arXiv:2603.02712v1 [cs.CV] 3 Mar 2026

从“绘制什么”到“如何绘制”：用于自回归图像生成的受约束推理

Ruxue Yan[1] Xubo Liu[1] Wenya Guo[1] Zhengkun Zhang[2] Ying Zhang[1] Xiaojie Yuan[11]
南开大学计算机科学学院[2]百度公司**通信作者：**yanruxue@dbis.nankai.edu.cn

摘要

近年来，随着思维链和强化学习的引入，自回归图像生成取得了进展。然而，当前方法仅通过重写输入提示来指定要描绘的“绘制什么”细节，却从根本上未能推理如何构建整体图像。这一固有局限导致了持续存在的问题，例如空间歧义会直接造成不真实的物体重叠。为弥补这一不足，我们提出***CoR-Painter***，这是一种开创性的框架，通过引入受约束推理来引导自回归生成，开创了“如何到什么”范式。具体而言，该方法首先从输入提示中推导出一组视觉约束，以确定“如何绘制”；这些约束明确规定空间关系、关键属性和构图规则。随后，这些约束引导生成详细描述（“绘制什么”），为准确的视觉合成提供结构合理且连贯的依据。此外，我们提出双目标 ***GRPO***策略，专门优化文本受约束推理和视觉投影过程，以确保整个生成流程的连贯性和质量。在 *T2I-CompBench*、*GenEval* 和 *WISE* 上进行的大量实验表明，我们的方法达到了最先进的性能，空间指标显著提升（例如，在 *T2I-CompBench* 上提升 *+5.41%*）。

1. 引言

文生图已成为计算机视觉领域的一个重要研究方向。大语言模型（LLM）[31, 33, 41]的成功推动了通过下一词元预测实现自回归视觉生成的发展。与此同时，思维链（CoT）推理 [37] 也被拓展到这一领域：相关方法首先生成对输入提示的详细描述，然后通过强化学习（RL）优化生成的描述 [25, 28, 29]，

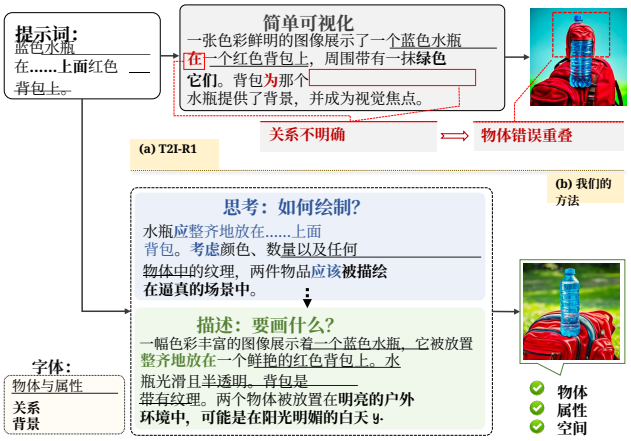


图 1. （a）基于思维链的方法 T2I-R1 [15], 和（b）我们的 CoR-Painter 的结构与输出示意图。

提升输出质量和语义一致性。

现有的基于思维链的方法 [6, 15, 36] 主要只是将输入提示改写为更详细的描述，以此引导图像生成。然而，这一过程仅仅着眼于扩展“描述什么”的细节，却缺少对“如何”连贯地构建整体图像所必需的推理。这不可避免地带来一个关键缺陷：无法建立一致的生成蓝图，导致局部看似合理的细节在全局层面相互冲突。如图1(a)所示，T2I-R1 [15] 能够生成准确的细节，例如水瓶和背包的材质、形状与风格，但缺少一套连贯的空间约束，无法将这些元素组织成逻辑一致的图像。尽管“在……上”和“为……提供背景”等单个关系单独来看都合理，但由于缺乏整体结构约束，物体的**空间关系解释变得模糊**。描述中的这种模糊性会直接传递到图像合成过程中。为了同时满足这些模糊的空间线索，

模型会受到误导，生成多个冗余物体，并以不合理的方式相互重叠，如图1(a)所示——这是明显的**视觉伪影**，严重损害了图像的语义完整性。

为了解决这一问题，我们提出了**CoR-Painter**，一种约束推理图像生成框架，其采用“如何到什么”范式，明确引入逻辑连贯、空间组织良好的约束来指导详细描述。受人类绘画过程的启发——艺术家会先确定构图和布局，再添加细节——CoR-Painter 首先明确“如何”组织图像的整体内容与空间结构，从而形成一组关于焦点和排列方式的明确约束。随后，这些约束会引导对“什么”细节进行推理，最终生成结构良好、内容详尽的视觉描述，作为图像生成的直接映射。在这一“如何到什么”范式的驱动下，如图1(b) 所示，**CoR-Painter 首先输出基于约束的指令，**例如“应整齐地放在……上方”以定义空间关系，“考虑颜色、数量和纹理”以突出关键视觉特征，以及“应描绘在户外场景中”以设定背景。在这些约束的引导下，模型**随后生成细致入微的描述**：例如，“一个蓝色水瓶整齐地放在鲜艳的红色背包上”，~~并进一步描述~~物体外观，例如“光滑、半透明的水瓶”和“有纹理的背包”，整体场景设定为“明亮的户外环境，可能是晴天”。这一结构化的受约束文本推理过程确保描述既具有全局一致性，又包含局部细节，从而有效消除空间歧义和物体错位。因此，生成图像具有更高的语义准确性和视觉连贯性，这一点可从图1(b) 中正确的空间排列和更出色的背景描绘得到印证。

在模型优化方面，我们采用**双**目标**GRPO**（**DO-GRPO**）策略，专门优化两个核心目标：1）第一个目标是强化文本推理过程中的语义对齐与逻辑连贯性，确保模型准确理解并组织输入提示中的关键信息。2）第二个目标是增强图像生成过程中的视觉—文本一致性，引导模型在合成过程中忠实再现详细描述与空间布局。这种双目标优化有效协调推理过程与生成过程，从而显著提升最终输出的语义完整性与质量。我们的贡献可总结为：

•我们提出了 CoR-Painter，这是一种“如何到什么”的框架，通过优先考虑结构约束（“如何”），再进行视觉细节刻画（“什么”），以确保全局一致性。

• 我们引入双目标 GRPO，分别提供专门的奖励，以强化文本推理的准确性与图像生成的保真度。

• 在 T2I-CompBench、GenEval 和 WISE 上进行的大量实验展现了新的最先进性能，尤其是在增强空间关系方面。

2. 相关工作

2.1. 图像生成中的自回归模型

受大语言模型中自回归生成成功的启发，[31, 33, 41]，自回归（AR）范式逐渐应用于图像生成：将图像离散化为序列化令牌，并通过下一词元预测逐步生成，以产出高质量图像。LlamaGen [1] 是早期探索之一，表明自回归模型能够在图像生成中取得具有竞争力的结果。后续研究主要聚焦于提升效率、分辨率和语义一致性。AiM [17]利用 Mamba 架构优化长序列建模并加速推理；Token-Shuffle [20] 减少令牌数量，以支持更高分辨率的生成；CTF [11] 采用由粗到细的令牌预测来提升质量；GigaTok [40]使用语义正则化来处理复杂的视觉词元器。Open-MAGVIT2 [19] 通过超大型码本和子词元预测提升语义一致性。较新的研究提出了下一尺度/Next-X [26]预测框架，以及基于大规模连续令牌的自回归模型（NextStep-1） [30], ，显著提升了生成效率、质量和高分辨率生成能力。然而，在优化中间推理以及确保复杂场景中的语义一致性方面，仍存在挑战。

2.2. 图像生成中的思维链推理与强化学习

基于大语言模型中的思维链推理，研究人员将推理与强化学习（RL）融入自回归图像生成，并进一步发展出用于引导生成过程的结构化框架。BiCoT-GRPO [15] 将语义级和词元级推理与生成奖励相结合。PARM 和 PARM++ [12]通过利用潜力评估奖励和反思机制进行逐步评估，增强了自回归生成。GoT [9] 和 GoT-R1 [6] 将语义-空间推理与基于强化学习的奖励相结合，以提升构图对齐和空间对齐。研究还聚焦于奖励机制和训练方法，以强化推理引导的生成。SUDER [13] 和 CoRL [16]探索自监督双重奖励与协同强化学习，以实现多模态优化。FocusDiff [22] 利用强化学习实现细粒度文本图像对齐，处理语义细微差异，

