LONG-TEXT-TO-IMAGE GENERATION VIA COMPOSITIOINAL PROMPT DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

While modern text-to-image models excel at generating images from intricate prompts, they struggle to capture the key details when the prompts are expanded into descriptive paragraphs. This limitation stems from the prevalence of short captions in their training data. Existing methods attempt to address this by either fine-tuning on long-prompt data, which generalizes poorly to even longer inputs; or by projecting the oversize inputs into normal-prompt domain and compromising fidelity. We propose a compositional approach that enables pre-trained models to handle long-prompts by breaking it down into manageable components. Specifically, we introduce a trainable PromptDecomposer module to decompose the long-prompt into a set of distinct sub-prompts. The pre-trained T2I model processes these sub-prompts in parallel, and their corresponding outputs are merged together using concept conjunction. Our compositional long-text-to-image model achieves performance comparable to those with specialized tuning. Meanwhile, our approach demonstrates superior generalization, outperforming other models by 7.4% on prompts over 500 tokens in the challenging DetailMaster benchmark.

1 INTRODUCTION

Compositionality is fundamental to human intelligence—the ability to understand novel concepts by decomposing them into familiar primitives and to build complex systems from simple components. This "divide and conquer" strategy is also common in creative activities. An artist, for instance, rarely materializes an intricate scene holistically. Instead, they might independently perfect the rendering of a rustic wooden house and the surrounding trees, ensuring each element is realized with care before integrating them into a cohesive whole. In stark contrast, text-to-image (T2I) models (Rombach et al., 2022) attempt to render the entire scene simultaneously in a single, monolithic process.

This paradigm works well for concise prompts but falters when the input becomes a descriptive paragraph. While a model may excel at rendering "a house in the middle of a forest" it often fails when the prompt expands, detailing the terracotta roof tiles, the weathered white panels of the house, and the striking contrast cast by the afternoon sun. This failure stems from a fundamental conflict between the nature of long-form text and the models' training paradigm. T2I models are predominantly trained on vast datasets of images paired with short, concise captions. They learn to map phrases to visual features but are undertrained on interpreting the narrative flow and distributed details of a paragraph (Bai et al., 2024). Even modern models using powerful text-encoders struggle on these inputs, missing more than a half of the specified objects (Jiao et al., 2025).

Existing methods attempt to bridge this domain gap with two main strategies (Figure 2). The most direct approach involves fine-tuning the T2I model on long-captioned images (Bai et al., 2024; Wu et al., 2025b). This is computationally prohibitive and risks "catastrophic forgetting" of the pre-trained knowledge. Furthermore, these tuned models often generalize poorly to prompts even longer. A second strategy adopts projection-based methods to compress a long-prompt into the compact semantic space that the T2I model understands (Hu et al., 2024; Liu et al., 2025). While efficient, forcing a rich paragraph through a narrow keyhole is inherently lossy, sacrificing the very details that make the long-prompt compelling. These limitations reveal an open question: how to utilize model's knowledge on short prompts to render the long paragraphs?

In this paper, we advocate the idea of compositionality for long-text-to-image generation. Instead of forcing the pre-trained model to follow the entire paragraph at once, we decompose it into a

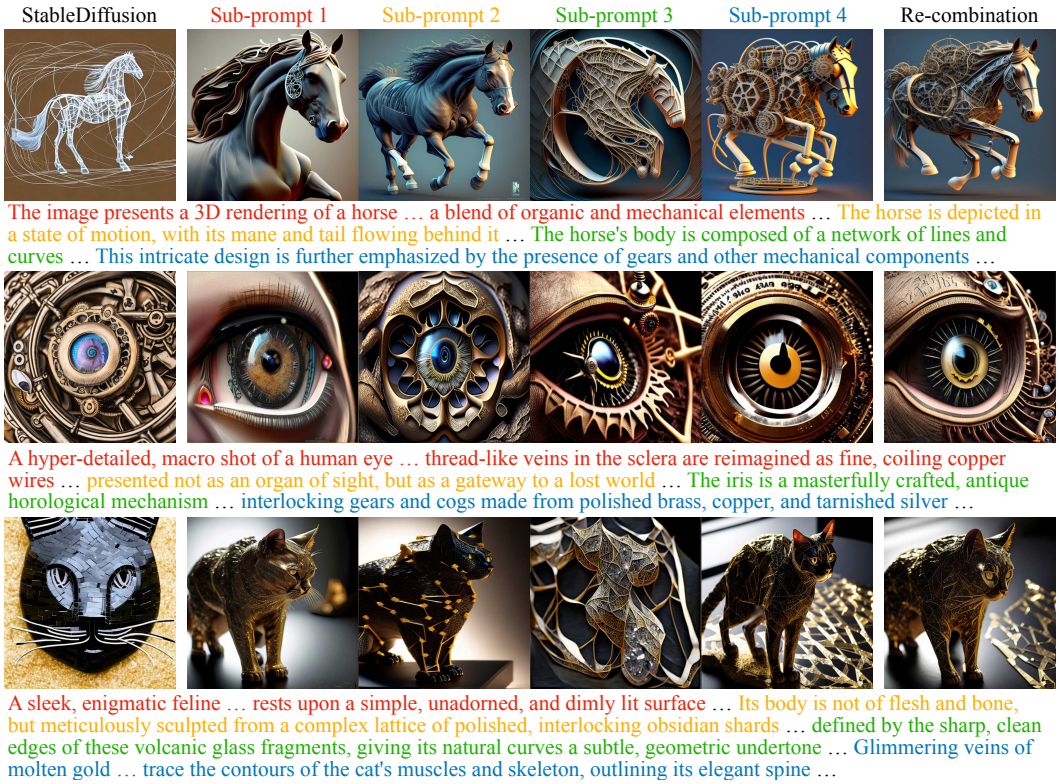

Figure 1: **Decomposing Long-prompt for Compositional Generation.** We decompose the long-prompt into manageable sub-prompts, each depicting parts of the original input. Model outputs on each of the decomposed sub-prompts are then re-combined into a final, cohesive image.

set of manageable sub-prompts. The final image is generated from the factorized distribution of the decomposed sub-prompts. Our approach draws inspiration from the compositional generative modeling, which can generalize constituent models to new tasks beyond individual capacity (Du & Kaelbling, 2024; Du et al., 2023). This compositional strategy offers two unique advantages. First, it allows the pre-trained model to operate within its domain of expertise—processing concise concepts—thereby eliminating the need for expensive tuning and preserving its powerful prior knowledge. Second, it ensures higher fidelity to the original input by distributing the paragraph's rich information across multiple components.

How to obtain such decomposition then becomes the central challenge; a simple linguistic split is insufficient as it loses global context. Our key insight is to directly learn this decomposition in the representation space, guided by the T2I model itself. We introduce *PromptDecomposer*, an end-to-end trainable module that uses a set of learnable vectors to query and extract sub-prompt representations from the encoded paragraph. A pre-trained T2I model parallelly processes the decomposed representations, with the resulting noise predictions merged into a single coherent update at each denoising step. *PromptDecomposer* is trained with both the text-encoder and the T2I model frozen, learning to factorize the intricate long-prompt representation into components that are interpretable by the pre-trained model. Our compositional solution delivers performance comparable to tuning-based methods, and crucially, demonstrates superior generalization as prompt length increases, outperforming other methods by **7.4%** on prompts over 500 tokens. Our contribution can be summarized as:

1. We propose a compositional framework for long-text-to-image generation that utilizes pre-trained T2I model without expensive fine-tuning.
2. We introduce a trainable PromptDecomposer module to directly decompose long-prompt representations for compositional generation.
3. Results show our method achieves comparable performance to tuning-based methods on a challenging benchmark, while offering superior generalization to longer inputs.

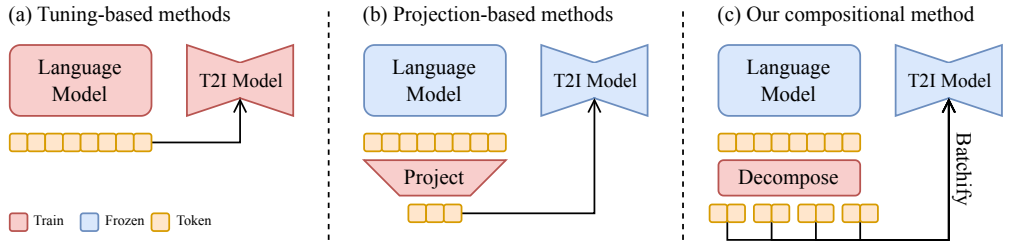

Figure 2: **Long-Text-to-Image Generation Strategies.** (a) Tuning-based methods adapt the T2I model to long-prompt inputs; (b) Projection-based methods map the long-prompt to compact space; (c) We decompose the long-prompt into several sub-prompts for compositional generation.

2 Related Work

2.1 Long-text-to-image Generation

Diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015) have significantly propelled visual generation. Integrated with text conditioning, these models can generate images with unprecedented diversity and quality from natural language descriptions (Rombach et al., 2022; Ramesh et al., 2022). Recent progress in model architecture (Peebles & Xie, 2023) and theoretical foundations (Liu et al., 2022b; Lipman et al., 2022) have enabled T2I models to scale to billions of parameters (Esser et al., 2024; Batifol et al., 2025). Despite this progress, a key limitation remains their difficulty in interpreting long, descriptive paragraphs (Jiao et al., 2025). This challenge often stems from the fixed context window of the text encoders (e.g., CLIP (Radford et al., 2021)), which can be overcome by using more powerful language models (LMs) (Zhao et al., 2024; Liu et al., 2025). However, adapting to the new input takes intensive tuning. An efficient strategy involves projecting the LM representations into the T2I model's original text embedding space (Hu et al., 2024; Liu et al., 2025). To systematically evaluate performance on this task, DetailMaster (Jiao et al., 2025) introduces a rigorous benchmark consists of intricate prompts with an average length of 284.9 tokens depicting complex scenes with multiple objects. It also provides a comprehensive, multi-stage evaluation pipeline leveraging multimodal models to analyze visual details.

