^0, z_0^1, \cdots z_0^m), \tag{7}$$

Here, we create an input $\hat{z}$ with hierarchical information $\{z_0^0, z_0^1, \cdots z_0^m\}$ with text P^i, reducing the need for the original prompt description on specific contents. And this process can be continued until we generate an image that covers the entire prompt.

$$z^{'} = \text{concat}(\mathcal{O}, z_t^{'}, \hat{z}_0^0, \hat{z}_0^1, \cdots \hat{z}_0^n), \tag{8}$$

where z_0 denotes the clear latent at timestep $t = 0$. This chained synthesis assembles all the generated results and also facilitates the hierarchical relationship generation.

3.3. Hierarchical Reward

To provide a comprehensive and fine-grained reward signal for the quality and accuracy of diffusion models, we propose a novel hierarchical reward mechanism. It not only focuses on global text-to-image alignment or aesthetic quality but also aims at the image's internals to quantify the accuracy of key subjects and the relationships between them. Our overall reward function, R_{total}, is a weighted sum of three core components: a global reward R_{global}, a subject-specific reward R_{subject}, and a subject relationship reward $R_{\text{relationship}}$.

$$R_{\text{total}} = R_{\text{global}} + R_{\text{subject}} + R_{\text{relationship}}, \tag{9}$$

Global Reward. The global reward R_{total} aims to assess the overall quality of the generated images from a holistic perspective. It consists of two main aspects: text-image alignment and human preference.

$$R_{\text{global}} = w_{\text{clip}} \cdot S_{\text{clip}} + w_{hps} \cdot S_{\text{hps}} \tag{10}$$

where S_{clip} andS_{hps} denote the image reward score from the CLIP model [8] and human-preference score [38].

Subject-Specific Reward. In many generation tasks, accurately depicting one or more key subjects is crucial. R_{subject} focuses on evaluating the fidelity and attributes of these specific subjects. For a specific subject, we first use the GroundingDino [39] to locate the positions of key subjects in the prompt, obtaining their bounding boxes, and then crop the subject. After the corresponding subject is located, we assess the similarity between the cropped part and the previously generated intermediate image using the DINOv2 [40]. We compute the embedding vector using cosine similarity $\cos(\cdot, \cdot)$ and the process can be expressed as:

$$S_{\text{DINOv2}} = \cos\left(\text{DINOv2}(I_{\text{cropped}}), \text{DINOv2}(I_{\text{ref}})\right) \tag{11}$$

where I_{cropped} and I_{ref} denote the cropped image and the reference image, respectively.

In addition, many key attributes (such as color and pose) of the subject are difficult to capture solely through embedding similarity. For this, we utilize the powerful Vision-Language Models (VLM) as the rule-based reward model to perform the rewarding process. For the N key subjects in the image, the subject reward R_{subject} is an aggregation of the scores for each subject:

$$R_{\text{subject}} = \frac{1}{N} \sum_{i=1}^{N} \left(w_{\text{dino}} \cdot S_{\text{DINOv2}}^{(i)} + w_{\text{vlm}} \cdot S_{\text{vlm}}^{(i)} \right) \tag{12}$$

Relationship Reward. Since the subject image is generated using an independent reference image, the following issues arise: 1) The rendered object appears to float or feel detached, with the interaction area appearing very stiff, resulting in an obvious pasted-on effect; 2) The object's proportions are significantly enlarged.

Therefore, the rewards focus on the relationship between the subject injection and the main content of the image, as well as whether the interaction is reasonable. We prompt VLM to focus on relationships between the given subjects.

$$R_{\text{relationship}} = \frac{1}{N} \sum_{i=1}^{N} \left(S_{\text{vlm}}^{(i)} \right) \tag{13}$$

In the practical implementation, we make the VLM to verify the hierarchical relationships and the relative proportions. Please refer to the supplementary for the prompts format.

3.4. Decay Stochasticity Schedule

Sampling SDEs. To enable exploration for reinforcement learning, we augment the standard reverse diffusion process with a controllable stochasticity term $\eta(t) \geq 0$. The resulting reverse-time SDE, evolving from $t = T$ to 0, is:

$$\mathrm{d}\mathbf{z}_t = \left[\mathbf{f}(\mathbf{z}_t, t) - g(t)^2 \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)\right] \mathrm{d}t + g(t)\eta(t)\mathrm{d}\mathbf{w}_t, \tag{14}$$

where $\mathbf{f}$ and g are the drift and diffusion coefficients, and $\mathbf{w}_t$ is a standard Wiener process. Our objective is to determine the optimal schedule for $\eta(t)$ to maximize sample diversity.

Optimal Stochasticity Scheduling. The denoising process has distinct temporal phases. Recent work [41] categorizes it into an early "generalization time" for establishing global structure and a later "collapse time" for refining details. This motivates concentrating exploration efforts early. We formalize this intuition by aiming to maximize sample diversity, quantified by $\text{Tr}(\text{Cov}(\mathbf{z}_0))$, under a fixed total stochasticity budget.

Theorem 1 (Optimal Stochasticity Allocation for Diversity Maximization)**.** *Consider the reverse SDE in Eq. 14 with a fixed budget* $\int_0^T \eta(t)^2 dt = C > 0$*. Under Assumptions 1 and 2 below, the schedule* $\eta(t)$ *that maximizes the final sample diversity* $Tr(Cov(\mathbf{z}_0))$ *is a monotonically decreasing function of time* t *(from* T *to* 0*).*

Assumption 1 (Linearizability)**.** The SDE dynamics can be analyzed by linearizing around a mean deterministic trajectory $\bar{\mathbf{z}}_t$.

Assumption 2 (Increasing Contractivity)**.** The Jacobian of the drift, $\mathbf{A}_t$, becomes increasingly contractive as $t \to 0$. This reflects the increasingly dominant role of the score function in collapsing the latents onto the data manifold, transitioning from "generalization time" to "collapse time".