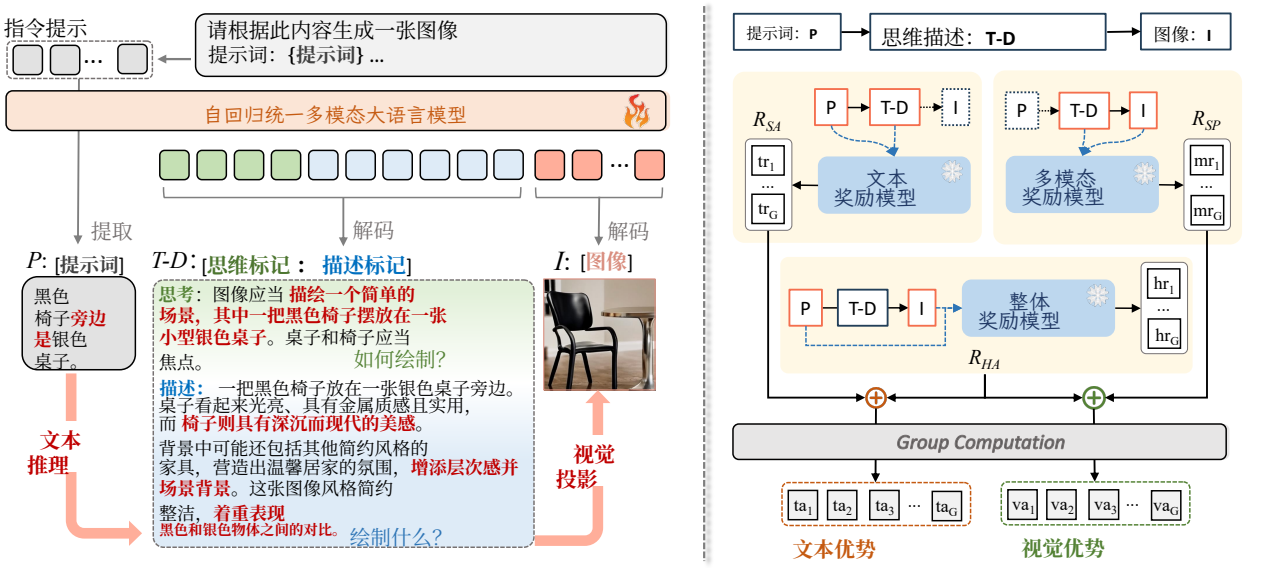


图 2. CoR-Painter 概览：(a) 文生图生成过程示意图；(b) 双目标 GRPO。 R_{SA}、 R_{SP} 和 R_{HA} 分别表示语义锚定奖励、语义投影奖励和整体对齐奖励。

尽管取得了这些进展，现有方法很少将文本推理与图像生成分开进行优化。我们的方法通过引入目标专属奖励进行专门优化，同时保持跨模态一致性，从而解决了这一问题。

3. 方法

本节将详细介绍 CoR-Painter：首先说明生成过程，然后概述如何通过强化学习训练模型，以生成高质量图像。

3.1. 图像生成流程

如前所述，我们的生成过程遵循图像生成的“如何到什么”范式。给定一个输入提示，我们依次从“如何绘制”和“什么要绘制”两个方面进行文本推理，生成约束引导指令和结构化逻辑描述，再将其映射到图像生成过程中，从而成为语言理解与最终视觉渲染之间的桥梁。该过程的流程如图2(a)所示。

我们使用 Janus-Pro [5] 作为底部模型（即，“自回归统一多模态大语言模型”），在共享表示空间中联合建模文本和图像 token。为了引导模型进行生成，我们基于原始提示词为多模态大语言模型设计了一条指令提示。该指令提示通过引入特定标签来定义两个组件，从而指定文本推理内容的结构： <思考></thought> 部分以及 <描述></description>部分。这一设计确保生成的推理与“如何绘制”和“绘制什么”的过程相一致。构建的指令提示如下：

原始提示词的指令提示

请根据以下提示词生成图像：“{提示词}”。让我们一步步思考。<thought> {列出提示词中提及的所有物体、属性和空间关系，并注明它们应遵循的任何约束。} </thought><description> {简要、准确地描述提示词中的所有元素应如何呈现。} </description>

遵循上述指令提示，模型首先根据原始提示词进行全局分析，以确定图像的整体构图和语义结构；随后针对主要物体、其属性及空间关系生成由约束引导的指令，作为思维标记，明确说明它们应当“如何”呈现。这种明确的指导为生成描述“描绘什么”的标记提供了全局布局信息先验，使模型能够以结构化的方式逐步完善细节，并生成能够呈现物体外观、相对位置和场景氛围的描述性文本。随后，这段详细且合乎逻辑的描述会映射到图像生成过程中。这种先进行明确思考、再进行详细描述的过程称为**文本推理**，

由此生成子序列 d。接下来，添加图像生成起始标记 <img start>，模型再基于 d生成图像 I。这一过程称为**视觉投影**。随后，针对这两个过程分别进行优化，具体如下。

3.2. 双目标 GRPO

群体相对策略优化（GRPO）。GRPO 是一种强化学习框架，通过引入基于群体的相对优势，而非显式价值估计，来增强语言模型的推理能力。给定提示词 p，模型会采样 G个响应 $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|p)$，并分别通过标量奖励 $\mathcal{R}_i$进行评估。每个样本的优势会在群体内进行归一化：

$$A_i = \frac{\mathcal{R}_i - \text{mean}(\{\mathcal{R}_i\}_{i=1}^{G})}{\text{std}(\{\mathcal{R}_i\}_{i=1}^{G})}. \tag{1}$$

随后，通过带有 KL 正则化的裁剪式 PPO 目标函数对策略进行优化：

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\big(r_i(\theta)A_i,\\ \text{clip}(r_i(\theta), 1-\varepsilon, 1+\varepsilon)A_i\big)\\ -\beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg],\end{aligned} \tag{2}$$

其中，$r_i(\theta) = \frac{\pi_\theta(o_i|p)}{\pi_{\theta_{\text{old}}}(o_i|p)}$，而 β 控制正则化强度。

双目标 GRPO GRPO 最初是为单模态文本推理设计的，而我们的任务需要对两种异构模态——文本推理和图像生成——进行联合优化。为此，我们将 GRPO 框架扩展为一种双目标强化学习策略，其中每个响应 o_i 由一段文本推理序列 s_i 和一个图像序列 t_i组成，即 $o_i = (s_i, t_i)$。文本推理过程对应**文本推理**，生成中间思考和描述 d；图像生成过程对应**视觉投影**，根据 d 生成图像令牌。对于每个生成过程，逐令牌概率比定义如下：

$$r_{i,j}(\theta) = \begin{cases} \dfrac{\pi_\theta(s_{i,j}|p, s_{i,<j})}{\pi_{\text{old}}(s_{i,j}|p, s_{i,<j})}, & 0 \le j \le |s_i|, \\ \dfrac{\pi_\theta(t_{i,j}|p, s_i, t_{i,<j})}{\pi_{\text{old}}(t_{i,j}|p, s_i, t_{i,<j})}, & |s_i| < j \le |s_i| + M. \end{cases} \tag{3}$$

这一分段比例确保图像令牌的生成明确地以完整的文本推理链为条件，从而维持跨模态的语义锚定。

目标特定奖励设计。基于上述设计，双目标 GRPO 引入模态特定奖励信号，以指导每个生成过程，同时通过维持共享的优势归一化来实现连贯优化。如图2 （b）所示，在文本推理过程中，我们应用语义锚定奖励（R_{SA}），由文本奖励模型计算，用于衡量生成的推理链相对于原始提示词的语义一致性和逻辑忠实度。在图像生成过程中，语义投影奖励（R_{SP}）由多模态奖励模型提供，用于评估文本推理在多大程度上被投射到视觉域中。此外，整体对齐奖励（R_{HA}）提供全局监督信号，以捕捉原始提示词与生成图像之间的语义对齐。在实践中，该奖励由三个独立模型计算，分别评估对象属性、空间关系和整体场景语义。为展示整体奖励，流程图将其表示为一个整体奖励模型，而各个方面的计算将在后续章节中详细介绍。此外，过程级优势在每个采样组内计算如下：

$$\begin{cases} A_i^{\text{text}} = \dfrac{(R_{\text{SA}} + R_{\text{HA}}) - \text{mean}(\{R_{\text{SA}} + R_{\text{HA}}\}_{i=1}^{G})}{\text{std}(\{R_{\text{SA}} + R_{\text{HA}}\}_{i=1}^{G})}, \\ A_i^{\text{img}} = \dfrac{(R_{\text{SP}} + R_{\text{HA}}) - \text{mean}(\{R_{\text{SP}} + R_{\text{HA}}\}_{i=1}^{G})}{\text{std}(\{R_{\text{SP}} + R_{\text{HA}}\}_{i=1}^{G})}, \end{cases} \tag{4}$$

其中，A_i^{text}和 A_i^{img} 分别表示文本推理和视觉投影过程的优势。

优化目标。双目标 GRPO 的最终目标是针对各自的优势，以适合每个过程的方式进行优化：

$$\begin{aligned}&\mathcal{J}_{\text{Dual}}(\theta) = \\ &\mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\Big(\sum_{j=1}^{|s_i|}\min\big(r_{i,j}A_i^{\text{text}}, \text{clip}(r_{i,j}, 1-\varepsilon, 1+\varepsilon)A_i^{\text{text}}\big)\\ &+\sum_{j=|s_i|+1}^{|s_i|+M}\min\Big(r_{i,j}A_i^{\text{img}}, \text{clip}(r_{i,j}, 1-\varepsilon, 1+\varepsilon)A_i^{\text{img}}\Big)\Big)\\ &-\beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg].\end{aligned} \tag{5}$$