2.2 Compositional Generative Modeling

Our work builds on the principle of compositional generative modeling, which constructs complex generative systems by combining simpler, specialized models rather than training a single monolithic one (Du & Kaelbling, 2024; Garipov et al., 2023). Conceptually, this approach treats each model as a soft constraint and uses optimization techniques to find outputs that have a high likelihood across all constituent models (Du et al., 2023; Yang et al., 2023). A key advantage of this approach is its data efficiency and generalization capability; by learning simpler, factorized distributions, a compositional system can generate valid samples for combinations of patterns unseen during training (Mahajan et al., 2024). In vision domain, compositional methods enable the generation of novel images with blended features (Du et al., 2020). For instance, composing T2I diffusion model outputs on different text prompts leads to a sample that is collectively described by all prompts (Liu et al., 2022a; Bradley et al., 2025; Bar-Tal et al., 2023; Yang et al., 2024). It is also possible to train a compositional generative system as a whole. This allows each constituent model to learn a compositional factor from data, which can then be recombined to synthesize novel combinations (Su et al., 2024; Liu et al., 2023). Similarly, we approach the challenge of long-text-to-image generation through a compositional lens, aiming to identify and model the compositional factors within a complex text prompt.

3 Methodology

Our approach achieves long-text-to-image generation by reframing it as a compositional task. Instead of training a monolithic model to interpret an entire paragraph, we decompose the paragraph into a set of sub-prompts that a pre-trained T2I model can readily understand. The final image is then synthesized by composing the model's outputs for each sub-prompt, a technique made possible by the insight that diffusion models can be treated as composable energy-based models.

3.1 PRELIMINARIES: COMPOSING DIFFUSION MODELS

Text-to-Image Diffusion Generation. A T2I diffusion model, $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$, generates an image $\boldsymbol{x}$ conditioned on a text prompt $\boldsymbol{c}$ by progressively denoising the input to decreased noise levels $\{\sigma_t\}_{t=1}^T$ (Ho et al., 2020). The model is trained to predict the noise $\boldsymbol{\epsilon}_t$ added to an image $\boldsymbol{x}$ at timestep t. Generation begins with pure Gaussian noise, $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \sigma_T^2 \boldsymbol{I})$, which the model iteratively refines by subtracting the predicted noise at each step. This process corresponds to score-based modeling (Song et al., 2020b), where the predicted noise is proportional to the time-dependent score function (the gradient of the log likelihood): $\boldsymbol{\epsilon}_\theta \propto -\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{c})$. Generation can thus be viewed as a form of Langevin dynamics (Du & Mordatch, 2019),

$$\boldsymbol{x}_{t-1} = \boldsymbol{x}_t + \frac{\sigma_t^2}{2} \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) + \sqrt{\sigma_t}\boldsymbol{\epsilon}. \tag{1}$$

where the learned score function at each timestep gradually guides a sample toward a high-density region of the target data distribution $p(\boldsymbol{x}|\boldsymbol{c})$.

Energy-Based Compositionality. The score-based view of diffusion models reveals a connection to Energy-Based Models (EBMs). An EBM defines a probability density via an unnormalized energy function, $p_\theta(\boldsymbol{x}) \propto e^{-E_\theta(\boldsymbol{x})}$, and uses the gradient of this energy function with Langevin dynamics for generation. A key advantage of EBMs is their inherent compositionality; sampling from a product of multiple data distributions is as simple as summing their energy functions:

$$p_{compose}(\boldsymbol{x}) \propto \prod_i p_\theta^i(\boldsymbol{x}) \propto e^{-\sum_i E_\theta^i(\boldsymbol{x})}, \tag{2}$$

yielding sample with high-likelihood across all constituent EBMs. As demonstrated by prior work, this logic can be extended to diffusion generation by drawing an line between the diffusion model and the gradient of an implicit energy function, $\boldsymbol{\epsilon}_\theta \approx \nabla_{\boldsymbol{x}_t} E_\theta(\boldsymbol{x}_t)$. To sample from the product of two distributions conditioned on $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$, one can simply sum their respective noise predictions,

$$\boldsymbol{\epsilon}_{composed}(\boldsymbol{x}_t, t) = \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_1) + \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_2) \propto \nabla_{\boldsymbol{x}_t} \log \left(p_t(\boldsymbol{x}_t|\boldsymbol{c}_1) \cdot p_t(\boldsymbol{x}_t|\boldsymbol{c}_2) \right), \tag{3}$$

in the score function of Equation 1. This operation, known as **concept conjunction** (Liu et al., 2022a), forms a new composite score that guides the generation process toward an image satisfying both prompts simultaneously. Notably, the synthesized sample won't have to be presented in either of the training data in $p(\boldsymbol{x}|\boldsymbol{c}_1)$ and $p(\boldsymbol{x}|\boldsymbol{c}_2)$. This principle allows us to construct novel scenes from familiar concepts, laying the cornerstone for our approach.

3.2 COMPOSITIONAL LONG-TEXT-TO-IMAGE GENERATION

The domain gap in input prompts is the core challenge of long-text-to-image generation. A descriptive paragraph, $\boldsymbol{C}$, is fundamentally an out-of-distribution input for pre-trained T2I model $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$. These models are trained on vast datasets like LAION (Schuhmann et al., 2022), which is dominated by short, label-like captions $\boldsymbol{c}$. Therefore the models primarily learn to map keywords and short-phrases to visual features, lack the ability of narrative comprehension. Our central hypothesis is that the complex conditional distribution $p(\boldsymbol{x}|\boldsymbol{C})$ described by the paragraph $\boldsymbol{C}$ can be effectively approximated by factorizing into a set of simpler distributions: $p(\boldsymbol{x}|\boldsymbol{C}) \propto \prod_i^N p(\boldsymbol{x}|\boldsymbol{c}_i)$, where each constituent distribution $p(\boldsymbol{x}|\boldsymbol{c}_i)$ is conditioned on a sub-prompt $\boldsymbol{c}_i$. Intuitively, a paragraph can be abstracted as a collection of phrases with each capturing a distinct feature. Leveraging the concept conjunction principle in Equation 3, we can construct a long-text-to-image generation model by composing a same pre-trained T2I model $\boldsymbol{\epsilon}_\theta$ with different sub-prompts:

$$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{C}) = \sum_{i=1}^{N} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}_i). \tag{4}$$

This composite score leads to an image that is collectively described by the sub-prompts $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$. Because the sub-prompts remain semantically concise, they can be readily processed by the pre-trained T2I model, avoiding resource-intensive fine-tuning. Furthermore, unlike projection-based methods that suffer from information loss, our factorized approach maintains high fidelity to the original paragraph by distributing its information across multiple sub-prompts $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$.

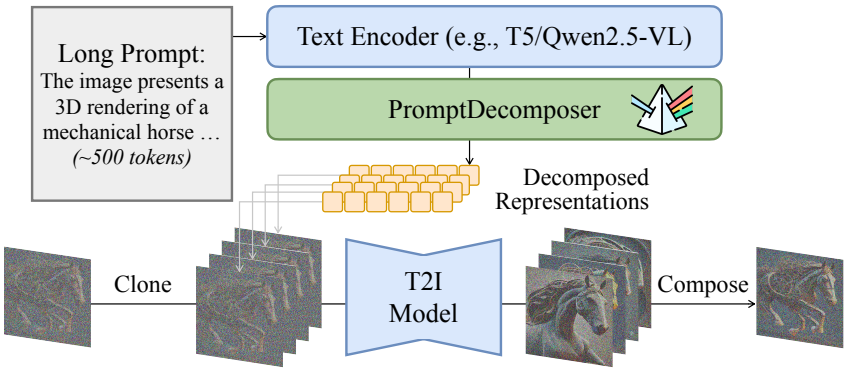

Figure 3: **Compositional Long-Text-to-Image Generation Model.** The input long-prompt is first encoded by the pre-trained T2I model's text-encoder. Our PromptDecomposer module then extracts the decomposed sub-prompt representations from the encoded long-prompt. Current noisy latent is first cloned into a batch according to the number of decomposed sub-prompts and parallelly processed by the T2I model. Finally, the noise predictions conditioned on different sub-prompts are merged into a composed diffusion step through concept conjunction.

3.3 Unsupervised Long-Prompt Decomposition

To obtain the sub-prompts $\{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$, one appealing option is to utilize LLMs to analyze and break down the paragraph. However, Equation 3 lacks explicit spatial control over each sub-prompt input, resulting in global context inconsistency and local concept blending. We propose to learn such decomposition directly in the textual representation space via a trainable *PromptDecomposer* (ψ) module. The T2I model utilizes the module's output to form the noise prediction as per Equation 4. Crucially, the entire composed model is trained end-to-end in an unsupervised manner, using the diffusion loss calculated on the composite score,

$$\mathcal{L}(\psi) = \mathbb{E}_{\boldsymbol{x},t}\left[\left|\sum_{i=1}^{N} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t, \boldsymbol{c}_i) - \boldsymbol{\epsilon}\right|^2\right], \psi(\boldsymbol{C}_{LM}) = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}. \tag{5}$$

By training on the frozen T2I model, the PromptDecomposer learns to distribute the information into sub-prompts $\{\boldsymbol{c}_i\}$ that are explicitly optimized for the compositional generation process in Equation 3. This end-to-end training allows the learned decomposition to effectively capture spatial relationships and global attributes (eg., lighting, style) that are critical for consistency but lost in linguistic splitting.