Proof of Theorem 1. Under Assumption 1, the covariance matrix $\mathbf{\Sigma}_t = \text{Cov}(\mathbf{z}_t)$ evolves according to the Lyapunov equation: $\frac{d\mathbf{\Sigma}_t}{dt} = \mathbf{A}_t\mathbf{\Sigma}_t + \mathbf{\Sigma}_t\mathbf{A}_t^\top + g(t)^2\eta(t)^2\mathbf{I}$. The final covariance, starting from $\mathbf{\Sigma}_T = \mathbf{0}$, is the solution integrated backwards in time:

$$\mathbf{\Sigma}_0 = \int_0^T \mathbf{\Phi}(0,s)\left(g(s)^2\eta(s)^2\mathbf{I}\right)\mathbf{\Phi}(0,s)^\top ds, \tag{15}$$

where $\mathbf{\Phi}(t,s)$ is the state transition matrix for the linearized system. We thus solve the variational problem:

$$\max_{\eta(t)^2\geq 0} \int_0^T \eta(s)^2 W(s) ds, \quad \text{s.t.} \int_0^T \eta(s)^2 ds = C, \tag{16}$$

where the weight $W(s) = \text{Tr}\left(g(s)^2\mathbf{\Phi}(0,s)\mathbf{\Phi}(0,s)^\top\right)$. The optimal solution allocates the budget proportionally to the weight $W(s)$.

The weight $W(s)$ quantifies the impact of a perturbation at time s on the final variance. Per Assumption 2, the system's contractivity increases as time progresses to 0. This means the dynamics suppress perturbations more aggressively at later times. Consequently, the propagator $\mathbf{\Phi}(0,s)$ experiences less total suppression for larger s (earlier times), implying $W(s)$ is a monotonically decreasing function as s goes from T to 0. To maximize the objective, the budget $\eta(s)^2$ must be allocated where $W(s)$ is highest, i.e., at earlier times. Therefore, the optimal schedule $\eta(t)$ is monotonically decreasing. □

In our framework, we instantiate this principle on Rectified Flow. We augment its deterministic ODE with our principled stochastic schedule, implemented as a cosine decay:

$$\eta(t) = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{\pi(T_{\max} - t)}{T_{\max}}\right)\right), \quad t \in [0, T_{\max}]. \tag{17}$$

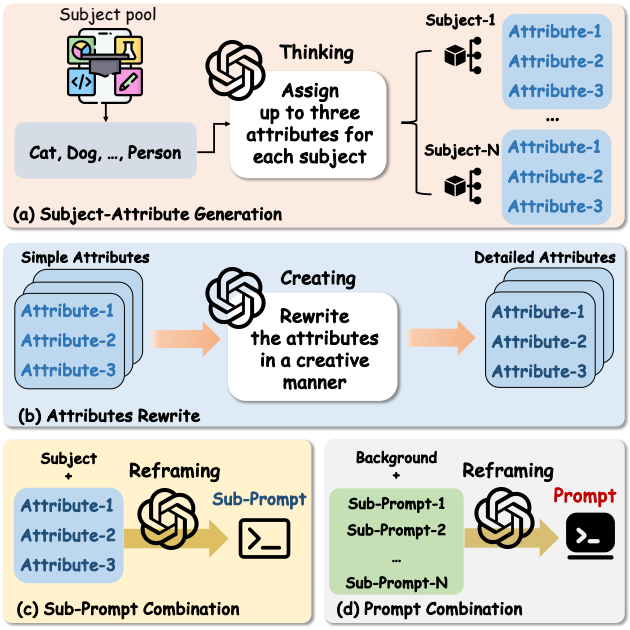


Figure 3. Illustration of our proposed HiCoPrompt dataset. It features clearly defined multiple hierarchical relationships, as well as thoroughly described concrete subjects and attributes.

This schedule focuses exploration on the crucial early stages of structural formation while preserving fidelity during the final refinement phase, thereby providing an optimal solution to the exploration-exploitation trade-off.

4. HiCoPrompt Dataset

To construct a hierarchical compositional prompt dataset for challenging text-to-image generation, we propose the HiCoPrompt dataset. This dataset defines multiple subjects with multiple attributes for a single image, which is absent in other existing T2I benchmarks. Specifically, each prompt in the HiCoPrompt describes multiple subjects with structured specific attributes (e.g., clothing or holding). Each attribute is modified by descriptive qualifiers. The comparison with other T2I benchmarks is present in Tab. 1. The benchmark comprises 3k prompts for testing and 12k prompts for RL training. The dataset will be available.

Subject Generation. The initial phase focuses on establishing a foundation of core elements. We first select a diverse set of Primary Subjects (e.g., cat, king, astronaut) chosen for their high potential for visual representation and modification. For each subject, we employ ChatGPT [42] to generate essential quantifiable core attributes. These attributes serve as standard slots for injecting descriptive details, thereby ensuring their richness and uniqueness.

Attribute Rewriting. The second phase introduces descriptive variation by applying different modifiers and detailed descriptions. We ask the LLM to rewrite the attribute in an imaginative form. The rewriting transforms a simple category into a specific object with complex concepts.

Benchmark	Prompts	Sources	#Subjects/Prompt	Subject Description	Hierarchy	Composition	Length
DrawBench [43]	200	Human	1-4	Simple	×	×	15
T2I-CompBench [44]	5,000	Template	1-4	Simple	×	×	10
T2I-CompBench++ [45]	8,000	Template	1-4	Simple	×	×	14
DPG-Bench [46]	1,065	Template	1-3	Simple	×	×	84
LongBench-T2I [47]	500	LLM	9	Detailed	×	✓	683
HiCoPrompt(Ours)	3,000	LLM	4-12	Detailed	✓	✓	219

Table 1. Comparison of other T2I benchmarks. The primary challenge of HiCoPrompt compared to other datasets lies in its detailed and specific descriptions for each subject, ensuring that even subjects within the same category exhibit distinct characteristics. At the same time, there exists a clear hierarchical relationship between subjects.