通过双目标设计，文本推理和视觉投影分别根据各自的优势进行优化，在提升语义连贯性和视觉保真度的同时，使推理与生成保持一致。

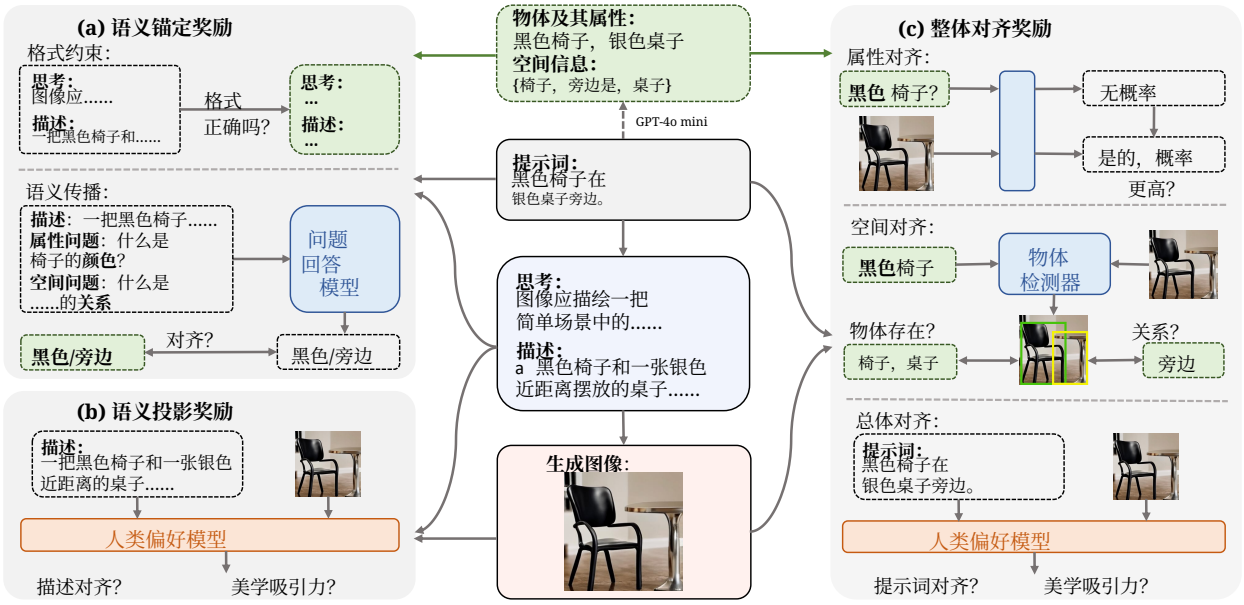


图3. 奖励计算概览。

3.3. 奖励计算

如上所述，我们的奖励设计包含三个组成部分，分别针对生成的不同方面。我们使用 GPT-4o-mini 从提示词中预先提取关键物体、属性、空间关系和数量，将其作为确定性参照，从而精确评估模型对提示词核心语义的保留程度。

语义锚定奖励（SAR）用于确保文本推理过程中的语义保真度。如图3（a）所示，格式奖励通过检查是否提取了思考和描述标签，确保模型执行“如何绘制”推理；若两者均存在，则给予较高分数，若仅存在其中一个，则给予较低分数。公式如下：

$$R_{\text{Format}}(d_i) = w_1 \cdot f_{\text{thought}} + w_2 \cdot f_{\text{description}}, \tag{6}$$

其中，f_{thought} 和 $f_{\text{description}}$ 表示是否提取了有效标签，权重 w_1和 w_2 均设为0.5。

对于语义传播，预提取元素会转换为问题，再由问答模型（Llama2- 7b-chat [32]）作答，奖励根据其与参考答案在词元层面的一致性进行衡量：

$$R_{\text{Prop}}(d_i) = \frac{1}{K}\sum_{k=1}^{K} \text{match}(\text{QA}(d_i, q_k), a_k), \tag{7}$$

其中 q_k 和 a_k 表示第 k个问题及参考答案，match(·)计算词重叠度或词元级相似度分数。

语义锚定奖励 R_{SA} 由格式奖励和语义传播奖励组成：

$$R_{\text{SA}}(d_i) = R_{\text{Format}}(d_i) + R_{\text{Prop}}(d_i), \tag{8}$$

语义投影奖励（SPR）用于评估详细描述 d_i 在生成图像中的实现忠实度 I_i，以及图像的美学质量。如图3（b）所示，我们使用人类偏好模型（HPSv2 [2]）来计算语义对齐度和美感：

$$R_{\text{SP}}(d_i, I_i) = \text{HPM}(d_i, I_i), \tag{9}$$

其中 HPM(·) 表示 HPSv2 给出的分数

整体对齐奖励（HAR）。用于评估提示词 p 与生成图像 I_i之间的整体跨模态对齐情况，并结合三项评估。首先，视觉问答模型（GIT） [34] 会检查根据提示词生成的属性问题是否与图像内容相符：

$$R_{\text{VQA}}(p, I_i) = \frac{1}{K}\sum_{k=1}^{K} P(\text{QA}(I_i, q_k) = a_k^{\text{ref}}), \tag{10}$$

其中，q_k 和 a_k^{ref}分别是第 k 个属性问题及其参考答案，K 是问题总数。其次，GroundingDino [18]检测提示词中的对象，并

表1. 在 T2I-CompBench 和 GenEval 数据集上的结果。

方法	T2I-CompBench						GenEval						
	属性绑定			物体关系。		整体复杂度							
	颜色	形状	纹理	空间	非空间		单个 Obj.	Two Obj.	计数	颜色	位置	颜色属性	
基于扩散的方法													
SD-v1.5 [27]	37.58	37.13	41.86	16.27	31.12	32.16	0.43	0.97	0.38	0.35	0.76	0.04	0.06
SD-XL [24]	58.79	46.87	52.99	29.57	31.19	33.84	0.55	0.98	0.74	0.63	0.67	0.34	0.36
PixArt-α [4]	66.90	49.27	64.77	20.64	**31.97**	34.33	0.48	0.98	0.50	0.44	0.80	0.08	0.07
SD3 [7]	80.94	58.64	72.97	54.55	31.43	38.91	-	-	-	-	-	-	-
Stable v3 [8]	81.32	58.85	73.34	56.02	31.40	39.00	-	-	-	-	-	-	-
基于*AR*的方法													
LlamaGen [1]	42.02	39.67	51.03	19.75	30.50	30.81	0.32	0.71	0.34	0.21	0.58	0.07	0.04
Show-o [39]	56.00	41.00	46.00	43.82	30.00	34.10	0.68	0.98	0.80	**0.66**	0.84	0.31	0.50
Infinity [3]	73.79	46.50	59.19	40.67	30.76	38.06	0.73	-	0.85	-	-	0.49	0.57
Janus-Pro-7B [5]	63.59	35.28	49.36	35.96	30.85	37.42	0.80	0.99	0.89	0.59	0.90	0.79	0.66
DDT-LLaMA [23]	72.80	51.40	64.20	-	-	-	0.66	0.99	0.64	0.56	0.87	0.39	0.48
基于*AR*的方法 + 强化学习													
Show-o+PARM [12]	75.00	56.00	66.00	46.58	31.00	38.02	0.69	0.97	0.75	0.60	0.83	0.54	0.53
EMU3 [35]	75.44	57.06	71.64	52.33	30.84	37.00	0.64	0.99	0.76	0.38	0.85	0.45	0.40
GoT-R1-7B [6]	81.39	55.49	73.39	51.27	31.69	40.45	0.75	0.99	0.94	0.50	0.90	0.46	0.68
T2I-R1 [15]	81.30	58.52	72.43	57.13	30.90	39.93	0.79	0.99	0.91	0.53	0.91	0.76	0.65
Janus-FocusDiff-7B [22]	83.00	60.30	72.80	58.43	30.99	40.98	0.85	0.99	0.95	0.63	0.93	0.85	0.75
CoR-Painter（我们的方法）	**84.47**	**61.94**	**76.93**	**63.84**	31.44	**41.18**	**0.87**	**1.00**	**0.96**	0.63	**0.93**	**0.90**	**0.77**

检查它们是否存在，以及它们在图像中的空间位置是否正确：

$$R_{\text{Det}}(p, I_i) = \frac{1}{N_o} \sum_{m=1}^{N_o} \text{match}(\text{Detect}(I_i, \text{obj}_m), \text{obj}_m^{\text{ref}}), \tag{11}$$

其中 q_k 和 a_k^{ref}分别是第 k个属性问题及其参考， K 是问题的总数量。最后，评估提示词与图像之间的整体语义与美感对齐：

$$R_{\text{Align}}(p, I_i) = \text{HPM}(p, I_i), \tag{12}$$

其中HPM(·)表示人类偏好模型（HPSv2）给出的分数。形式上，集成奖励的计算方式为：

$$R_{\text{HA}}(p, I_i) = R_{\text{VQA}} + R_{\text{Det}} + R_{\text{Align}}. \tag{13}$$