3.4 Implementation Details

Our compositional approach functions as a general framework to generalize T2I models to long-prompts outside training data distributions. Here we present two distinct PromptDecomposer module designs tailored to the type of text-encoder employed by the underlying T2I model.

Bidirectional Text-Encoder. For full attention Transformer text-encoder (Vaswani et al., 2017) like the T5 (Raffel et al., 2020) in Stable Diffusion-3.5 (Peebles & Xie, 2023) or FLUX (Batifol et al., 2025), we implement PromptDecomposer as a Perceiver network (Jaegle et al., 2021). As illustrated in Figure 10, since the text-encoder outputs a fixed length hidden state, we employ N learnable vectors to query the encoded long-prompt $\boldsymbol{C}_{LM}$ through multiple cross-attention layers. These queries thus learn to extract distinct semantic components from the global context, projecting them into sub-prompt representations optimized for compositional generation.

Decoder-only Language Model. For T2I models built upon decoder-only LLMs, such as Qwen-Image (Wu et al., 2025a) with Qwen2.5-VL (Bai et al., 2025), we leverage the LLM's inherent reasoning capabilities to generate decomposed representations directly. As illustrated in Figure 11, we replicate the input tokens N times and prepend a trainable component token $\langle|\texttt{comp}_i|\rangle$ to each segment i. These N segments are concatenated into a single contiguous input sequence, allowing the model to reason over the full context while generating distinct representations for each component. The final output is chunked into N pieces corresponding to the sub-prompt representations.

Table 1: **Evaluations on the DetailMaster Benchmark (Jiao et al., 2025).** We perform comparisons within two groups: Long-Text-to-Image generation methods built on Stable Diffusion-1.5, and state-of-the-art T2I models including SDXL, Stable Diffusion-3.5, FLUX and Qwen-Image. Numbers are reported in percentage accuracies and the best results in each group are marked in **Bold**.

Model	Character Presence	Character Attributes			Character Location	Scene Attributes			Spatial Relation
		Object	Animal	Person		Background	Light	Style	
Long-Text-to-Image Methods Built on SD-1.5									
StableDiffusion-1.5	19.12	84.40	76.62	80.73	8.66	24.53	69.27	84.47	7.18
LLM4GEN	19.43	82.99	78.00	81.67	9.48	28.32	68.08	50.28	8.04
LLM Blueprint	18.69	81.40	76.25	76.53	**18.40**	56.69	83.28	67.07	14.16
ELLA	25.57	82.38	78.75	80.33	15.04	69.15	83.12	44.17	15.17
LongAlign	25.88	85.54	83.28	83.85	14.12	78.60	87.33	70.49	21.24
PromptDecomposer	**28.21**	84.78	83.24	84.54	16.57	82.45	**92.48**	64.10	20.88
PromptDecomposer$_{(w/\ tuning)}$	25.99	**86.05**	**86.21**	**86.16**	16.21	**90.96**	91.16	**84.93**	**24.47**
State-of-the-art T2I Models with Modern Architectures									
ParaDiffusion$_{(SDXL)}$	28.63	87.40	85.34	84.66	20.62	84.83	93.59	72.16	25.95
StableDiffusion-3.5	39.01	87.60	87.57	89.55	31.91	93.82	92.53	95.31	39.36
FLUX-Dev.	42.02	91.14	89.61	90.23	38.18	**95.73**	96.91	95.28	44.94
Qwen-Image	40.46	90.21	89.13	91.29	40.14	92.00	96.93	91.53	47.02
PromptDecomposer$_{(Qwen)}$	**46.84**	**91.55**	**90.36**	**93.53**	**41.49**	94.62	**97.32**	**95.62**	**49.23**

Table 2: **Quantitative comparisons of generated image quality.** We employ various models to assess images based on semantic alignment and human aesthetics. Best results are marked in **Bold**.

Model	CLIPScore	DenScore	PickScore	VQAScore	HPSv3
Long-Text-to-Image Methods					
ELLA	30.89	20.34	20.72	73.30	6.78
LongAlign	**33.43**	**22.35**	24.43	82.01	**13.26**
PromptDecomposer$_{(SD1.5)}$	32.56	22.24	**24.50**	**83.22**	13.03
State-of-the-art T2I Models					
StableDiffusion-3.5	**34.97**	22.37	21.63	86.12	**13.39**
FLUX-Dev.	33.30	22.56	21.89	86.19	13.17
Qwen-Image	33.85	22.25	20.98	85.02	8.56
PromptDecomposer$_{(Qwen)}$	34.12	**22.93**	**22.04**	**86.21**	12.05

4 EXPERIMENT

4.1 EXPERIMENTAL SETUP

Training. We implement our compositional approach on two pre-trained T2I models to demonstrate generalizability across varying architectures. PromptDecomposer-SD1.5 is built on the widely-used Stable Diffusion-1.5 (SD-1.5) (Rombach et al., 2022) backbone. Consistent with prior work (Hu et al., 2024), we replace the original CLIP text-encoder with T5-XL (Raffel et al., 2020) to accommodate the token length of descriptive paragraphs. We also develop a PromptDecomposer on Qwen-Image (Wu et al., 2025a) to validate our approach on a large-scale modern architecture.

We conduct training on the dataset provided in LongAlign (Bai et al., 2024). This dataset comprises approximately 2 million images re-captioned by LLaVA-Next (Liu et al., 2024) or ShareCaptioner (Chen et al., 2024) to ensure descriptive textual input. We adopt an AdamW optimizer (Loshchilov & Hutter, 2017) with batch size 192 and learning rate $1.0e^{-5}$. This training takes about 20 hours on 4 A100 GPUs. For PromptDecomposer-Qwen, we employ LoRA fine-tuning on the text encoder for approximately 2,500 steps with batch size 24 and learning rate $5.0e^{-5}$.

Evaluation. We adopt the DetailMaster benchmark (Jiao et al., 2025) to comprehensively assess long-text-to-image performance. DetailMaster is a challenging benchmark consists of prompts with 284.89 tokens on average, evaluating generation quality across five dimensions. Specifically, **Character Presence** verifies how many characters in the prompt are successfully generated, and **Character Attributes** measures whether their features (e.g., color, shape) match the text description, with the accuracies computed separately for object, animal, and person categories. **Character Locations** checks if these characters are positioned correctly. **Scene Attributes** evaluates adherence to overall scenic instructions in terms of background, lighting, and style. Finally, **Spatial Relation** quantifies the model's ability to reflect the specified spatial and interactive relationships between the characters.

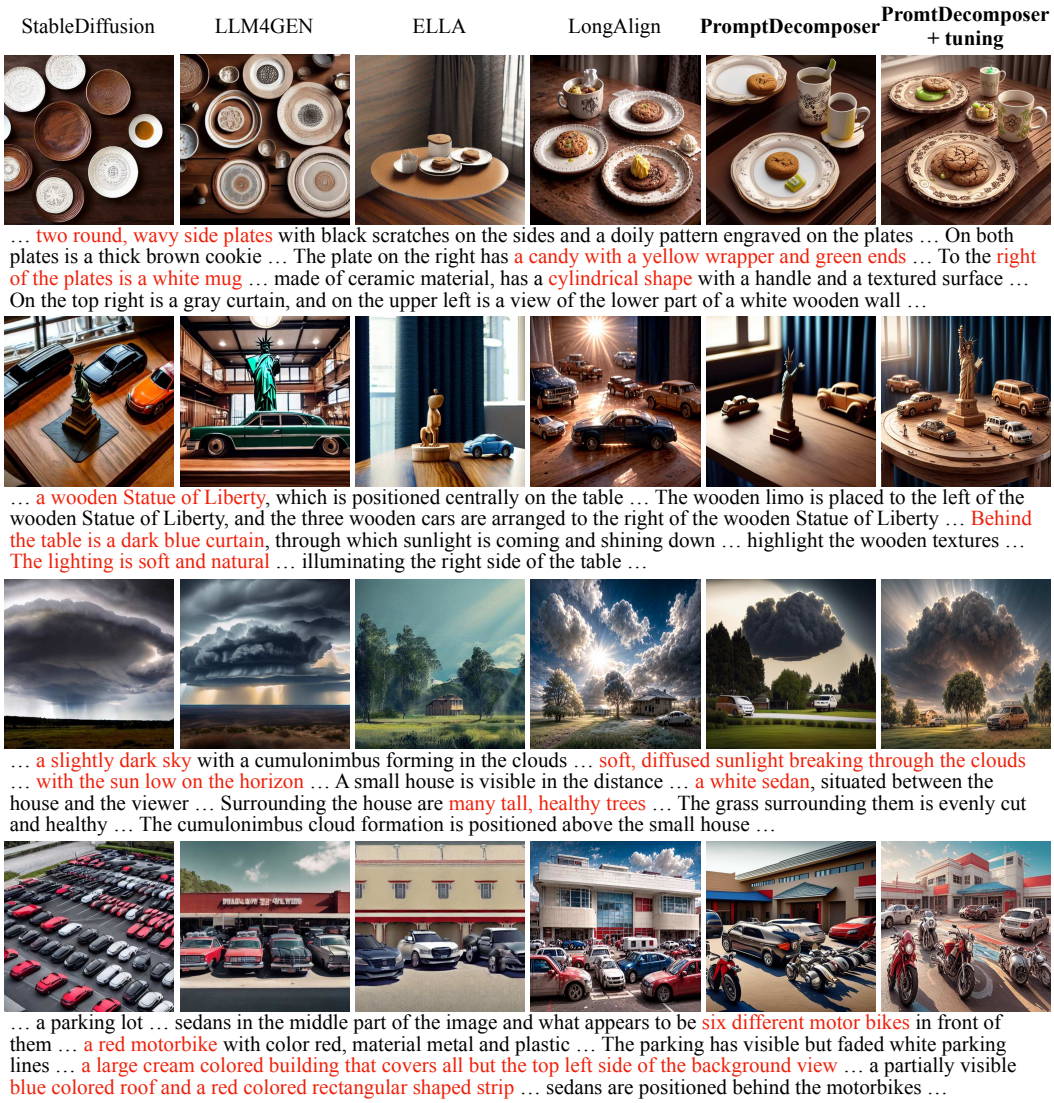

Figure 4: **Long-Text-to-Image Generation Samples.** Our PromptDecomposer accurately captures the attributes and spatial relationships of objects described in the complex scene, with the image quality further enhanced by slightly tuning the T2I model on the decomposed sub-prompts.