Methods	$\text{Acc}_{\text{exist}}$ ↑	$\text{Acc}_{\text{attribute}}$ ↑	$\text{Acc}_{\text{relationship}}$ ↑	CLIP Score↑	HPSv2↑
SDXL [13]	0.2672	0.0825	0.0760	0.2676	0.2379
Stable-Diffusion-3 [7]	0.3669	0.2991	0.3026	0.2774	0.2810
FLUX.1-dev [5]	0.4456	0.3805	0.4035	0.2628	0.2974
FLUX.1-Krea [6]	0.5172	0.5992	0.6400	0.2641	0.2952
Qwen-Image [48]	0.6292	0.6907	0.7829	0.2687	0.2913
MS-Diffusion [23]	0.4552	0.1274	0.1960	0.2556	0.2330
OminiGen [24]	0.5818	0.6791	0.6789	0.2646	0.2893
OminiControl [25]	0.4032	0.6583	0.3044	0.2607	0.2886
UNO [26]	0.4400	0.6923	0.3697	0.2691	0.2698
HiCoGen(ours)	**0.7127**	**0.7673**	**0.8203**	**0.3192**	**0.3357**

Table 2. Comparison of the proposed HiCoGen and other text-to-image models. HiCoGen outperforms other text-to-image models or subject-driven generative models in all metrics.

Sub-Prompt Reframing. The full composite prompt is assembled by combining multiple subjects, each with a structured mix of these attributes. We require the LLM to construct the final prompt by incorporating the necessary prepositions or actions between subjects to integrate the distinct sentences. The construction is present in Fig. 3

5. Experiment

5.1. Experimental Setup

We employ our experiment with flow-based diffusion models FLUX and the UNO LoRA weights for subject-driven generation. All experiments are conducted with 8 A100 80G GPUs. We employ the AdamW optimizer with a fixed learning rate of 1e-4 for fine-tuning. The rank and alpha of the LoRA are both set to 512, and the number of samples is set to 16. We employ Qwen2.5-VL-3B as both the parse and rewrite LLM and the reward model. The η_{max} is set to 1.0 in our decaying stochasticity scheduler. Please refer to the supplementary for more details.

5.2. Evaluation Metrics

To further investigate how the RL fine-tuning improves the proposed HiCoGen, we design three corresponding metrics to evaluate the results based on GPT-4o [42]. We prompt the GPT-4o to perform reasoning on three dimensions: 1) Does the subject/attribute exist? 2) Does the attribute align with the given description in detail? 3) Are the subject's relationships (e.g., interaction, proportion) accurate? In addition, we employ the commonly used CLIP score [8] and HPSv2 [38] score for comparison.

5.3. Quantitative Results

We primarily compare our HiCoGen with two categories of methods: end-to-end text-to-image generative models (e.g., SDXL [13], SD3 [7], FLUX [5]) and subject-driven text-to-image generative models (e.g., MS-Diffusion [23], UNO [26]). For these subject-driven models that only support single-subject generation, we employed a multi-step approach to inject the new subject with the prompt into the image one by one. As shown in Tab. 2, HiCoGen outperforms other text-to-image models in all metrics. For subjects existing and attributes accuracy, HiCoGen achieves a 9% improvement compared to Qwen-Image, which is known for handling complex prompts. HiCoGen also surpasses FLUX and Qwen-Image in the visual appeal metric HPSv2 by 0.04. It is worth noting that using a subject-driven generative model (e.g., UNO) in such tasks preserves attribute details in the prompt better than other text-to-image models, although concept missing still occurs.

5.4. Qualitative Results

In Fig. 4, we present visual results comparing our HiCoGen with other text-to-image methods. The conventional T2I model is good at depicting a single subject. Both FLUX and FLUX-Krea achieve good visual results in the HiCoPrompt.