4. 实验

4.1. 主要结果

我们在 3 个广泛使用的基准测试上进行了实验：T2I-Compbench [14], GenEval [10] 和 WISE [21]。结果如下：

对于 T2I-Compbench 数据集 [14], ，我们评估生成图像在属性、物体关系和复杂场景方面的表现。将原始 UniDet [42] 替换为

表 2. WISE 数据集上的结果。

方法	文化	时间	空间	生物	物理	化学	全部
VILA-U [38]	0.26	0.33	0.37	0.35	0.39	0.23	0.31
Show-o [39]	0.28	0.4	0.48	0.30	0.46	0.30	0.35
EMU3 [35]	0.34	0.45	0.48	0.41	0.45	0.27	0.39
Janus-Pro-7B [5]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
T2I-R1 [15]	0.56	0.55	0.63	0.54	0.55	0.30	0.54
CoR-Painter（我们的方法）	**0.61**	**0.56**	**0.66**	**0.55**	**0.64**	**0.33**	**0.58**

开放词汇 GroundingDINO [18] 后，结果显著提升——例如，T2I-R1 的空间分数从 33.78% 升至 57.13%，证明了检测器的有效性。如表1所示，我们的 CoR-Painter 在几乎所有类别上都优于现有方法，其中空间关系指标提升了 5.41%，表现出强大的布局推理能力和精准的物体定位能力。非空间分数略低于 PixArt-α (↓0（0.0025），但考虑到 CLIP 分数相近（0.30–0.31），这一差距可以忽略不计。总体而言，这些结果充分证明了我们的生成式推理链与 DO-GRPO 相结合的有效性。

在 GenEval数据集上，我们在多种任务上评估模型，包括单物体、双物体、计数、颜色、位置和属性绑定场景。

表 3. 推理链结构和奖励设计的消融实验。

组件	属性绑定			对象关系		复杂
	颜色	形状	纹理	空间	非空间	
Janus-Pro-7B（基线） [5]	63.59	35.28	49.36	35.96	30.85	37.42
我们的方法的完整链	**84.47**	**61.94**	**76.93**	**63.84**	**31.44**	**41.18**
(a) 不含思考（“如何绘制”）	81.14	58.04	74.92	60.85	30.90	40.32
(b) 不含描述（“绘制什么”）	78.99	53.00	69.03	54.66	30.53	39.02
我们的方法-完整奖励	**84.47**	**61.94**	**76.93**	**63.84**	**31.44**	**41.18**
(c) 不含 SAR：HAR⇔SPR+HAR	83.46	61.17	75.23	62.89	30.57	40.94
(d) 不含 SPR：SAR+HAR⇔HAR	82.25	60.58	74.85	61.25	29.85	40.08
(e) 不含 SAR 和 SPR：HAR⇔HAR	81.31	59.38	73.85	59.89	29.27	39.51
我们的方法-Full HAR	**84.47**	**61.94**	**76.93**	**63.84**	**31.44**	**41.18**
(f) 无视觉问答	81.95	58.65	73.42	63.27	30.92	39.96
(g) 无检测	82.00	60.25	72.90	58.38	30.99	39.47
(h) 无HPM	81.74	59.38	73.85	60.89	29.79	40.02

如图 1所示，CoR-Painter 在整体表现上达到最先进水平，尤其是在空间定位任务上，比此前最先进的方法 Janus-FocusDiff 高出 5%。我们的方法更注重全局布局连贯性，而非严格的数量控制，因此计数指标有所波动，略低于 Show-o。总体而言，CoR-Painter 在该基准测试上表现出色，能够准确安排物体位置，同时保持语义和视觉一致性。

在WISE数据集上 [21], 文本提示需要领域知识和常识推理，包括时间、地理和自然科学，才能生成准确的图像。这对模型的推理能力提出了更高要求。以往模型的焦点在于将提示词改写为详细描述，但对于需要知识推理的隐式提示，它们无法推断出需要扩展的物体，从而导致性能不佳。相比之下，CoR-Painter通过约束引导过程推断目标对象，从而实现正确的图像生成。如表2所示，CoR-Painter优于现有方法，证明了其在知识驱动图像生成中的有效性。

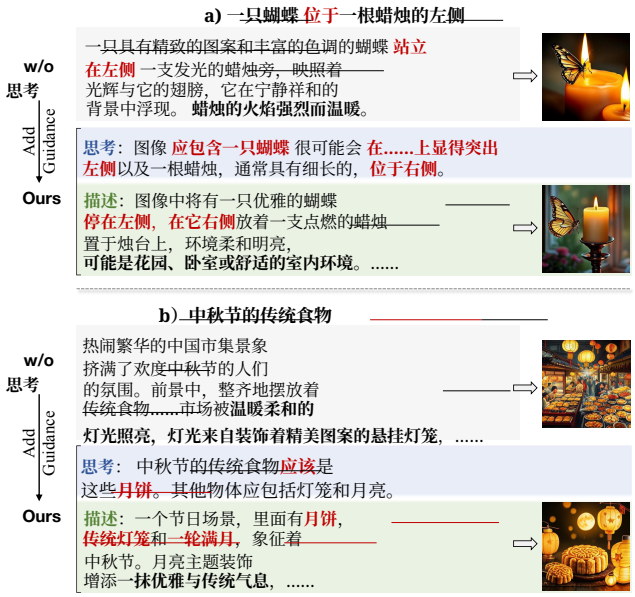


图 4. 融入思维过程前后，描述与生成图像的对比。

4.2. 消融研究

为了进一步验证“如何到什么”范式和 DO-GRPO 的有效性，我们对推理链结构和奖励设计进行了消融实验。

推理链结构。如表 3 （a–b 行）所示，去除思维过程会降低语义准确性，而省略描述过程则会降低 per-

由于缺乏结构化映射而导致的保真度下降。两种组成部分对于语义保真度和空间一致性都至关重要。图4 中的可视化表明，思考过程阐明了物体属性和关系，例如，蝴蝶和蜡烛——并改进了对隐含物体的推理，例如“中秋节的传统食物应该是月饼”，从而生成更加完整且一致的输出图像。

奖励设计。如表 3 （c）–（e）所示，移除 SAR 会削弱语义迁移，同时影响属性和空间性能；而移除 SPR 则会降低文本-图像对齐。当二者都被移除、仅保留 HAR 时，性能会显著下降。对单个 HAR 奖励的消融实验（第（f）–（h）行）同样显示出明显下降，进一步凸显了每种奖励的重要性。定性



图5. CoR-Painter与近期方法在典型案例上的定性比较。

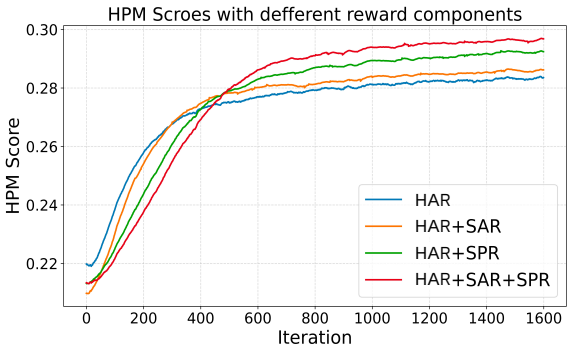


图6. 不同奖励消融实验下的HPM分数。

分析 HPM 奖励可知，它反映了提示词-图像对齐和美感。如图 6所示，同时使用 SAR 和 SPR 时，HPM 得分最高；移除任一奖励后，得分都会下降。这些结果表明，双目标设计能够有效优化文本和图像生成，提升推理能力、图像质量以及空间关系。

4.3. 案例研究

我们通过分析案例，直观展示了我们的方法相较于三种近期方法的有效性：Janus-Pro（我们的基线）、T2I-R1（此前最先进的基于思维链的方法）和 Janus-FocusDiff（此前最先进的基于自回归的方法）。如图 5 所示，Janus-Pro 无法生成部分目标对象，导致图像不完整，例如在案例（b）中只生成了花瓶。T2I-R1 则能够生成丰富的视觉细节，但存在布局不合理和物体数量错误的问题，例如，案例（b）中人物与花瓶之间的空间关系有误，案例（c）中生成了两瓶葡萄酒而非一瓶。Janus-Focus 对基本细节处理得很好，但在复杂场景中表现不佳，例如，案例（e）中柜子的颜色不正确，案例（f）中物体数量也有误。相比之下，我们的 CoR-Painter 首先为全局元素及其位置生成约束，以指导后续的详细描述。这使其能够精确生成属性，例如案例（a）和（c）中物体的颜色和纹理，并形成连贯的空间布局，例如案例（b）中人物和花瓶的位置。它在包含多个物体的复杂场景中也表现出色，如案例（e）；对于物体数量和颜色各异的情况，也能准确处理，例如案例（f）中的山羊和绵羊。

5. 结论

在本文中，我们提出了 CoR-Painter，这是一种开创“如何到什么”范式的新型框架，旨在增强文生图。该方法首先生成关于“如何绘制”的指令性约束，然后推理“绘制什么”，从而生成结构化且详细的视觉描述，有效解决不明确的空间关系和不对齐的物体位置问题。我们进一步引入 DO-GRPO，分别优化文本生成和图像生成，提升整体质量与语义一致性。在多个基准测试上的大量实验表明，我们的方法始终优于现有方法，证明了其生成高质量、语义对齐图像的有效性。