4.2 Long-Text-to-Image Generation

Long-prompt Following. Table 1 summarizes the benchmark evaluations of DetailMaster, where we examine the effectiveness of our method against specialized Long-Text-to-Image generation methods and SOTA baselines. PromptDecomposer-SD1.5 outperforms other methods by **2.33% on Character Presence** and **1.53% on Character Location**, demonstrating the efficiency of our PromptDecomposer in processing descriptive paragraphs. Moreover, since the decomposed sub-prompts remain in the pre-trained model's expected input domain, our method can be used in conjunction with other tuning methods to further enhance the results. Our model outperforms LongAlign across all metrics by 4.65% on average using the same tuning method.

Enhancing SOTA Models. Despite employing powerful text-encoders (eg., T5-XXL or Qwen2.5-VL) in modern architectures, Table 1 also shows that more than a half of the characters described in the prompt are completely omitted by even the strongest FLUX model. Our compositional method can further enhance the performance of these SOTA models. For Qwen-Image that leverages an MLLM as its text-encoder, our method improves the **Character Presence by 6.38%** and **Character Attributes by 1.60%** on average. This highlights the long-text-to-image generation as a fundamental challenge originates from the scarcity of long-captioned training data, instead of the text-encoder.

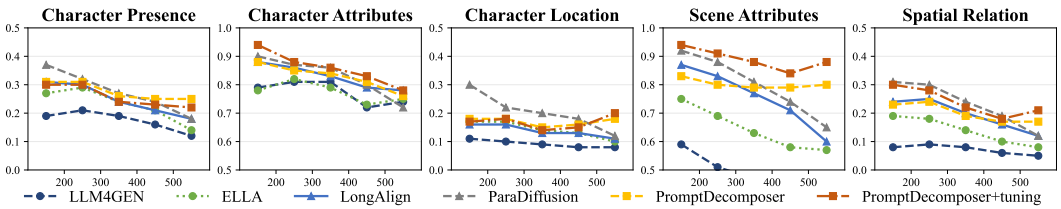

Figure 5: **Generalization to Longer Prompts.** Tuning-based methods (triangle mark) struggle with longer prompts unseen during fine-tuning, and projection-based (round mark) methods suffer from information loss. By decomposing the long input into sub-prompts, our compositional methods (square mark) maintain robust performance across various input lengths.

Figure 6: **Compositional Generalization.** Images are generated from a same prompt rewritten into various lengths. Our method consistently captures the key elements from overwhelmed information.

Image Generation Quality. We employ preference models to assess the generated image quality. We choose three CLIP-based models (CLIPScore (Hessel et al., 2021), DenScore (Bai et al., 2024), PickScore (Kirstain et al., 2023)) to evaluate overall text-image alignment, as well as more powerful MLLMs (VQAScore (Lin et al., 2024), HPSv3 (Ma et al., 2025)) for finer analysis of visual details. Quantitative comparisons are presented in Table 2. On a SD-1.5 backbone, PromptDecomposer-SD1.5 matches the quality of the reward-tuning model LongAlign while consistently surpassing ELLA. Crucially, our approach extends this advantage to SOTA models. **PromptDecomposer-Qwen** achieves the best results among modern baselines on DenScore (**22.93**), PickScore (**22.04**), and VQAScore (**86.21**), outperforming strong baselines including SD-3.5 and FLUX-Dev. Notably, our compositional strategy drastically improves the base Qwen-Image model on **HPSv3 from 8.56 to 12.05**, confirming that our framework effectively resolves the capacity bottleneck in modern foundation models to enhance long-prompt adherence. We present quantitative comparisons in Figure 4 & 7. PromptDecomposer-Qwen successfully interprets the intricate relationships and attributes among six teddy bears while other SOTA models struggling.

4.3 Improved Generalization to Longer Prompts

T2I models are known to generalize poorly with long-prompts because of their scarcity in training data. To evaluate such generalization, we analyze the compared long-text-to-image models' performance according to input prompt length. Specifically, we partition test prompts in DetailMaster into five bins: <200 tokens, 200–300, 300–400, 400–500 and >500 tokens. As shown in Figure 5, LongAlign performs well on prompts under 300 tokens, which constitute the majority of its training data. However, its performance degrades sharply on longer prompts, dropping by up to 30% for those over 500 tokens. Although this degradation is mitigated in projection-based methods, their capacities are constrained by the fixed context window. In contrast, PromptDecomposer maintains robust performance across all prompt lengths despite being trained on the same dataset, achieving an average improvement of **7.4%** on prompts exceeding 500 tokens. This result highlights the improved generalization endowed by compositional generative modeling.

Figure 6 provides a visualization of this improved generalization. We progressively expand a base prompt with more details and compare the generated images. As prompt lengthens, elements such as

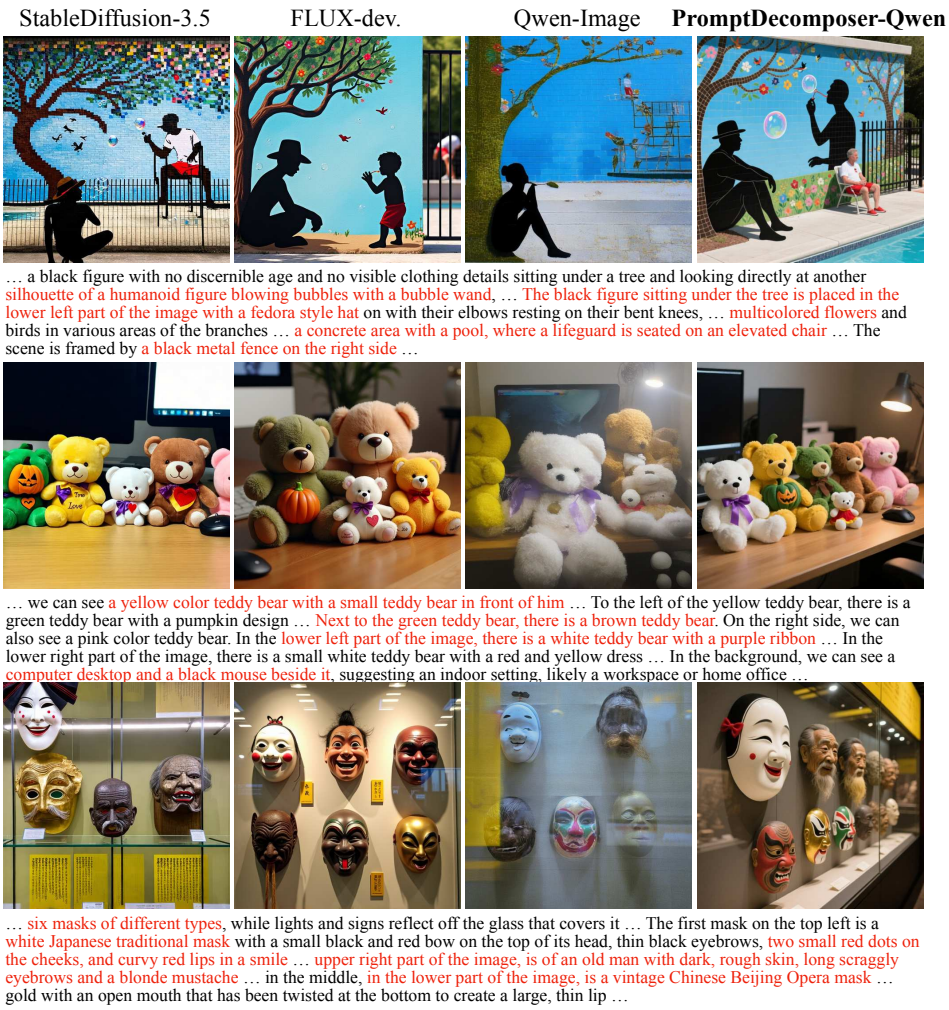

Figure 7: **Image Samples from SOTA Models.** Our PromptDecomposer can further enhance the long-prompt following ability of the SOTA Qwen-Image to accurately render complex scenes.