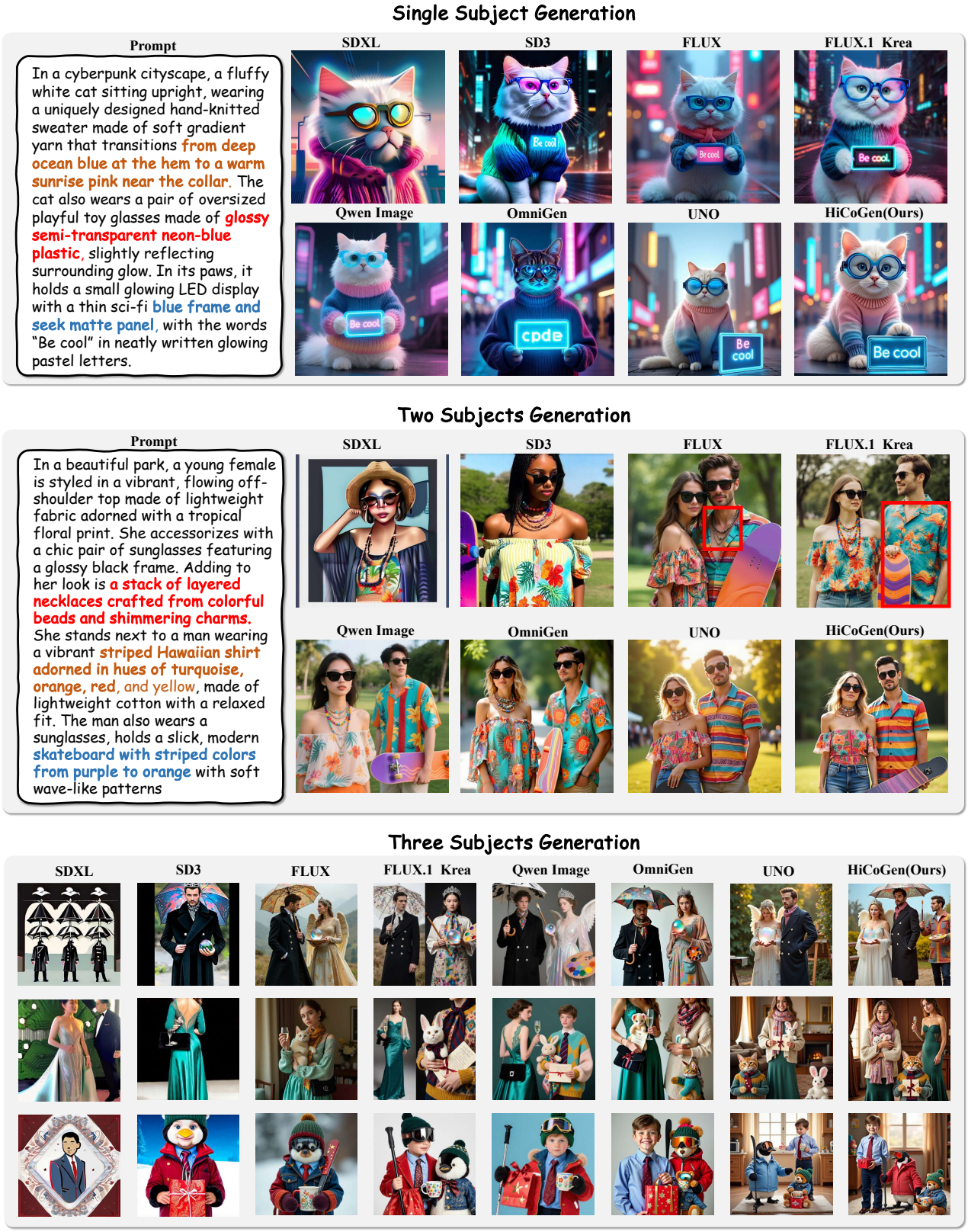


Figure 4. The visual result of our HiCoGen. HiCoGen greatly mitigates the issue of concept missing or confusion in image generation. When handling prompts with clear hierarchical relationships and multiple complex subjects, it significantly outperforms other T2I models.

Reward			$\text{Acc}_{\text{exist}}$ ↑	$\text{Acc}_{\text{attribute}}$ ↑	$\text{Acc}_{\text{relationship}}$ ↑	CLIP Score↑	HPSv2↑
Global-Level	Subject-Level	Relationship-Level					
✓	✓	✓	**0.7128**	0.7672	**0.8202**	0.3192	**0.3357**
	✓	✓	0.6828	**0.7743**	0.7810	0.2670	0.2720
✓		✓	0.6924	0.6206	0.7164	**0.3208**	0.3346
✓	✓		0.7038	0.7558	0.7689	0.3106	0.3327

Table 3. The ablation studies of using different rewards in HiCoGen.

Models	$\text{Acc}_{\text{exist}}$ ↑			$\text{Acc}_{\text{attribute}}$ ↑			$\text{Acc}_{\text{relationship}}$ ↑		
	1	2	3	1	2	3	1	2	3
Qwen-Image	**0.93**	0.80	0.36	0.93	0.61	0.54	**0.89**	0.85	0.61
UNO	0.62	0.42	0.28	0.94	0.62	0.53	0.52	0.34	0.25
HiCoGen	0.92	**0.81**	**0.41**	**0.95**	**0.74**	**0.62**	**0.89**	**0.86**	**0.71**

Table 4. Different number of subjects in generated images.

However, there are some inconsistencies between the image and the given prompt. For example, they fail to realize the description "*from deep ocean blue at the hem to a warm sunrise pink near the collar*" and FLUX-Krea incorrectly adds an extra paw of the cat to "*hold the LED board.*" When dealing with multiple subjects, text-to-image models tend to produce images with conceptual confusion. For example, FLUX-Krea [6] and Qwen-Image [48] incorrectly apply the "*floral clothing*" attribute to a different subject, or FLUX assigns the "*stacked necklaces*" to the wrong subject.

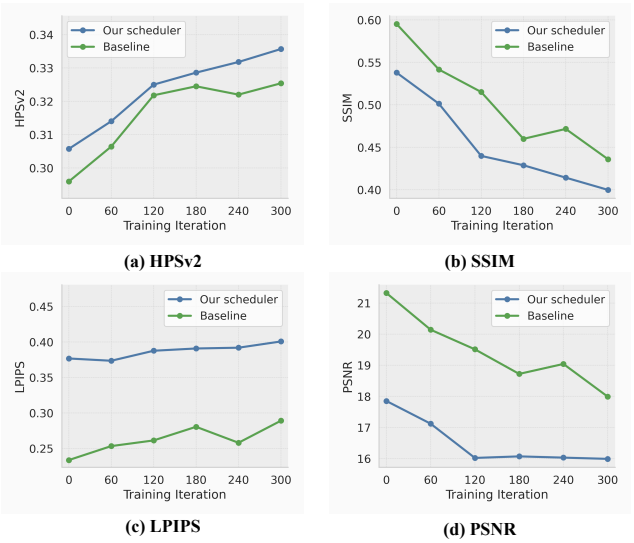
5.5. Ablation Study

For the ablation studies, we conduct experiments on our HiCoGen to analyze the contributions of different reward components. As shown in Tab. 3, the results show that combining global, subject, and relationship rewards yields the best overall performance. Removing the subject reward significantly decreases attribute accuracy (about 14% decline), while removing the relationship reward weakens relationship construction (about 4% reduction in accuracy).

5.6. Discussion

The accuracy of the generated subject. In qualitative results, we have preliminarily demonstrated that the more subjects involved within the prompt, the more challenging for the generative model to generate the corresponding images. When focused on generating 1-2 subjects, the quality and accuracy of the generated images are both excellent. However, when the number of subjects increases to 3, the accuracy shows a significant decline (51% in existing accuracy and 33.5% in attributes). Despite this, HiCoGen still outperforms Qwen-Image and UNO, as shown in Tab 4.

Stochasticity in Diffusion RL. In Fig. 5, we give the curve of similarity between the sampled results using different stochasticity strategies. We employ SSIM, PSNR and LPIPS to measure the similarity between different samples, as well as the result of the trained models. Note that the lower values of the SSIM and PSNR indicate lower similarity between the sampled results. Compared to the baseline, the image sampled from our pipeline has lower similarity, which means the diversity of the samples, thus leading to a higher value of HPSv2 score.



(a) HPSv2 (b) SSIM (c) LPIPS (d) PSNR

Figure 5. Similarity between samples in diffusion GRPO during the training process. The samples obtained from baseline are highly similar, which reduces the diversity of the samples.