参考文献

[1] 用于可扩展图像生成的 Llama。arXiv:2406.06525。2, 6[2] 人类偏好分数 v2。arXiv:2306.09341。5[3] Infinity：扩展按位自回归建模以实现高分辨率图像合成。在 CVPR，2025。6[4] Junsong Chen、Jincheng Yu、Chongjian Ge、Lewei Yao、Enze Xie、Yue Wu、Zhongdao Wang、James Kwok、Ping Luo、Huchuan Lu 等。Pixart-α：用于逼真文生图合成的快速扩散变换器训练。arXiv 预印本 arXiv:2310.00426，2023。6[5] Xiaokang Chen、Zhiyu Wu、Xingchao Liu、Zizheng Pan、Wen Liu、Zhenda Xie、Xingkai Yu 和 Chong Ruan。Janus-pro：通过数据与模型扩展实现统一多模态理解与生成。arXiv 预印本 arXiv:2501 17811，2025。3, 6, 7 。 。[6] Chengqi Duan、Rongyao Fang、Yuqing Wang、Kun Wang、Linjiang Huang、Xingyu Zeng、Hongsheng Li 和 Xihui Liu。GoT-R1：通过强化学习释放多模态大语言模型的视觉生成推理能力。arXiv 预印本 arXiv:2505.17022，2025。1, 2, 6[7] Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas M¨uller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel、Dustin Podell、Tim Dockhorn、Zion English 和 Robin Rombach。扩展用于高分辨率图像合成的整流流变换器。在第 41 届国际机器学习大会论文集，2024。6[8] Patrick Esser、Sumith Kulal、Andreas Blattmann、Rahim Entezari、Jonas M¨uller、Harry Saini、Yam Levi、Dominik Lorenz、Axel Sauer、Frederic Boesel 等。扩展用于高分辨率图像合成的整流流变换器。在第四十一届国际机器学习大会，2024。6[9] Rongyao Fang、Chengqi Duan、Kun Wang、Linjiang Huang、Hao Li、Shilin Yan、Hao Tian、Xingyu Zeng、Rui Zhao、Jifeng Dai 等。GoT：通过强化学习释放多模态大语言模型用于视觉生成与编辑的推理能力。arXiv 预印本 arXiv:2503.10639，2025。2[10] Dhruba Ghosh、Hannaneh Hajishirzi 和 Ludwig Schmidt。GenEval：用于评估文生图对齐的以物体为中心的框架。在 NeurIPS，2023。6, 1[11] Ziyao Guo、Kaipeng Zhang 和 Michael Qizhe Shieh。通过由粗到细的令牌预测改进自回归图像生成。arXiv 预印本 arXiv:2503 16194，2025。2 。 。[12] Ziyu Guo、Renrui Zhang、Chengzhuo Tong、Zhizheng Zhao、Rui Huang、Haoquan Zhang、Manyuan Zhang、Jiaming Liu、Shanghang Zhang、Peng Gao 等。我们能否使用思维链生成图像？让我们逐步验证并强化图像生成。arXiv 预印本 arXiv:2501.13926，2025。2, 6, 1[13] Jixiang Hong、Yiran Zhang、Guanzhong Wang、Yi Liu、Ji-Rong Wen 和 Rui Yan。通过双重自奖励强化多模态理解与生成。arXiv 预印本 arXiv:2506.07963，2025。2[14] Kaiyi Huang、Kaiyue Sun、Enze Xie、Zhenguo Li 和 Xihui Liu。T2I-CompBench：一个综合基准测试用于开放世界组合式文生图。在 NeurIPS，2023。6，1[15] Dongzhi Jiang、Ziyu Guo、Renrui Zhang、Zhuofan Zong、Hao Li、Le Zhuo、Shilin Yan、Pheng-Ann Heng 和 Hong- sheng Li。T2i-r1：通过协同语义级与令牌级思维链强化图像生成。arXiv 预印本 arXiv:2505.00703，2025。1，2，6[16] Jingjing Jiang、Chongjie Si、Jun Luo、Hanwang Zhang 和 Chao Ma。用于统一多模态理解与生成的协同强化学习。arXiv 预印本 arXiv:2505.17534，2025。2[17] Haopeng Li、Jinyue Yang、Kexin Wang、Xuerui Qiu、Yuhong Chou、Xin Li 和 Guoqi Li。基于 Mamba 的可扩展自回归图像生成。arXiv 预印本 arXiv:2408 12245，2024 2 . .[18] Shilong Liu、Zhaoyang Zeng、Tianhe Ren、Feng Li、Hao Zhang、Jie Yang、Qing Jiang、Chunyuan Li、Jianwei Yang、Hang Su 等。Grounding DINO：将 DINO 与接地预训练相结合，实现开放集目标检测。载于《欧洲计算机视觉会议》，第 38–55 页。Springer，2024。5，6[19] Zhuoyan Luo、Fengyuan Shi、Yixiao Ge、Yujiu Yang、Limin Wang 和 Ying Shan。Open-MAGVIT2：一个旨在推动自回归视觉生成普及的开源项目。arXiv 预印本 arXiv:2409.04410，2024。2[20] Xu Ma、Peize Sun、Haoyu Ma、Hao Tang、Chih-Yao Ma、Jialiang Wang、Kunpeng Li、Xiaoliang Dai、Yujun Shi 等。Token-Shuffle：利用自回归模型实现高分辨率图像生成。arXiv 预印本 arXiv:2504.17789，2025。2[21] Yuwei Niu、Munan Ning、Mengren Zheng、Weiyang Jin、Bin Lin、Peng Jin、Jiaqi Liao、Chaoran Feng、Kunpeng Ning、Bin Zhu 等。WISE：一种融入世界知识的文生图语义评估方法。arXiv 预印本 arXiv:2503.07265，2025。6，7，1[22] Kaihang Pan、Wendong Bu、Yuruo Wu、Yang Wu、Kai Shen、Yunfei Li、Hang Zhao、Juncheng Li、Siliang Tang 和 Yueting Zhuang。FocusDiff：通过强化学习提升自回归视觉生成的细粒度文本-图像对齐。arXiv 预印本 arXiv:2506.05501，2025。2，6[23] Kaihang Pan、Wang Lin、Zhongqi Yue、Tenglong Ao、Liyu Jia、Wei Zhao、Juncheng Li、Siliang Tang 和 Hanwang Zhang。利用离散扩散时间步令牌进行生成式多模态预训练。载于《计算机视觉与模式识别会议论文集》，第 26136–26146 页，2025。6[24]Dustin Podell、Zion English、Kyle Lacey、Andreas Blattmann、Tim Dockhorn、Jonas M¨uller、Joe Penna 和 Robin Rombach。SDXL：改进潜在扩散模型，实现高分辨率图像合成。arXiv 预印本 arXiv:2307.01952，2023。6[25] Rafael Rafailov、Archit Sharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn。直接偏好优化：你的语言模型其实是一个奖励模型。《神经信息处理系统进展》，36：53728–53741，2023。1[26] Sucheng Ren、Qihang Yu、Ju He、Xiaohui Shen、Alan Yuille 和 Liang-Chieh Chen。超越下一令牌：用于自回归视觉生成的 Next-x 预测，