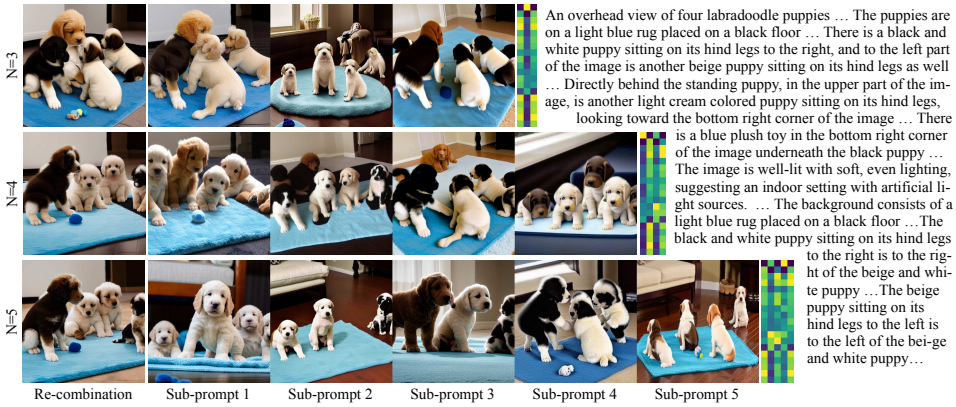

Figure 8: **Semantic Decoupling with Finer Decomposition Granularity.** We visualize the recombination and individual generation results of each sub-prompt by using different number of sub-prompts (N). A smaller N requires each sub-prompt to encode more information and incurs semantic coupling, while a large N allowing each sub-prompt to focus on different aspects.

the house and the yard gradually vanish in LongAlign's outputs. Our model, however, successfully integrates the additional details without overwriting existing concepts, consistently rendering all the key elements regardless of prompt length.

4.4 Ablation Study

Number of Sub-prompts. To investigate the impact of decomposition granularity, we conduct an ablation study on the number of sub-prompts (N). We train two additional PromptDecomposer with N=3 and N=5, alongside our primary version with N=4.

We quantitatively compare these variants' long-prompt generalization ability in Figure 9 (solid bars), which demonstrates a clear improvement from more decomposed sub-prompts. We further visualize this trend in Figure 8, where we can see a finer-grained decomposition (N=5) effectively reduces the semantic load on each sub-prompt. Conversely, smaller N forces each sub-prompt to encode more information, leads to individual generation with high resemblance to the composite output. This is also confirmed in their similarity scores to the input long-prompt: a repeated pattern can be observed in the similarity matrix of N=3, suggesting a similar content in sub-prompts a diminished effect of decomposition.

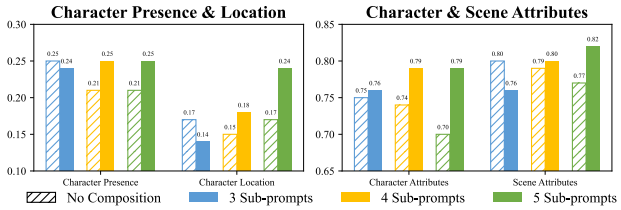

Figure 9: **Improved Generalization from Composition.** We compare model generalization of different numbers of sub-prompts in decomposition (solid bars), as well as the same capacity versions without composition (hatched bars).

Table 3: Ablation Study.

	Character Presence	Character Attributes	Character Location	Scene Attributes	Spatial Relation
w/o Composition	28.98	83.63	16.16	78.92	20.97
w/ Composition	29.49	82.97	17.10	85.34	22.22
Composition via Split	14.01	72.48	6.44	58.69	5.18

Compositionality. We design non-compositional baselines to isolate the benefit of composition. These baselines are created by training PromptDecomposer with one learnable query which corresponds to an unitary long-text-to-image generation model. We increase the query vector's size accordingly to match the total parameter counts. As illustrated in Figure 9, performance of the unitary models (hatched bars) degrades from their compositional counterparts except the case of 3 sub-prompts decomposition. This variant has a diminished compositional effect due to the coarse decomposition (see Section 4.3). Consequently, it enjoys less of the benefit from compositional generation. We also compare the performance gain after tuning in Table 3, where we can see tuning the composed model is more effective. This is because the pre-trained T2I model is more familiar with the decomposed sub-prompts, thus it takes less training to align to these inputs.

5 Conclusion

In this paper, we address the long-prompt generalization problem in T2I models. This fundamental challenge originates from the scarcity of long-captioned images in training data, which hinders T2I models from learning to render the complex narrative flow of a descriptive paragraph.

We propose a compositional approach that leverages pre-trained T2I models' expertise on concise prompts to extend their capacities. We introduce a trainable PromptDecomposer module to directly extract and decompose sub-prompts in the textual representation space. Crucially, this module is trained in an unsupervised manner on the frozen T2I model. By distributing the rich semantic load across multiple sub-prompts, our approach demonstrates superior adherence to detailed instructions and enhanced generalization to increased prompt lengths. Empirically, PromptDecomposer outperforms other long-text-to-image generation methods with a 7.4% improvement on the longest prompts in the DetailMaster benchmark. Furthermore, our approach is also applicable to modern architectures, with non-trivial improvements on the Qwen-Image model which employs an LLM text-encoder.

The primary limitation of our method lies in the concept conjunction lacking explicit spatial control over the generation process. As a result, our method remains data-driven for decomposing the generative distributions. Future work could explore more advanced composition approaches. Another promising direction is to decompose the input prompts adaptively according to their complexity. Although we find a fixed decomposition granularity is robust in the normal-length prompts, using fewer components for concise prompts could improve the efficiency of compositional generation.

Large Language Models Usage Disclosure

LLMs were employed in a limited capacity for writing optimization. Specifically, the authors provided their own draft text to the LLM, which in turn suggested improvements such as corrections of grammatical errors, clearer phrasing, and removal of non-academic expressions. LLMs were also used to inspire possible titles for the paper. While the system provided suggestions, the final title was decided and refined by the authors and is not directly taken from any single LLM output. In addition, LLMs were used as coding assistants during the implementation phase. They provided code completion and debugging suggestions, but all final implementations, experimental design, and validation were carried out and verified by the authors. Importantly, LLMs were NOT used for generating research ideas, designing experiments, or searching and reviewing related work. All conceptual contributions and experimental designs were fully conceived and executed by the authors.

Ethics Statement

This research was conducted in adherence to the ICLR 2026 Code of Ethics. We specifically address the following ethical considerations:

- **Data Usage:** Our work utilizes publicly available datasets that have undergone anonymization to protect individual privacy. We have handled all data in accordance with their specified terms of use.
- **Model Bias:** Our method builds upon existing open-source Text-to-Image models. We acknowledge that these foundational models may reflect societal biases present in their training data. While a full audit of these biases is beyond the scope of our work, we highlight the importance of downstream evaluation for fairness before any real-world application of our method.
- **Societal Impact:** We recognize that Text-to-Image technology has the potential for misuse, such as the generation of misinformation. The aim of our research is to contribute positively to creative applications. We advocate for the responsible development of generative models and support community-wide efforts to establish safeguards against potential harms.

Reproducibility statement

To ensure the reproducibility of our results, we provide our source code of the implementation of our proposed method in the supplementary material. All critical hyperparameters, training configurations and datasets details for our models can be found in Section 4.1. The computational infrastructure used for our experiments is also detailed in this section.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*, 2024. 1, 6, 8, 16

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation.(2023). *URL https://arxiv. org/abs/2302.08113*, 2023. 3

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025. 3, 5

Arwen Bradley, Preetum Nakkiran, David Berthelot, James Thornton, and Joshua M Susskind. Mechanisms of projective composition of diffusion models. *arXiv preprint arXiv:2502.04549*, 2025. 3

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024. 6

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*, 2023. 16

Yilun Du and Leslie Kaelbling. Compositional generative modeling: A single model is not all you need. *arXiv preprint arXiv:2402.01103*, 2024. 2, 3

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019. 4

Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020. 3

Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023. 2, 3

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. 3, 15

Timur Garipov, Sebastiaan De Peuter, Ge Yang, Vikas Garg, Samuel Kaski, and Tommi Jaakkola. Compositional sculpting of iterative generative processes. *Advances in neural information processing systems*, 36:12665–12702, 2023. 3

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 8

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 1, 3, 6, 15

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021. 5

Qirui Jiao, Daoyuan Chen, Yilun Huang, Xika Lin, Ying Shen, and Yaliang Li. Detailmaster: Can your text-to-image model handle long prompts? *arXiv preprint arXiv:2505.16915*, 2025. 1, 3, 6

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 8, 16

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024. 8

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

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

Mushui Liu, Yuhang Ma, Zhen Yang, Jun Dan, Yunlong Yu, Zeng Zhao, Zhipeng Hu, Bai Liu, and Changjie Fan. Llm4gen: Leveraging semantic representation of llms for text-to-image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5523–5531, 2025. 1, 3

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pp. 423–439. Springer, 2022a. 3, 4

Nan Liu, Yilun Du, Shuang Li, Joshua B Tenenbaum, and Antonio Torralba. Unsupervised compositional concepts discovery with text-to-image generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2095, 2023. 3

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022b. 3

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

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. *arXiv preprint arXiv:2508.03789*, 2025. 8

Divyat Mahajan, Mohammad Pezeshki, Charles Arnal, Ioannis Mitliagkas, Kartik Ahuja, and Pascal Vincent. Compositional risk minimization. *arXiv preprint arXiv:2410.06303*, 2024. 3

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. 3, 5

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 3

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. 5, 6

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

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. 1, 3, 6

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 4

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. 3

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a. 16

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*, 2020b. 4

Jocelin Su, Nan Liu, Yanbo Wang, Joshua B Tenenbaum, and Yilun Du. Compositional image decomposition with diffusion models. *arXiv preprint arXiv:2406.19298*, 2024. 3