6. Conclusion

Text-to-image diffusion models often fail to represent all concepts in long, complex prompts. In this study, we propose HiCoGen, which applies Chain of Synthesis to alleviate the difficulty of the model in generating an image from complex text. It applies an LLM to parse and rewrite the complex prompt into multiple sub-prompts to guide each intermediate image generation. Then, these intermediate images are fed to a subject-driven generative model with the prompt to assemble these concepts until all the concepts in the complex prompt are performed in the final image. To ensure accurate CoS, we introduce a hierarchical reward for diffusion RL fine-tuning, which provides rewards at the global-level, subject-level, and relationship-level. In addition, we prove that optimal sample diversity is achieved by concentrating stochasticity in the early generation stages. Finally, we propose a new text-to-image benchmark, HiCoPrompt, which challenges text-to-image models to produce images with all the concepts appearing in complex prompts. Experiments on the HiCoPrompt demonstrate the effectiveness of HiCoGen.

References

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[2] 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. *Neural Information Processing Systems*, 35:36479–36494, 2022. 1

[3] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2

[5] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 6

[6] Sangwu Lee, Titus Ebbecke, Will Beddow Erwann Millon, Le Zhuo, Iker García-Ferrero, Liam Esparraguera, Mihai Petrescu, Gian Saß, Gabriel Menezes, and Victor Perez. Flux.1 krea [dev]. https://github.com/krea-ai/flux-krea, 2025. 6, 8

[7] 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 *International Conference on Machine Learning*, 2024. 1, 2, 3, 6

[8] 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, 2021. 1, 4, 6

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 1

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Neural Information Processing Systems*, 35: 24824–24837, 2022. 2

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[12] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, et al. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025. 2

[13] 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. 2, 6

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[15] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[16] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: a unified diffusion model for controllable visual generation in the wild. In *Neural Information Processing Systems*, pages 42961–42992, 2023.

[17] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet ++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025.

[18] Hongji Yang, Wencheng Han, Yucheng Zhou, and Jianbing Shen. Dc-controlnet: Decoupling inter-and intra-element conditions in image generation with diffusion models. *International Conference on Computer Vision*, 2025.

[19] Yuhang Ma, Shanyuan Liu, Ao Ma, Xiaoyu Wu, Dawei Leng, and Yuhui Yin. Hico: Hierarchical controllable diffusion model for layout-to-image generation. *Neural Information Processing Systems*, 37:128886–128910, 2024. 2

[20] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. 2

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2): 3, 2022. 2

[23] Xierui Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 2, 6

[24] Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Computer Vision and Pattern Recognition*, pages 13294–13304, 2025. 6

[25] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal

control for diffusion transformer. In *International Conference on Computer Vision*, pages 14940–14950, 2025. 6

[26] Shaojin Wu, Mengqi Huang, Wenxu Wu, Yufeng Cheng, Fei Ding, and Qian He. Less-to-more generalization: Unlocking more controllability by in-context generation. *arXiv preprint arXiv:2504.02160*, 2025. 2, 6

[27] 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. *Neural Information Processing Systems*, 35:27730–27744, 2022. 2

[28] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024. 2

[29] 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 *Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 2

[30] 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. *Neural Information Processing Systems*, 36: 79858–79885, 2023. 2

[31] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Neural Information Processing Systems*, 36: 53728–53741, 2023. 2

[32] Kelin Yu, Sheng Zhang, Harshit Soora, Furong Huang, Heng Huang, Pratap Tokekar, and Ruohan Gao. Genflowrl: Shaping rewards with generative object-centric flow in visual reinforcement learning. In *International Conference on Computer Vision*, pages 13183–13192, 2025. 2

[33] Ziyu Guo, Renrui Zhang, Chengzhuo Tong, Zhizheng Zhao, Rui Huang, Haoquan Zhang, Manyuan Zhang, Jiaming Liu, Shanghang Zhang, Peng Gao, et al. Can we generate images with cot? let's verify and reinforce image generation step by step. *arXiv preprint arXiv:2501.13926*, 2025. 2

[34] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023. 2

[35] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025. 3

[36] 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. 3

[37] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 3

[38] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Better aligning text-to-image models with human preference. *arXiv preprint arXiv:2303.14420*, 1(3), 2023. 4, 6

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 4

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 4

[41] Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024. 4

[42] ChatGPT. Professional communication. Accessed: 12 July 2024. 5, 6

[43] Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Shin'ichi Satoh. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Computer Vision and Pattern Recognition*, pages 14277–14286, 2023. 6

[44] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Neural Information Processing Systems*, 36:78723–78747, 2023. 6

[45] 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. 6

[46] 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. 6

[47] Yucheng Zhou, Jiahao Yuan, and Qianning Wang. Draw all your imagine: A holistic benchmark and agent framework for complex instruction-based image generation. *arXiv preprint arXiv:2505.24787*, 2025. 6

[48] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 6, 8

HiCoGen: Hierarchical Compositional Text-to-Image Generation in Diffusion Models via Reinforcement Learning