用于自回归视觉生成的预测。arXiv 预印本 arXiv:2502.20388，2025。2[27]Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Bj¨orn Ommer。利用潜在扩散模型进行高分辨率图像合成。载于《IEEE/CVF 计算机视觉与模式识别会议论文集》，第 10684–10695 页，2022。6[28] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Rad- ford 和 Oleg Klimov。近端策略优化算法。arXiv 预印本 arXiv:1707.06347，2017。1[29] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、YK Li、Yang Wu 等。DeepSeekMath：推动开放语言模型数学推理能力的极限。arXiv 预印本 arXiv:2402.03300，2024。1[30] NextStep 团队、Chunrui Han、Guopeng Li、Jingwei Wu、Quan Sun、Yan Cai、Yuang Peng、Zheng Ge、Deyu Zhou、Haomiao Tang 等。NextStep-1：迈向大规模连续令牌自回归图像生成。arXiv 预印本 arXiv:2508.10711，2025。2[31]Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timoth´ee Lacroix、Baptiste Rozi`ere、Naman Goyal、Eric Hambro、Faisal Azhar、Aur´elien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lample。Llama：开放且高效的基础语言模型。CoRR，abs/2302.13971，2023。1，2[32] Hugo Touvron、Martin 等。Llama 2：开放基础模型与微调聊天模型。arXiv 预印本 arXiv:2307 09288，2023 5 . . [33] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi 等。Llama 2：开放基础模型与微调聊天模型。CoRR，abs/2307.09288，2023。1，2[34] Jianfeng Wang、Zhengyuan Yang、Xiaowei Hu、Linjie Li、Kevin Lin、Zhe Gan、Zicheng Liu、Ce Liu 和 Lijuan Wang。GIT：用于视觉与语言的生成式图像到文本转换器。arXiv 预印本 arXiv:2205.14100，2022。5[35] Xinlong Wang、Xiaosong Zhang、Zhengxiong Luo、Quan Sun、Yufeng Cui、Jinsheng Wang、Fan Zhang、Yueze Wang、Zhen Li、Qiying Yu 等。EMU3：下一词元预测就是你所需要的一切。arXiv 预印本 arXiv:2409.18869，2024。6[36] Yi Wang、Mushui Liu、Wanggui He、Longxiang Zhang、Zi- wei Huang、Guanghao Zhang、Fangxun Shu、Zhong Tao、Dong She、Zhelun Yu 等。MINT：统一生成模型中的多模态思维链，助力提升图像生成。arXiv 预印本 arXiv:2503.01298，2025。1[37] Jason Wei、Xuezhi Wang、Dale Schuur mans、Maarten Bosma、Brian Ichter、Fei Xia、Ed H. Chi、Quoc V. Le 和 Denny Zhou。思维链提示能够激发大语言模型的推理能力。载于《神经信息处理系统进展 35：神经信息处理系统年度会议 2022》，NeurIPS 2022，美国路易斯安那州新奥尔良，2022年11月28日至12月9日，2022。1[38] Yecheng Wu、Zhuoyang Zhang、Junyu Chen、Haotian Tang、Dacheng Li、Yunhao Fang、Ligeng Zhu、Enze Xie、Hongxu Yin、Li Yi 等。VILA-U：整合视觉理解与生成的统一基础模型。arXiv 预印本 arXiv:2409.04429，2024。6 [39] Jinheng Xie、Weijia Mao、Zechen Bai、David Junhao Zhang、Weihao Wang、Kevin Qinghong Lin、Yuchao Gu、Zhijie Chen、Zhenheng Yang 和 Mike Zheng Shou。Show-o：一种统一多模态理解与生成的单一变换器。发表于 *ICLR*，2025 年。6[40] Tianwei Xiong、Jun Hao Liew、Zilong Huang、Jiashi Feng 和 Xihui Liu。GigaTok：将视觉分词器扩展至 30 亿参数以实现自回归图像生成。*arXiv* 预印本 *arXiv:2504.08736*，2025 年。2[41] An Yang、Baosong Yang、Binyuan Hui、Bo Zheng、Bowen Yu 等。Qwen2 技术报告。*CoRR*，abs/2407.10671，2024 年。1，2[42] Xingyi Zhou、Vladlen Koltun 和 Philipp Kr¨ahenb¨uhl。简易多数据集检测。发表于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 7571–7580 页，2022 年。6

从“什么”到“如何”：用于自回归图像生成的受约束推理

补充材料

6. 更多实验细节

6.1. 基准测试

我们使用以下三个权威基准测试对我们的方法进行了系统评估：T2I-Compbench[14], GenEval [10]和 WISE [21]。这些基准测试使我们能够全面测试构图文本理解和世界知识推理等关键能力。

T2I-CompBench 采用 6,000 条精心设计的文本提示，专门评估模型的复杂组合推理能力。该基准测试的测试框架涵盖三个核心维度：属性绑定、物体关系和复杂组合。每个维度进一步细分为六个子类别，分别考察模型在颜色、形状、纹理、空间关系、非空间关系和复杂组合任务方面的表现。

GenEval 在 6 个核心任务中采用 553 条系统设计的提示，评估细粒度的文本到图像对齐能力。这个以物体为中心的基准测试使用由 MS-COCO 物体和基本属性填充的模板，评估单个/多个物体生成、计数、颜色绑定、空间关系和属性绑定等能力。

WISE 专注于评估模型整合世界知识并进行推理的能力。该基准测试包含 1,000 条测试提示词，涵盖文化常识、时空推理和自然科学等领域。在这一基准测试中取得优异表现，要求模型不仅理解字面含义，还能推断文本描述的具体现实世界场景或实体。为获得更一致的评估结果，我们对原始评估指令作了适当调整。

6.2. 提示设置

为了更清晰地说明我们的生成流程，我们提供了训练和推理过程中使用的完整指令提示。该提示引导模型完成“如何绘制”和“绘制什么”这两个过程，从而为图像合成建立一条结构化的推理路径。此处展示了我们在实验中使用的完整提示：

原始提示词的指令提示

请根据此提示生成图像：“{提示词}”。让我们一步一步思考图像中应该出现什么。

<思考> {列出提示词中提到的物体、属性和空间关系，并注明它们应遵循的任何约束。请保持简洁、客观，并聚焦于具体的视觉细节——约 2–3 个短句。}</思考>

<描述> {简要、准确地可视化提示词中的所有元素。包含每个物体；如有提供，请说明其视觉属性（颜色、数量、形状、纹理），并阐明它们之间的关系（例如空间关系）。请简洁作答（$\leq$ 50字），且不要添加提示词以外的任何内容。} </描述>

请明确按以下格式输出：思考： [思考内容]描述： [描述内容]

6.3. 训练设置。

我们的实验配置采用了经过严格筛选的数据集，其中包含 6,786 条文本提示，全部取自 T2I-CompBench 的训练划分以及 [12] 中确立的方法。完整的训练语料恰好包含 条提示，训练过程中未纳入任何视觉数据。

表 4. 模型架构和优化参数

参数	值
学习率	1e-6
贝塔系数 β	0.01
组大小 G	8
无分类器引导尺度	5
最大梯度范数	1.0
批次大小	8
训练步数	1,600
梯度累积步数	2
图像分辨率 $h \times w$	384×384

7. 更多定性结果

7.1. 详细直观比较

我们进一步提供了 CoR-Painter 与基于思维链的最先进方法 T2I-R1 之间的详细视觉对比，以直观展示它们在

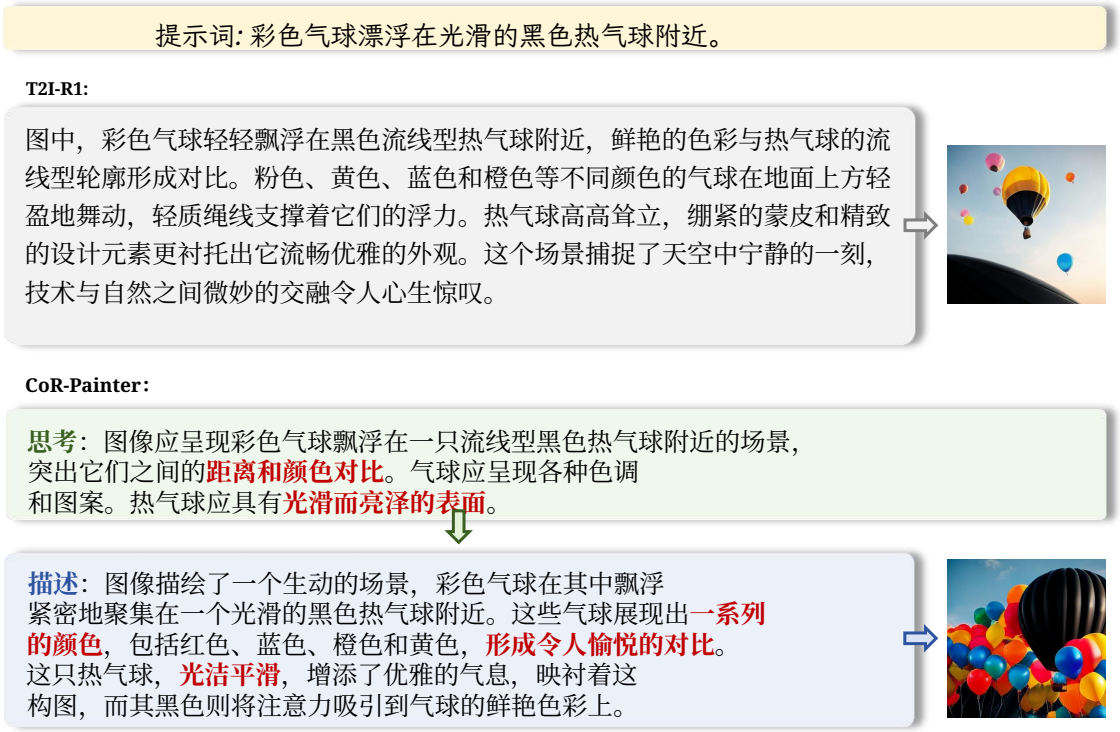


图7. T2I-R1 与我们的 CoR-Painter 之间的详细比较。

图像生成。这些比较涵盖多种提示词类型，包括空间关系、多对象属性绑定、基于知识的推理和复杂场景构图，展示了模型处理各类提示词的能力。如图7-10所示，将文本推理过程与对应图像直接对比可以看出，CoR-Painter 如何通过思维过程中生成的约束指令，系统地引导描述生成，确保物体属性、数量和空间布局得到妥善组织。这种“如何到什么”范式不仅为目标对象和空间安排提供明确指导，还能在知识密集型场景中保持逻辑连贯性。生成的图像在语义准确性、空间关系和视觉一致性方面均有明显提升，有效缓解了布局模糊导致的物体重叠、计数错误等常见问题，同时增强了整体场景完整性和可解释性。

7.2. 更多思维消融可视化

如图 11-14所示，我们比较了有无思维过程时的文本推理过程及相应的生成图像。为清楚说明消融研究的输入，我们还提供了不含思维过程的版本所使用的完整提示词，以便读者直接了解模型接收了什么信息，以及这些信息如何影响生成的描述和图像。