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a. 5, 6

Weijia Wu, Zhuang Li, Yefei He, Mike Zheng Shou, Chunhua Shen, Lele Cheng, Yan Li, Tingting Gao, and Di Zhang. Paragraph-to-image generation with information-enriched diffusion model. *International Journal of Computer Vision*, pp. 1–22, 2025b. 1

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2096–2105, 2023. 16

Sherry Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. Probabilistic adaptation of black-box text-to-video models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

Zhutian Yang, Jiayuan Mao, Yilun Du, Jiajun Wu, Joshua B Tenenbaum, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Compositional diffusion-based continuous constraint solvers. *arXiv preprint arXiv:2309.00966*, 2023. 3

Shihao Zhao, Shaozhe Hao, Bojia Zi, Huaizhe Xu, and Kwan-Yee K Wong. Bridging different language models and generative vision models for text-to-image generation. In *European Conference on Computer Vision*, pp. 70–86. Springer, 2024. 3

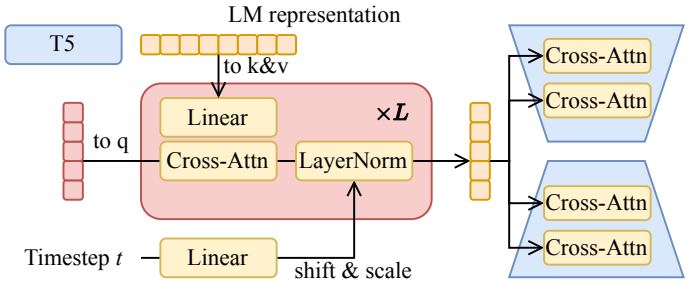

Figure 10: **Architecture Details of our PromptDecomposer.** Our PromptDecomposer is built on the efficient model design of ELLA (Hu et al., 2024), which use a learnable vector to query the long-prompt representation from a LM (T5) through L transformer blocks. The final output is then used as the textual condition in the pre-trained T2I model.

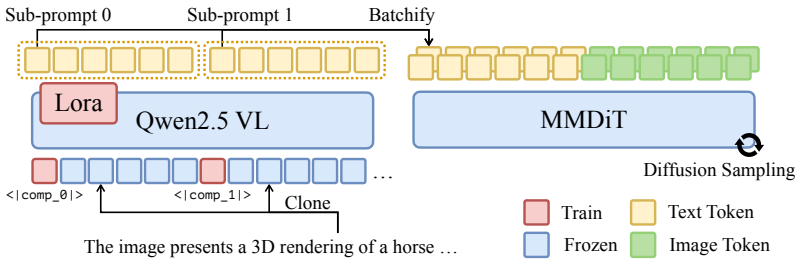

Figure 11: **Applying PromptDecomposer on Qwen-Image.** We leverage the powerful text encoders in modern T2I architecture by applying LoRA and tune these text encoders to directly output decomposed representations using Equation 5.

A Implementation Details

A.1 PromptDecomposer Architecture

Our PromptDecomposer borrows the model design from ELLA (Hu et al., 2024), which contains a series of transformer blocks with a learnable query and the LM-encoded long-prompt as key-value. This architecture can efficiently extract textual condition for pre-trained T2I model from the intricate LM output. Furthermore, ELLA also introduces a time-aware adaptive layer normalization layer. This component leverages the diffusion timestep to modulate the hidden features within each transformer block, as illustrated in Figure 10. The temporal information facilitates the model to extract fine-grained textual conditions that are specific to different stages of the denoising process. The final output vector from these blocks is then sent to the cross-attention layers in T2I model, serving as the textual condition. We inherit most of their design in our PromptDecomposer, except that we remove the time-aware layer normalization on the query inputs which we found leads to mode collapse in the learnable vectors. For PromptDecomposer-SD1.5, we use a total of 6 transformer blocks and 64 tokens in each of the N learnable queries. As for PromptDecomposer-SD3.5, we add an additional cross-attention layer in each transformer block, as illustrated in Figure 3. This additional layer accommodate the extra textual condition to handle the multi text-encoder in StableDiffusion-3.5 (Esser et al., 2024). We only use 3 transformer blocks to balance the overall parameter count in PromptDecomposer-SD3.5, and 128 tokens in each learnable query.

This design requires the hidden dimension of PromptDecomposer to match that of the T2I model's cross-attention layers. When adapting our method to larger T2I models like StableDiffusion 3.5, which features a 4096-dimensional cross-attention layer, this requirement leads to a substantial increase in module size.

As a workaround, we can amortize the parameters in the PromptDecomposer module by leveraging the capacities of the powerful text encoders in modern T2I architectures. For T2I models built on decoder-only LLMs, such as the Qwen-Image with Qwen2.5-VL, we leverage the LLMs' reasoning capabilities to directly generate decomposed representations. We first replicate the input

tokens of the long prompt by N times. Then, we introduce a set of trainable component tokens, $\langle|\texttt{comp}_0|\rangle \dots \langle|\texttt{comp}_{N-1}|\rangle$, which are prepended to each replicated token segment. The expanded prompt is processed as a single contiguous sequence using the causal attention mechanism of Qwen2.5-VL. The output representations are subsequently chunked into N samples corresponding to decomposed representations. Additionally, we introduce LoRAs to the LLM to tune its behavior for this specific task, optimizing the entire system via the compositional objective defined in Equation 5. The entire system is illustrated in Figure 11.

For models relying on T5-XXL (SD3.5 and FLUX), we similarly apply LoRA fine-tuning to the text encoder to directly synthesize decomposed representations. To accommodate the multi-encoder architecture of SD3.5, we let PromptDecomposer processes CLIP representations separately. The outputs from the T5 encoder are chunked into N segments and concatenated with the processed CLIP embeddings. This configuration is highly parameter-efficient, requiring only 160M trainable parameters, a $10\times$ reduction compared to the prior design, while delivering pronounced improvements under the same training budget.

A.2 Reward Tuning Strategy

Since the decomposed sub-prompts representations from our PromptDecomposer remain in the pre-trained T2I model's expected input domain. Our compositional long-text-to-image generation model can be tuned efficiently with other tuning methods. Using reward models for tuning T2I models have been widely explored recently (Kirstain et al., 2023; Wu et al., 2023; Bai et al., 2024). These models are trained on collected human preference data, and are able to measure how well the input image is aligned with text description as well as human aesthetic. We adopt the reward tuning model from LongAlign (Bai et al., 2024), which is optimized on long-caption data to provide more holistic reward signal. We apply the reward tuning algorithm from Clark et al. (2023), which uses gradient-checkpointing to back-propagate the reward signal calculated on the final generation result:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}_0}\left[1 - \mathcal{R}(\boldsymbol{x}_0, \boldsymbol{C})\right] = \mathbb{E}_{\boldsymbol{x}_0}\left[1 - \mathcal{C}_{image}(\boldsymbol{x}_0) \cdot \mathcal{C}_{text}^T(\boldsymbol{C})\right], \tag{6}$$

where $\boldsymbol{x}_0$ is generated from our compositional long-text-to-image model using a DDIM sampler (Song et al., 2020a). For computational efficiency, we generate images with 50 sampling steps in the training loop, where we randomly choose 5 steps to calculate gradients and update model parameters within the memory constraint of our device.

A.3 Inference Efficiency

We analyze the computational overhead of our proposed compositional generation approach. Theoretically, the default 4-component setting implies a $4\times$ increase in total Floating Point Operations (FLOPs). However, practical inference latency does not scale linearly with FLOPs due to hardware parallelism.

By implementing the compositional generation as a batch operation within the denoising loop, we utilize the GPU's parallel processing capabilities more effectively. This larger batch size saturates the tensor cores, mitigating the cost of the additional components. Consequently, the actual inference time increases by a factor of roughly $2\times$, rather than the theoretical $4\times$. On our hardware with A100 GPU, our method runs at 10 iterations/second, compared to LongAlign's speed of 22 iterations/second.

B Evaluation on Standard T2I Benchmark

To verify that our compositional approach effectively handles prompt with standard length, we evaluated its performance on standard T2I benchmarks T2I-CompBench and GenEval. As presented in Table 4, our method demonstrates robust capability in fundamental generation tasks. Specifically, we achieve the best performance on T2I-CompBench, securing the best results in the 'color', 'shape', and 'texture' metrics (ranking first in 3 out of 5 categories). Furthermore, on the GenEval benchmark, our approach remains highly competitive, achieving the second-best result with only a marginal performance difference compared to the leading baseline, LongAlign. These results confirm that our method enhances long-prompt generation capabilities without compromising fidelity or semantic alignment in standard text-to-image tasks.

Table 4: Standard T2I benchmark results on T2I-CompBench and GenEval.

Models	Color	Shape	Texture	Spatial	Numeracy	GenEval
StableDiffusion-1.5	0.3647	0.3768	0.4095	0.5064	0.3197	0.4418
LLM4Gen	0.5084	0.4167	0.5085	**0.6254**	**0.3828**	0.4083
ELLA	0.6269	0.4250	0.5585	0.5713	0.3013	0.4971
LongAlign	0.5654	0.4693	0.5259	0.5698	0.3683	**0.5075**
PromptDecomposer	**0.7113**	**0.5204**	**0.6253**	0.6015	0.3701	0.4960

C Additional Results

As demonstrated in the Table 5, the LoRA-adapted-PromptDecomposers improve their baseline across the DetailMaster benchmark except the some of the scene attribute metrics. This is likely due to the chunk operation on T5 output, which may pose a risk to global information retention Similarly, we observe slight performance degradation of PromptDecomposer on the FLUX model. We also attribute this to the chunking operation on T5 outputs. We hypothesize that these limitations can be addressed through advanced module designs or by scaling training resources to support our original token-resampler design (as used in SD-1.5).