Supplementary Material

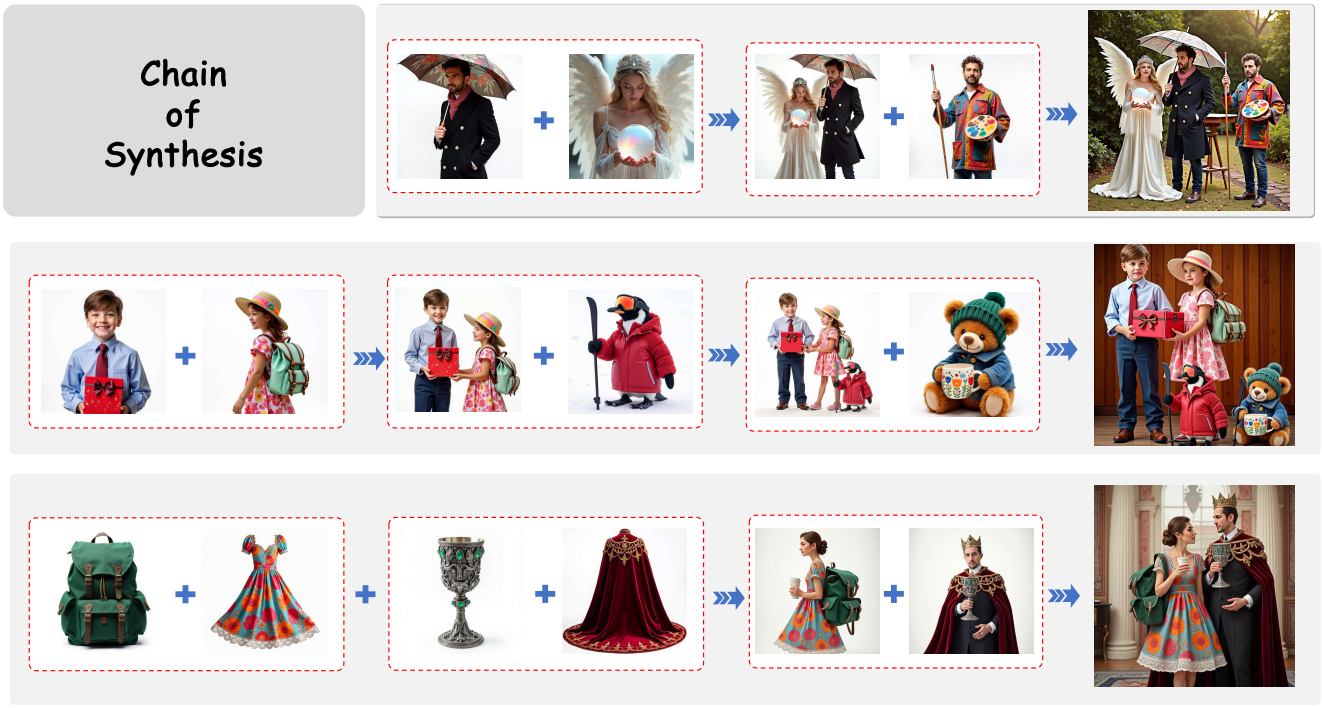


Figure 6. The intermediate results of the Chain of Synthesis.

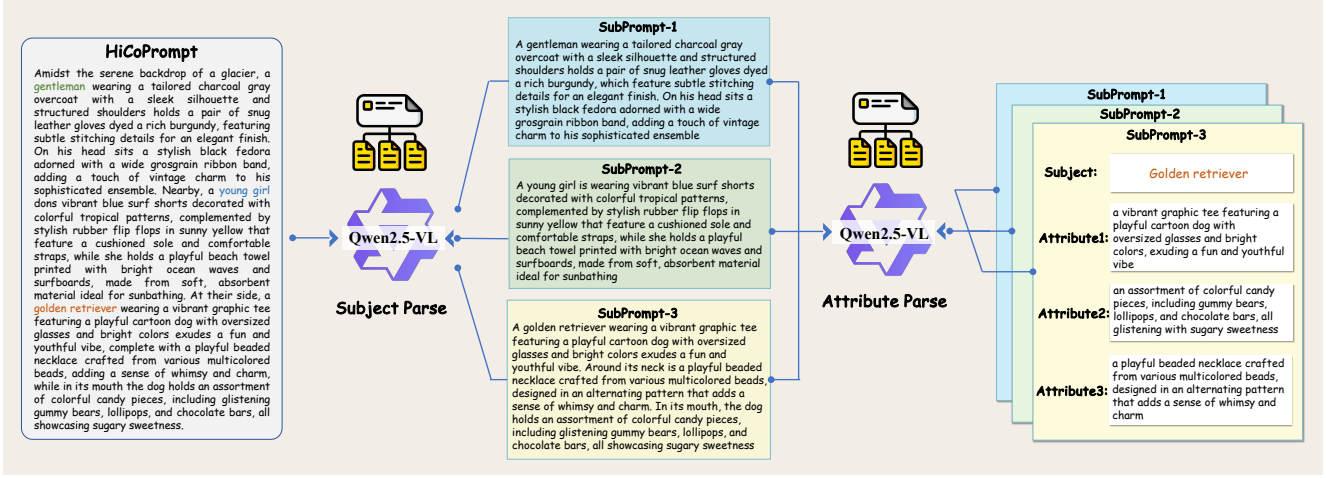


Figure 7. The intermediate prompt of the Chain of Synthesis (zoom in for details).

7. Implementation details

The weights of the reward are set to $w_{\text{clip}} = 0.7, w_{\text{hps}} = 1.4, w_{\text{dino}} = 0.7, w_{\text{vlm}} = 0.7$. The size of the trainset and testset is set to 12,000 and 3,000, respectively. The number of de-noise steps is set to 16 when sampling. The $\eta_{\max}$ and $\eta_{\min}$ are set to 1.0 and 0. The output image size is set to 512×512, with the reference image size being 320×320. The total training step is set to 300 with a batch size of 12.