消融指令提示词

请根据以下提示词生成一幅图像：“{提示词}”。让我们逐步思考图像中应出现什么内容。<description> { 请对提示词中的所有元素进行简短、准确的可视化描述。包括每个物体；如果提示词中给出了视觉属性（颜色、数量、形状、纹理），请明确说明；并阐明它们之间的关系（例如空间关系）。请保持简洁（≤ 50字），不要添加提示词之外的任何内容。} </description>请明确按以下格式输出：描述： [描述内容]

这些视觉结果阐明了思维过程中生成的约束指令如何确保空间布局连贯、物体间的逻辑关系一致，清晰地凸显了思考引导的推理对整体图像质量的贡献。移除此模块后，生成的描述缺乏结构指导，导致空间排列明显变差，物体间关系也受到破坏。这项消融研究通过视觉结果证实了我们的“如何到什么”范式对于维持生成图像的结构完整性和语义一致性起着关键作用。

提示词： 绿色植物在白色花瓶旁边。

T2I-R1：

一株叶片纤细、鲜艳的绿色植物立在一只纯白花瓶旁，两者相得益彰。



CoR-Painter：

思考：图像中应包含一株绿色植物，旁边放着一个白色花瓶，两者置于家庭环境中。植物应当**生机盎然、看起来健康**，叶片呈绿色，且**应靠近**花瓶，**两者之间没有可见的间隙或空隙**。

描述：一株鲜艳的绿色植物，长着茂盛的叶子，**略微偏右处立着**一只纯白花瓶。它们位于一间舒适明亮的房间里，房间铺着木地板，地上放着一块柔软的图案地毯。花瓶很可能是**陶瓷制成，表面有华丽的设计**；这株植物**或许是多肉植物或室内植物**，看起来安然地待在它的位置上。



Prompt：开裂的矩形倚靠着光泽圆柱和剥落的三角形 .

T2I-R1：

图像描绘了一个杂乱的场景，其中有三个不同的物体。第一个是一块倚靠在光泽圆柱上的开裂的矩形。这个米色矩形上有锯齿状裂纹图案，横贯其表面，边缘锐利，因磨损而显得陈旧。旁边立着一个闪亮的灰色圆柱体。它表面光滑、有光泽，能反射光线，与矩形形成鲜明对比，突显其流畅简洁的外形。圆柱体在倾斜的矩形旁显得坚固。最后，还有一个暗绿色的剥落三角形物体。它位于圆柱体附近，粗糙的纹理上可见剥落层，呈现出破旧的外观。剥落的三角形边缘尖锐且呈斜角，底部略微凹陷，为场景增添了破败荒废之感。



CoR-Painter：

思考：图像应描绘一个开裂的矩形**倚靠着**一个光泽圆柱。此外，一个剥落的三角形清晰可见，矩形和圆柱体都具有**不同的纹理**。

描述：图像中，一个竖直放置、开裂的矩形**倚靠着**一个光泽的、垂直放置的圆柱体。矩形表面布满了**裂缝和破损**，呈现出**粗糙纹理**。一块剥落的三角形增添了一层场景右侧的杂乱，与之在颜色和纹理上形成对比的是**圆柱体和矩形原本光泽的表面**。



图8. T2I-R1 与我们的 CoR-Painter 之间的详细比较。

提示词： 条纹地毯在木制长凳旁边。

T2I-R1：

场景展示了地板上的一条条纹地毯，紧邻木制长凳。地毯呈现交替的明暗条纹图案，柔软质地表明它可能由羊毛或其他天然纤维制成。木制长凳由精心挑选的抛光木材制成，展现出质朴而结实的外观。长凳紧邻地毯摆放，营造出协调而温馨的氛围。这幅图像令人联想到一个舒适而宜人的空间，非常适合放松或社交。



CoR-Painter：

思考：图像中应包含一块带有鲜明彩色条纹的条纹地毯，放在一张质朴外观的木制长凳旁。地毯**表面应具有可见纹理**，而长凳**应带有水平或垂直板条，以增添视觉特色**。场景还应尽可能采用自然光或室内光照，营造出**逼真氛围**。

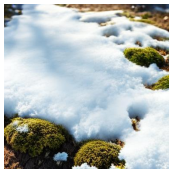
描述：图像展示了**一条带有鲜明彩色条纹的条纹地毯**，放置在木制长凳旁边。地毯具有可见纹理，而长凳则有**水平或垂直板条**，呈现出质朴外观。场景的照明**来自自然光或室内光**，营造出逼真氛围。



提示词： 平滑的白色雪覆盖着粗糙的棕色地面。

T2I-R1：

被雪覆盖的场景中，白色的雪层平滑地铺在纹理粗糙的棕色地面上。雪看起来松软蓬松，呈现出柔和、轻盈的质感，并反射着阳光，在大地色调的地面上形成明亮的对比。周围的苔藓和草从雪中露出，展现出自然冬季场景中多样的纹理和颜色。雪覆盖着相对平坦的景观，而粗糙的棕色地面则成为平滑白雪的鲜明背景。这幅图像中看不到其他元素或物体



CoR-Painter：

思考：图像应呈现一片以积雪地面为主的景观。雪地将平滑且呈白色，覆盖着粗糙的棕色地形，**没有树木或建筑物等其他显著特征**。

描述：一幅**积雪的棕色景观**的视觉呈现，其中**平滑的白色雪层均匀覆盖着粗糙的棕色地形**。



图9. T2I-R1 与我们的 CoR-Painter 之间的详细比较。

提示词： 一名滑雪者正沿着白雪覆盖的山坡滑下。

T2I-R1：

在一幅滑雪者沿着积雪覆盖的山坡滑行的图像中：1. 一名滑雪者正在运动，滑雪板掠过雪面。2. 背景中有一座积雪覆盖的山峰，积雪正向下倾泻。3. 滑雪者的站姿和装备清晰可见，包括头盔、护目镜和滑雪杖。4. 通过运动线或模糊细节表现速度与动感。5. 冬日氛围寒冷，颜色柔和，以白色、灰色和蓝色为主。6. 空中飘着雪花或吹雪，营造出狂风卷雪的环境氛围。



CoR-Painter：

思考：图像应呈现一名正在运动的滑雪者，他穿着鲜艳的滑雪服，背景中有被雪覆盖的山峰和雪道。滑雪者的**装备应包括**滑雪板和护目镜，以及头盔和手套等御寒配件。图像**应包含一些元素**，例如太阳、云，或**山脚下的村庄**，具体取决于整体氛围。

描述：一名滑雪者沿着积雪覆盖的山坡滑下。滑雪者**身着色彩鲜艳的服装，包括头盔、护目镜、围巾和手套**，滑雪装备包括多块滑雪板和滑雪杖。画面还描绘了背包和滑雪靴等御寒配件，并将**焦点放在积雪山峰与雪道构成的壮丽景观上**。背景可以根据需要包含积雪覆盖的湖泊、树木或山脚下的村庄等自然景物。画面还可以呈现**阳光**、云或其他大气元素。



提示词：一位园丁正在修剪一棵美丽的盆景树。

T2I-R1：

一位园丁正仔细修剪一棵枝叶繁茂、色彩鲜艳的盆景树，叶片优雅而造型别致。盆景树错综复杂的枝条宛如一幅马赛克，交织着浓郁的绿色和深红色，在阳光下还闪烁着些许金色。园丁戴着防护修剪手套，专注而勤勉地进行修剪。场景设在宁静的花园露台上，四周环绕着盛开的花朵和繁茂的植物，营造出放松而和谐的氛围。



CoR-Painter：

思考：画面应包含一位园丁正在修剪一棵美丽的盆景树，并着重表现细节
例如**树木精巧的形状、园丁的工具以及宁静的花园环境**。

描述：一位园丁**熟练地修剪着一棵迷人的盆景树**，展现出其**纤细的枝条和繁茂的绿色枝叶**。园丁站在**一座宁静的花园环境**中，周围环绕着其他植物、木长椅，以及柔和的阳光透过叶片洒落。



图 10. T2I-R1 与我们的 CoR-Painter 的详细对比。

提示词: 一只 蝴蝶位于蜡烛左侧。

无需思考:

描述: 一只蝴蝶，带有精致的
图案和丰富的色调，位于左侧
一支发光蜡烛的左侧，映照着
它的翅膀上的光辉，同时它显现于
宁静祥和的背景中。
蜡烛的火焰强烈而温暖，
投下闪烁的金色光芒，
光芒在细腻的纹理间舞动，
蝴蝶翅膀上。稳定的
蜡烛的光辉柔和地映亮
边缘处鲜艳的色素
翅膀，为宁静的场景
增添了一丝诗意的优雅。



添加指导:

思考: 图像中应包含一只蝴蝶，其
突出的形状和颜色可能会**在左侧**
显眼。 **一根蜡烛，通常具有纤细的**圆柱形主体，
散发着光，**位于右侧**。蜡烛的火焰
强烈而温暖。

描述: 图像将呈现一只优雅的蝴蝶
栖息在左侧，其右边，放着一支点燃的蜡烛
置于烛台或烛座上，周围环境光线柔和，
可能是**花园、卧室或舒适的室内空间**。这只
蝴蝶展现出鲜艳的颜色，翅膀上有精致的
图案，而蜡烛散发出温暖的火焰，营造出
夜晚宁静祥和的氛围。