In Figure 13 we present qualitative comparisons between SD-3.5, FLUX, Qwen-Image and LoRA-adapted PromptDecomposer applied on these models. Leveraging the powerful Qwen2.5-VL text encoder backbone, PromptDecomposer-Qwen delivers exceptional long-prompt generation quality with faithful details following. In Table 6 we evaluate images generated on the DetailMaster benchmark using CLIPScore, DenScore, PickScore and HPSv3, where the best results are all obtained by LoRA-adapted PromptDecomposer except on the CLIPScore metric which is unreliable in capturing long-prompt semantics. Moreover, we conduct an user study on the 40 prompts over 400 tokens from the DetailMaster benchmark. Specifically, we shortlist some key concepts from the lengthy prompts, and ask the users to select one best image per group of samples according to human perception quality and adherence to the concepts. Images generated by our PromptDecomposer-Qwen gains the widest range of popularity (23.4%) compared to the strong baseline presented by FLUX-Dev. (18.3%).

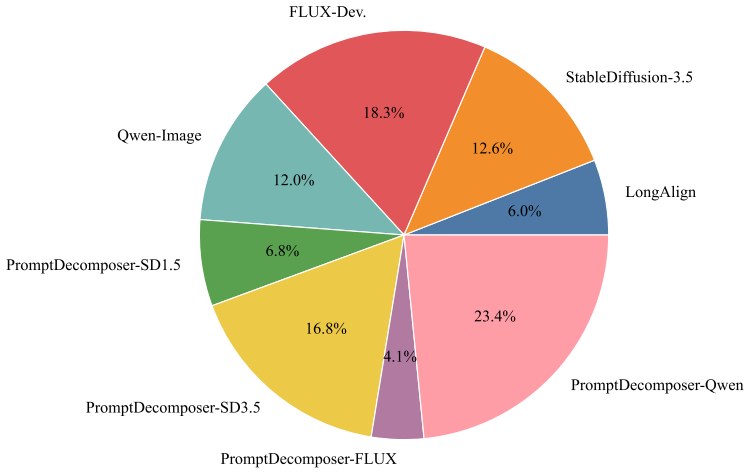

Figure 12: **User Study on the Generation Image Quality.** We shortlist key concepts from each lengthy prompts, and ask the user to select the best image in each batch according to human perception quality and adherence to the key concepts.

D Full text prompts for Image generation

In this section we provide the full long-prompt that is used for generating figures in this paper.

Table 5: DetailMaster benchmark results of **PromptDecomposer** on SD1.5, SD3.5, Qwen-Image and FLUX.

Model	Character Presence	Character Attributes			Character Location	Scene Attributes			Spatial Relation
		Object	Animal	Person		Background	Light	Style	
StableDiffusion-1.5	15.26	24.82	11.48	11.99	7.39	22.02	65.81	83.91	5.75
PromptDecomposer-SD1.5	23.37	28.39	27.95	15.72	15.30	78.17	89.21	74.95	17.33
StableDiffusion-3.5-M	31.19	31.55	32.03	27.54	26.69	87.89	92.32	94.70	28.90
PromptDecomposer-SD3.5	33.03	30.28	35.37	31.21	27.62	87.16	91.96	92.14	31.98
Qwen-Image	31.63	41.01	36.22	24.15	31.01	92.29	96.34	91.41	37.14
PromptDecomposer-Qwen	36.84	38.17	40.50	29.95	32.55	85.87	95.43	94.70	37.80
FLUX-Dev.	34.33	38.49	38.40	32.30	31.62	92.84	95.80	94.70	35.31
PromptDecomposer-FLUX	31.05	36.70	34.46	27.83	26.69	85.69	93.24	91.96	30.29

Table 6: Quantitative evaluations of image generation quality on large-scale T2I models.

Models	CLIPScore	DenScore	PickScore	HPSv3
StableDiffusion-3.5 Medium	34.97	22.37	21.63	13.39
PromptDecomposer-SD3.5	32.97	**25.01**	21.49	**13.52**
Qwen-Image	33.85	22.25	20.98	8.556
PromptDecomposer-Qwen	**34.12**	**22.93**	**22.04**	**12.05**
FLUX-Dev.	33.30	22.56	21.89	13.17
PromptDecomposer-FLUX	32.10	22.36	21.63	12.78

For generating Figure 1:

1. The image presents a 3D rendering of a horse, captured in a profile view. The horse is depicted in a state of motion, with its mane and tail flowing behind it. The horse's body is composed of a network of lines and curves, suggesting a complex mechanical structure. This intricate design is further emphasized by the presence of gears and other mechanical components, which are integrated into the horse's body. The background of the image is a dark blue, providing a stark contrast to the horse and its mechanical components. The overall composition of the image suggests a blend of organic and mechanical elements, creating a unique and intriguing visual.

2. A hyper-detailed, macro shot of a human eye, presented not as an organ of sight, but as a gateway to a lost world of intricate craftsmanship. The iris is a masterfully crafted, antique horological mechanism, a complex universe of miniature, interlocking gears and cogs made from polished brass, copper, and tarnished silver. Each metallic piece is exquisitely detailed, with tiny, functional teeth that seem to pulse with a slow, rhythmic, and almost imperceptible life. The vibrant color of the iris is replaced by the warm, metallic sheen of the gears, with ruby and sapphire jewels embedded as tiny, gleaming pivots. At the center, the pupil is not a void but the deep, dark face of a miniature clock, its impossibly thin, filigreed hands frozen at a moment of profound significance. The delicate, thread-like veins in the sclera are reimagined as fine, coiling copper wires, connecting the central mechanism to the unseen power source at the edge of the frame. The entire piece is captured under a soft, focused light that highlights the metallic textures and casts deep, dramatic shadows within the complex machinery, suggesting immense depth. The background is a stark, velvety black, ensuring nothing distracts from the mesmerizing, mechanical soul of the eye.

3. A sleek, enigmatic feline, a cat of indeterminate breed, is the central figure, poised in a state of serene contemplation. Its body is not of flesh and bone, but meticulously sculpted from a complex lattice of polished, interlocking obsidian shards. Each piece is perfectly fitted against the next, creating a mosaic of deep, lustrous black that absorbs the light. The cat's form is defined by the sharp, clean edges of these volcanic glass fragments, giving its natural curves a subtle, geometric undertone. Glimmering veins of molten gold run through the cracks between the shards, glowing with a soft, internal heat that pulses rhythmically, like a slow heartbeat. These golden rivers trace the contours of the cat's muscles and skeleton, outlining its elegant spine, the delicate structure of its paws, and the graceful curve of its tail. Its eyes are two brilliant, round-cut rubies, catching an unseen light source and casting a faint, crimson glow. The whiskers are impossibly thin strands of spun platinum, fanning out from its muzzle with metallic precision. The entire figure rests upon a simple,

Table 7: Image Quality Assessment on LongAlign dataset.

Metrics	SD-1.5	LLM4GEN	ELLA	LongAlign	PromptDecomposer
CLIPScore	0.3462	0.3362	0.3310	0.3568	0.3519
DenScore	0.2047	0.2028	0.2112	0.2587	0.2596
PickScore	0.2083	0.2069	0.2052	0.2306	0.2308
HPSv3	9.174	7.620	5.631	12.72	12.61

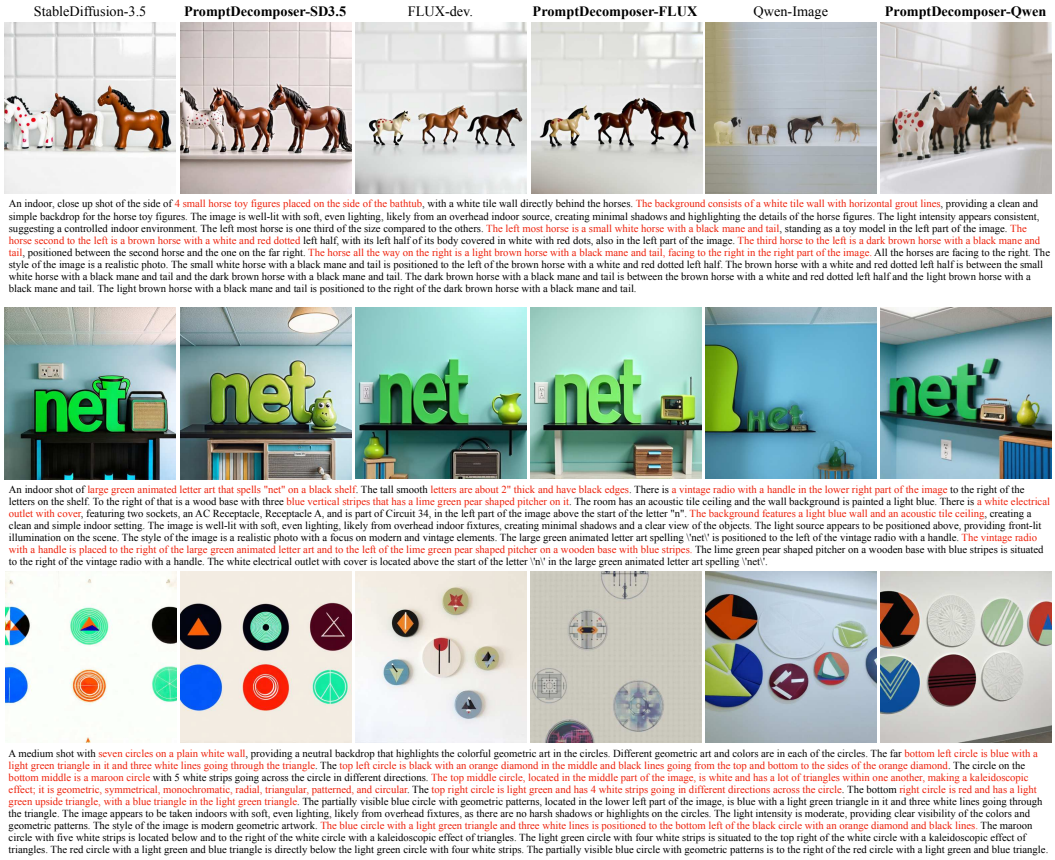

Figure 13: **Qualitative Samples on DiT Models.** Leveraging the powerful text encoders in modern architectures, our PromptDecomposer effectively interpret object attributes and relationships in an intricate paragraph.

unadorned, and dimly lit surface, ensuring that all focus remains on the cat's extraordinary construction—a masterful fusion of natural grace and exquisite, dark craftsmanship.