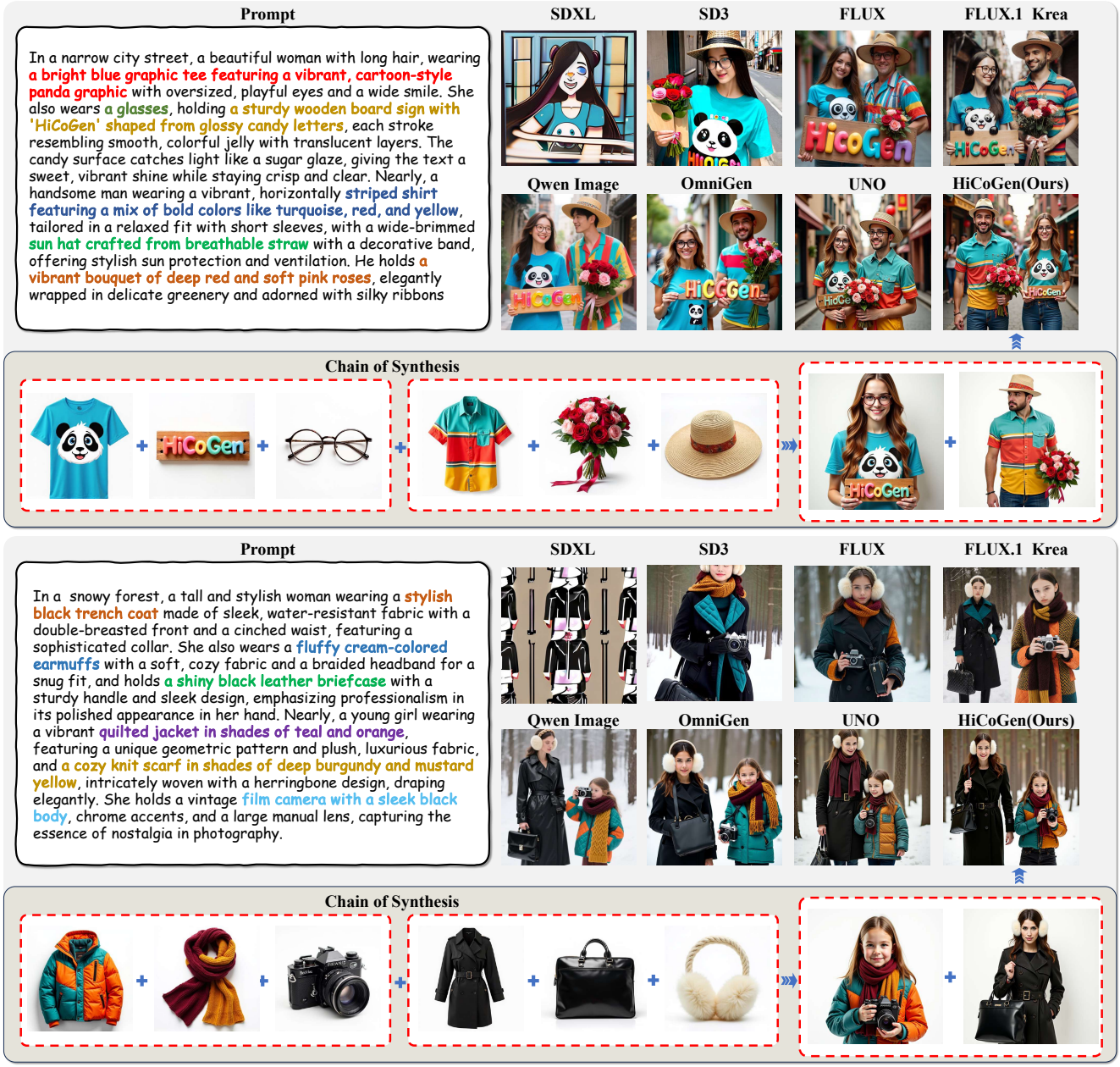


Figure 8. More visual results of the HiCoGen

8. Chain of Synthesis

Fig. 6 shows the case of how the HiCoGen performs Chain of Synthesis. The first row showcases the CoS of three subjects in HiCoGen. It first generates two subjects in one intermediate output, and then includes the third subject in the final output. Besides, we also present the scaling results of four subjects in the second row. The third row shows the result of the HiCoGen on how to generate subjects with hierarchical relationships. For example, it generates clothing and items of the subject, and then generates the final image. Fig. 7 indicates the whole workflow of the parse and rewrite LLM. For a long and complex HiCoPrompt, it is first parsed according to subjects (e.g., gentleman, young girl, and golden retriver). Then, the sub-prompt is continued to be parsed if it contains multiple attributes that can be generated. Thus, we replace the long and complex prompt with a structured prompt composed of several semantic units.

9. More Visual Results

Fig. 8 presents additional visual results produced by HiCoGen, as well as the results from other methods. We also illustrate the CoS process involved in the generation. Through CoS, HiCoGen generates all the concepts from the HiCoPrompt.

Subject Generation

Role:
Please be very creative and generate 20 groups of components for the given character.

Follow these rules:
1. You will be given a <character>, you need to create its wearing, holding, and accessory items.
2. These items must exist in the real world (no fantasy or fictional materials).
3. Do not repeat the same words between outputs; use your imagination and common sense of real life.
4. Each item must contain **no more than two words**.
5. Output multiple unique sets if possible.
6. Given a final complete and brief prompt for text-to-image generation

Output format:
[character]: <character name>
[clothing1]: <real-world clothing>
[holding1]: <real-world object>
[accessory1]: <real-world accessory>

Example:
[character]: dog
[clothing1]: superman's costume
[holding1]: sign
[accessory1]: goggles
[brief_prompt1]: a dog wearing a superman's costume with a goggles, holding a sign

[clothing2]: space suit
[holding2]: book
[accessory2]: glasses
[brief_prompt2]: a dog wearing a space suit with glasses, holding a book
...
(Up to [asset20])

[character]: character

Attribute Rewrite

Role:
Please be very creative and generate a detailed prompt for text-to-image generation.

Follow these rules:
1. You will be given 3 {assets}, you need to create an asset (detailed subject prompt) based on the {assets}.
2. These descriptions can refer only to appearance descriptions/or to certain brands. e.g., "Elon Musk in pajamas", "a tiger in a black hat", "A Mercedes sports car", "A blonde", "A rotten wooden door", "a book with cover written 'Magic'"
3. Do not repeat the same words between outputs; use your imagination and common sense of real life.
4. Describe this asset in one sentence. No more than 40 words
5. Focus on the asset itself, and must contain the asset in each output.
6. Do not describe its background and environment.
7. You may rewrite it based on its contextual meaning within the original prompt.
8. Do not include any users, wearers, or owners. Describe only the object itself.

Example:
[input_asset1]: book
[input_asset2]: lab coat
[input_asset3]: necklace
[original_prompt]: a beautiful woman wearing a lab coat with a necklace and holding a book.