提示词： 中秋节的传统食物。

无思考:

描述: 热闹非凡的中国传统
中国市集景象，人们
正沉浸在中秋节的氛围中。
前景中，整齐地摆放着
以众多为特色的传统小吃摊
由月饼、幸运饼干制成的菜肴，
以及其他标志性的地方特色美食。
集市被温暖柔和的灯光照亮
灯光来自装饰着精致图案的悬挂的灯笼
营造出宁静而喜庆的
氛围。当地家庭和朋友们聚集在一起
围在摊位旁，分享着欢声笑语
一边交谈，一边享受丰富多样的
传统菜肴风味。



添加指导:

思考: 中秋节的传统食物
节日**应该是月饼**。物体包括
月饼，**小灯笼和一轮满月**。 视觉
属性**应包含柔和的暖黄色**
与庆典相关。月饼的
饼皮可以饰以传统图案，
灯笼上则装饰着星形图案。

描述: 一个节日场景，其中有**月饼，**
传统小灯笼和一轮满月
象征着中秋节。 **月亮**
主题装饰增添了一丝优雅和
传统气息，邀请人们参与
欢乐的庆典。



图11. 有无思维过程的图像对比。

提示词： 左侧一个又高又细的矩形，右侧一个又短又宽的矩形

不含思考：

描述： 生成一幅高而窄的矩形图像
以及一个矮而宽的矩形，窄的矩形位于左侧
宽的矩形位于右侧。图像中有一个高而
窄的矩形位于左侧，呈鲜艳的绿色
其中有六扇间距均匀、垂直排列的窗户。
右侧的矮而宽的矩形呈醒目的
蓝色色调的窗户共八扇，间距宽松，水平
排列。蓝色横幅上写着
“Eccentric Boutique”，优雅的白色文字横跨
两个矩形之间。狭长的绿色矩形延伸
斜穿蓝色矩形，营造出动感十足的
布局，在熙熙攘攘的环境中格外引人注目
城市景观，与种类繁多的
的建筑物。



添加指导：

思考：图像**应包含两个矩形**，一个又高又细、呈左-向朝向，以及一个更短、更宽的结构，呈朝右的朝向，**二者尺寸具有明显对比**。

描述：画面将呈现一个高高的左侧有一个狭长的矩形，另有一个右侧有一个短而宽的矩形，两者采用 **颜色区分来突出显示它们的尺寸，** 以 **水平对齐**于可能的背景之上：一面实心墙或房间。



提示词： 一种与墨西哥沙漠和文化息息相关、以带刺的质地闻名的耐旱植物。

无思考：

描述： 一种耐旱植物，与
墨西哥沙漠和文化息息相关，以其带刺的
质地闻名。观察墨西哥沙漠时，一株
色彩鲜艳的植物坚韧地生长着，展现出一种
带刺的质地，令其能够抵御恶劣环境。
它呈现出浓郁的绿色，茎的质地各异，
从光滑到带有细密纹路不等，而它的
叶片边缘尖锐而狭长，并覆有一层
保护层，用以抵御
干旱环境。这种植物看起来
与丰富的墨西哥
文化有着深厚的联系，体现了韧性的精神，
以及面对逆境时的适应能力。它构成了
沙漠生态系统的重要组成部分，
为多样提供庇护所和营养
昆虫和小型哺乳动物。



添加指导：

思考：图像 **应包含一株仙人掌**，墨西哥沙漠的典型特征。植物应生长于 **有岩石的沙地中石块**，为墨西哥的氛围增添沙漠文化特色。色彩搭配应温暖，以棕色、绿色和橙色色调为主，以呈现沙漠环境和文化元素。

描述：一片棕色的景观，有沙子和岩石，展现出一种长着尖锐绿色刺的耐旱植物。 **仙人掌位于**沙漠中央，背景中有几朵橙色和黄色的沙漠花卉。 **仙人掌被描绘成**焦点元素，赋予其文化遗产的意蕴，并展现出在严酷环境中的韧性。



图 12. 有思维过程和无思维过程的图像对比。

提示词: 一个立方体冰块和一个圆柱形水瓶。

无思考：

描述： 一幅 3D 渲染图，展示了立方体冰块呈现为一个透明的浅蓝色立方体，带有锐利的边缘和冰块般的图案。底部连接着圆柱形水瓶，这是一个半透明蓝色容器，表面带有纹理，略显凹凸模仿了冷凝水的外观。这只水瓶部分隐藏在冰块后面，暗示了一种空间关系：一个物体部分遮挡了另一个物体。



添加指导：

思考：图像中应包含一个由以下物质构成的立方体冰块：**透明冰晶**，周围环绕着**一个圆柱形水瓶**瓶中装有清澈的静水。冰块和水瓶应摆放的位置应**使它们清晰可见，且彼此易于区分**。这些元素应置于没有其他干扰或杂物的室内环境中。

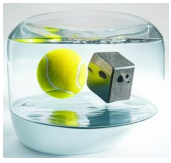
描述：图像中有一个由透明冰晶制成，置于平坦表面上。周围区域包括一个装满清澈的静水的圆柱形水瓶，**放置在距离冰块不远处**。内部的水瓶外部出现冷凝水，表明水的温度低于环境温度。自然室内环境光线充足，为场景增添了真实感。



提示词： 一个网球和一个铁块位于透明水箱中。

不含思考：

描述： 一个网球和一个铁块位于透明水箱中。一个透明水箱中装有一个网球和一个铁块，两者都漂浮在水面上方。网球呈深黄色，带有光滑的橡胶质感，而铁块则是金属灰色，带有粗糙的金属质感。水占据了水箱，营造出一种令人愉悦的美感对比，位于形成对比的颜色和材质。水坦克置于稳固的表面上，周围环绕着明亮的白色背景。这张图像捕捉到了这两个物体之间独特的互动及其对比特征。



添加指导：

思考：图像应包含一个透明的水箱、一个网球和一个放置在水箱内的网球 **应漂浮在水中**，而**铁块应位于底部**或水面附近。两个物体都应清晰可见，且水箱应装满水达到一定高度。

描述：一个装满水的透明玻璃水箱水中有**一个漂浮的网球和一块铁块在底部**或接近水面的位置。网球漂浮在水中，而铁块没有浸入水中，**形成了引人入胜的物体之间的视觉关系**。

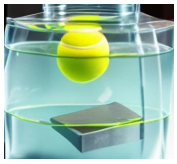


图13. 有无思维过程的图像对比。

提示词：橙色金字塔位于粉色半球和黄色圆柱体之间。

不加思考：

描述：一个色彩缤纷、充满活力的场景展现了一座位于……之间的橙色金字塔一个鲜艳的粉色半球和一个明亮的黄色圆柱体。金字塔以其尖锐角和几何形状，在半球体光滑的表面和圆柱体充满活力的形体。粉色半球体带有弯曲柔和的边缘，从一侧环绕着金字塔的底部，与金字塔的直线形成对比，金字塔。相反，黄色圆柱体凭借其圆柱形主体和光泽表面，被放置在更靠近观者的位置，突显其相较于其他物体的存在感与主导地位金字塔周围的元素。这种独特的色彩与形状的组合营造出一种视觉上引人入胜且充满动感的构图。



添加指导：

思考：图像应展示一个包含以下元素的场景：一个橙色金字塔、一个粉色半球和一个黄色圆柱体。**这三个不同的物体应当以一种突出其空间彼此关系的方式摆放。**

描述：图像呈现出鲜艳且色彩缤纷的场景中有**一个位于中央的橙色金字塔它位于一个粉色半球和一个黄色圆柱体**。金字塔鲜艳的橙色与半球的粉色以及圆柱体的黄色形成了**视觉对比**。粉色半球和黄色圆柱体位于场景的两端，衬托着金字塔。**整体构图给人一种平衡与和谐之感，形状和颜色**无缝融合，带来连贯统一的视觉体验。



提示词：一块蓬松地毯和一个皮革钱包。

不进行思考：

描述：这张图像描绘了一个柔软蓬松的地毯，触感纹理柔和外观主要呈现为各种色调由米色和白色组成。地毯旁边，有一个经典的皮革钱包，它的表面呈温暖浓郁的棕色，带有清晰可见的缝线细节。钱包直立摆放，与舒适的……形成对比，地毯的触感。这个可视化呈现了两者之间的对比在这两个项目之间时聚焦于它们各自不同的外观和特征。



添加指导：

思考：图像应包含**一块蓬松地毯**，上面呈现柔软的白色**类似地毯的纤维作为主要视觉元素。皮革钱包**也应采用光泽表面，并带有其标志性的设计，例如**皮革带**。

描述：图像呈现了一张蓬松的白色地毯，占据前景，**其柔软质地十分突出**，营造出温暖舒适的氛围。在蓬松纤维间，**一只时尚的棕色色调浓郁、配有皮革带的皮革钱包凭借微妙光泽形成视觉对比**。钱包的结构与工艺和地毯浑然一体，提升整体的奢华美感与舒适感。



图14. 有思维过程与无思维过程的图像对比。