For generating Figure 4:

1. A high angle shot of a brown wooden bench with several dishes on top of it. In the center and on the left are two round, wavy side plates with black scratches on the sides and a doily pattern engraved on the plates. On both plates is a thick brown cookie that's been crosscut at the top, located in the middle part of the image. The plate on the right has a candy with a yellow wrapper and green ends. To the right of the plates is a white mug with whipped cream on top that is similar to the glass plates. The cup, made of ceramic material, has a cylindrical shape with a handle and a textured surface. The white whipped cream on top is frothy and has an embossed design. Surrounding the wooden bench is a dark brown wooden floor. On the top right is a gray curtain, and on the upper left is a view of the lower part of a white wooden wall. The image is taken indoors with soft, warm lighting, likely from an overhead source, creating a cozy and inviting atmosphere. The lighting is evenly distributed,

with no harsh shadows, suggesting a relaxed time of day, possibly evening. The style of the image is a realistic photo with a warm, homely aesthetic. The brown wooden bench supports the two round, wavy side plates with black scratches and a doily pattern, which are placed side by side. The thick brown cookies crosscut at the top are positioned on top of the two round, wavy side plates, with one cookie on each plate. The candy with a yellow wrapper and green ends is located on the right plate, next to the thick brown cookie. The white mug with whipped cream on top is situated to the right of the two round, wavy side plates. The two round, wavy side plates are adjacent to each other, with the plate containing the candy being closer to the white mug with whipped cream on top.

2. An indoor top-down view of a wooden Statue of Liberty, which is positioned centrally on the table, covering a black marking on the table, on a wooden table with 3 wooden cars and 1 wooden limo next to it. The wooden limo is placed to the left of the wooden Statue of Liberty, and the three wooden cars are arranged to the right of the wooden Statue of Liberty. On the table, the black marking on the table is partially hidden by the wooden Statue of Liberty in the upper part of the image. Behind the table is a dark blue curtain, through which sunlight is coming and shining down on the right side of the table, casting a soft glow and creating gentle shadows that highlight the wooden textures. The lighting is soft and natural, suggesting it is daytime with sunlight filtering through the curtain, illuminating the right side of the table. The dark blue curtain is located behind the table, indicating it is not on the same plane as the objects on the table, and it is positioned at the back of the image. The style of the image is a realistic photo.
3. A long-shot view of a slightly dark sky with a cumulonimbus forming in the clouds, allowing rays of sunlight to pierce through, creating a striking contrast against the darkened landscape. The sky is bright blue, and the cumulonimbus cloud formation is a dark blue and gray, with soft, diffused sunlight breaking through the clouds, suggesting it is either early morning or late afternoon, with the sun low on the horizon. A small house is visible in the distance; it has tan panels, and it has a white metal roof. Parked in the lower right part of the image in front of the house is a white sedan, situated between the house and the viewer. Surrounding the house are many tall, healthy trees that are mostly shrouded in shadow; these trees have green leaves, a broad canopy, dense foliage, and provide natural shade, located around and behind the small house, creating a natural border. The grass surrounding them is evenly cut and healthy. The scene is somewhat dark, with rays of sunlight shining through the gathered clouds to illuminate the sky from above, enhancing the tranquil yet moody atmosphere. The cumulonimbus cloud formation is positioned above the small house, and the rays of sunlight are directed towards the area above the house and trees, capturing natural lighting and atmospheric conditions in a realistic photo style.
4. A front view of a parking lot with several vehicles parked including two dark colored sedans in the middle part of the image and what appears to be six different motor bikes in front of them. The bikes seem to range from a red motorbike with color red, material metal and plastic, typical features include two wheels, handlebars, seat, engine, exhaust pipe, and headlight in the right part of the image, a white motor bike in the left part of the image, a silver motor bike, another white motor bike, another silver motorbike that is silver in color, and another silver motorbike that is also silver in color. The parking has visible but faded white parking lines, and behind all of the vehicles are two handicap parking signs. Behind the handicap signs is a large cream colored building that covers all but the top left side of the background view, it has a partially visible blue colored roof and a red colored rectangular shaped strip that passes along the view of the building a couple of feet below the blue roof. The background features a large cream-colored building with a blue roof and a red strip, partially obscured by the parked vehicles. Two handicap parking signs are visible on the building's facade. The image appears to be taken during the day under natural light, with the light source positioned overhead, creating soft shadows beneath the vehicles. The lighting is bright and even, suggesting a clear sky with no direct sunlight causing harsh shadows. The style of the image is a realistic photo. The two dark colored sedans are positioned behind the motorbikes, with one slightly to the left and the other to the right. The red motorbike is to the right of the other motorbikes, closer to the right sedan. The white motorbike is to the far left, with the silver motorbike next to it. The another white motorbike is positioned between the first white motorbike and the silver motorbikes. The another silver motorbike is next to the another white motorbike, and the last silver motorbike is next to the red motorbike.

The cream colored building with a blue roof and red strip is behind all the vehicles, with the handicap signs in front of it.

For generating Figure 6 and Figure 8:

1. A high-angle side view of a black Yamaha Virago motorcycle facing the right side of the image parked on an black asphalt surface. The front of the motorcycle is turned slightly toward the top right corner of the image. The fenders, the fuel tank, and the handles of the motorcycle are black. The motorcycle has a brown leather seat. The engine, exhaust pipes, and handlebar are gray silver. There is a red tail light attached to the fender over the top of the rear wheel. The Virago logo is on the side of the gas tank. The motorcycle is facing a lawn area on the side of a house visible at the top of the image. There is a patch of grass and a walkway leading to a gray door near the top right corner of the image, there is a window on each side of the door. There are two blue chairs in the top right corner of the image. Visible in the top left corner of the image is the right side of the front of a gray Toyota C-HR SUV with metallic paint, a compact SUV shape, sleek headlights, a Toyota emblem, and a modern design. The background features a residential setting with a gray house, a lawn, a walkway, and two blue chairs near the top right corner. A gray Toyota C-HR SUV is partially visible in the top left corner. The image is taken outdoors under natural daylight, with soft lighting conditions suggesting it could be morning or late afternoon. The light source is positioned to the side, creating gentle shadows and highlighting the motorcycle's details. The style of the image is a realistic photo. The black Yamaha Virago motorcycle is positioned in front of the lawn area with a gray door and windows, indicating it is closer to the viewer than the house. The gray Toyota C-HR SUV is located to the left of the black Yamaha Virago motorcycle, suggesting it is parked parallel to the motorcycle but further away from the house. The two blue chairs are situated to the right of the lawn area with a gray door and windows, showing they are placed on the side of the house away from the motorcycle and the SUV. The lawn area with a gray door and windows is between the motorcycle and the two blue chairs, establishing it as a central point in the spatial arrangement of the scene.
2. An overhead view of four labradoodle puppies, three puppies are sitting and one puppy is standing with its right paw resting against the white barrier at the bottom of the image. The puppies are on a light blue rug placed on a black floor. The puppy standing is a beige and white puppy with curly fur, dark eyes, a small nose, and a fluffy appearance, its paw extended. There is a black and white puppy sitting on its hind legs to the right, and to the left part of the image is another beige puppy sitting on its hind legs as well. Directly behind the standing puppy, in the upper part of the image, is another light cream colored puppy sitting on its hind legs, looking toward the bottom right corner of the image. The three puppies in the front are looking up, the puppy behind them is looking toward the bottom right corner of the image. There is a blue plush toy in the bottom right corner of the image underneath the black puppy. The rug the puppies are on is not laying completely flat on the ground, its unintentionally folded up in some areas and folded over itself in the top right corner of the image. The background consists of a light blue rug placed on a black floor, with the rug showing some unintentional folds and overlaps. A blue plush toy is visible in the bottom right corner under the black puppy. The image is well-lit with soft, even lighting, suggesting an indoor setting with artificial light sources. The light appears to be front-lit, as there are no harsh shadows on the puppies. The style of the image is a realistic photo. The beige and white puppy standing with its right paw resting against the white barrier is in front of the light cream colored puppy sitting on its hind legs in the back. The black and white puppy sitting on its hind legs to the right is to the right of the beige and white puppy standing with its right paw resting against the white barrier. The beige puppy sitting on its hind legs to the left is to the left of the beige and white puppy standing with its right paw resting against the white barrier. The light cream colored puppy sitting on its hind legs in the back is behind the beige and white puppy standing with its right paw resting against the white barrier. The black and white puppy sitting on its hind legs to the right is next to the beige puppy sitting on its hind legs to the left.