[output_asset1]: an open ancient magic book with a thick dark brown tanned leather cover, adorned with hand-embossed golden runes and intricate patterns
[output_asset2]: a crisp white lab coat with embroidered name and pen-stained pocket
[output_asset3]: a whimsical necklace with animal-shaped pendants

Now, please generate the detailed prompt for the following assets:
[input_asset1]: {subject1}
[input_asset2]: {subject2}
[input_asset3]: {subject3}
[original_prompt]: {ori_prompt}

Prompt Reframing

Role:
You are a professional prompt structure optimization and synthesis expert, skilled at merging multiple prompt segments into a single, well-structured and logically coherent prompt.

Follow these rules:
Generate a unified prompt based on the provided ori_prompt (original prompt) and subprompts (refined sub-prompts). Please strictly follow these rules:
1. Preserve the internal structure, hierarchy, and semantics of each subprompt.
2. You may adjust sentence structures or logical order slightly during merging to ensure overall fluency and coherence.
3. Do not delete or rewrite any key information from the subprompts.
4. You may adapt the language style according to the tone or purpose of the ori_prompt (e.g., analysis, creation, reasoning, Q&A, etc.).
5. The final output should clearly reflect the role and integrated content of each subpart.
6. Output one complete prompt paragraph — no explanations or additional commentary are needed.
7. Do not describe its background and environment.

[input_asset1]: {subject1}
[input_asset2]: {subject2}
[input_asset3]: {subject3}
[original_prompt]: {ori_prompt}

10. Prompt

10.1. Dataset Creation Prompt

In this section, we demonstrate the prompt used to construct the dataset. First, a subject is randomly selected from a predefined character pool. Then, this subject will be assigned up to three attributes. After that, these attributes are rewritten into a more detailed version. Finally, these prompts are reframed into the final prompts by LLMs.

Subject Reward

Role
You are an expert AI assistant specializing in the objective evaluation of the consistency of subjects in two images. Assign a specific integer score to subject. More and larger differences result in a lower score.

Important Notes
- Provide quantitative differences in reason whenever possible.
- Ignore differences in the subject's background, environment, position, size, etc.
- Ignore differences in the subject's actions, poses, expressions, viewpoints, additional accessories, etc.
- Ignore the extra accessory of the subject in the second image, such as hat, glasses, etc.

You must adhere to the output format strictly.

The score ranges from 0 to 4:
- 0: No resemblance. The {subject} in Image2 does NOT appear in Image1 at all. No matching identifiable features.
- 1: Minimal resemblance. The {subject} in Image2 appears only minimally in the {subject} in Image1. Only trivial or coincidental similarities.
- 2: Moderate resemblance. The {subject} in Image2 partially matches the {subject} in Image1. Some major features match, but key identity traits differ.
- 3: Strong resemblance. The {subject} in Image2 is very similar to the {subject} in Image1. Most identity traits align, with minor differences.
- 4: Near-identical. The {subject} in Image2 is identical to the {subject} in Image1. Nearly identical; same character.

Output only the brief reason and score. DO NOT OUTPUT any other text.

Image1: [brief description of Image1]
Image2: [brief description of Image2]
Reason: [Brief Reason]
Output: [Score]

Relation Reward

Role:
You are a visual consistency evaluator. You will compare Image2 and Image1.

Your task is to check two things:
1) Content Preservation:
Are the key visual elements from Image2 present in Image1? Ignore differences in the subject's background, environment, position, size, etc.
Examples of key visual elements:
- For a character: clothing type, accessories.
- For clothing: color, pattern, shape, length, structure.
- For a held object: object type, shape
2) Correct Assignment:
The elements of the character from Image2 must appear on the same character in Image1.
They must not be assigned to another character or misplaced.

Then classify overall similarity using the scale below:

0 - Completely mismatched: Key elements are missing or assigned to the wrong subject. No meaningful match.
1 - Minimal similarity: Only small or generic resemblance. Most defining elements are missing or incorrectly reassigned.
2 - Partial similarity: Some important features match, but major elements are missing, altered, or incorrectly assigned.
3 - High similarity: Most key features are present and assigned correctly, with only minor inaccuracies.
4 - Fully consistent: All key features are preserved and assigned correctly. No confusion or mixing.

Output only the following format:

Image1: [brief description of Image1]
Image2: [brief description of Image2]
Reason: [brief explanation of match or mismatch]
Output: [0-4]

10.2. Reward Prompt

Accuracy of Existing

Check whether a specific object appears in the image.

Instruction:
Analyze the provided image carefully. Determine whether the specified item is present or visible in the image.

Object: {user_prompt}

Output format:
Answer "Yes" if the object is clearly present.

Answer "No" if the object is not visible. Do not assume or infer its presence based on context.

If uncertain, answer "Unclear" and briefly explain why.

Output only the brief description, reason and answer. DO NOT OUTPUT other text.

Description: [Brief Description of the Image]
Reason: [Brief Reason]
Output: [Yes / No / Unclear]

Accuracy of Attribute

Check whether the object in the image matches the attributes described in the given text.

Instruction:
Analyze the image carefully. Determine whether the object matches the specific details (such as color, shape, material, pattern, size, part structure, or other explicit attributes) described in the text. Object description: "{user_prompt}"

Output format:

Answer "Yes" if the object is clearly present and its visible attributes match the given text.

Answer "No" if the object is present but clearly does not match the described details, or the described object is not visible at all.

If uncertain, answer "Unclear" and briefly explain why.

Description: [Brief Description of the Image]
Reason: [Brief Reason]
Output: [Yes / No / Unclear]

Accuracy of Attribute

Check whether the specified object in the image interacts with the main subject in a correct and reasonable way according to the following rules.

Rules:
The object mentioned in the prompt must interact with the main subject in a reasonable, physically consistent way (no floating, no paste-on appearance).
The object must not appear in the hands or possession of another subject.
The object's size and proportion relative to the main subject must be correct and not distorted.

Instruction:
Analyze the provided image carefully and judge whether the described object satisfies all the above rules.
Object: "{user_prompt}"

Output format:
Output "Yes" if the object is visible and satisfies all three rules.
Output "No" if any rule is violated or the object is not visible.

Description: [Brief Description of the Image]
Reason: [Brief Reason]
Output: [Yes / No / Unclear]

10.3. Evaluation Prompt