PromptCoT: Align Prompt Distribution via Adapted Chain of Thought

Anonymous Author(s)
Affiliation
Address
email

Abstract

Diffusion-based generative models have exhibited remarkable capability in the production of high-fidelity visual content such as images and videos. However, their performance is significantly contingent upon the quality of textual inputs, commonly referred to as "prompts". The process of traditional prompt engineering, while effective, necessitates empirical expertise and poses challenges for inexperienced users. In this paper, we introduce PromptCoT, an innovative enhancer that autonomously refines prompts for users. The design of PromptCoT is based on the observation that, prompts resembling textual information corresponding to high-quality images within the training set tend to yield superior generation performance. As such, we fine-tune the pre-trained Large Language Models (LLM) using a curated text dataset comprising solely of high-quality visual content descriptions. By doing so, the LLM becomes capable of capturing the distribution of high-quality training texts, enabling it to generate aligned continuations and revisions to boost the original texts. Nonetheless, one drawback of pre-trained LLMs is their tendency to generate extraneous or irrelevant information. To enhance the alignment between the original text prompts and the refined counterparts, we leverage the Chain-of-Thought (CoT) mechanism. CoT can extract and amalgamate crucial information from the aligned continuation and revision, enabling reasonable inferences based on the contextual cues to produce a more comprehensive and nuanced final output. Considering computational efficiency, instead of allocating a dedicated LLM for prompt enhancement to each individual model or dataset, we integrate adapters that facilitate dataset-specific adaptation, leveraging a shared pre-trained LLM as the foundation for this process. By fine-tuning these adapters independently, we can adapt PromptCoT to new datasets with minimal increase in training cost and memory usage. We assess the performance of PromptCoT on widely-used latent diffusion models for image and video generation to validate the effectiveness. The results demonstrate significant improvements in key performance metrics.

1 Introduction

In recent years, deep generative models have made notable advancements, specifically with the introduction of diffusion probabilistic models (DPMs). These models have exhibited exceptional capabilities in generating a wide range of visually compelling and high-fidelity visual contents, such as images and videos, as evidenced by notable contributions in the literature [37, 12, 38, 36, 7, 28, 32, 30].

By harnessing textual inputs as conditional guidance, diffusion models have the ability to generate visual outputs that align with the corresponding input text, utilizing an iterative denoising procedure. This technological advancement has paved the way for revolutionary applications, including notable examples such as DALL-E 2 [28], Stable Diffusion [30], MagicVideo [50], among others.

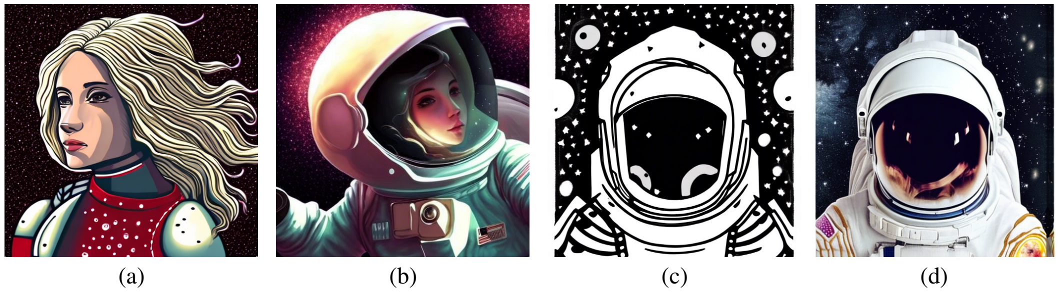

Figure 1: **Impacts of PromptCoT.** (a) and (c) shows the images generated with the original text prompts, and (b) and (d) show the images generated with the text prompts refined by PromptCoT. The text prompt for (a), (b), (c) and (d) are: 1) "highly detailed portrait of a hopeful pretty astronaut lady with a wavy blonde hair, by Jamini Roy , 4k resolution, nier:automata inspired, bravely default inspired, vibrant but dreary but uplifting red, black and white color scheme!!! ((Space nebula background))" ; 2) "Astronaut portrait of Silica from the game Bravely Default II by Jamini Roy", and 3) "highly detailed portrait of a hopeful pretty astronaut lady with a wavy blonde hair, by Pablo Picasso, 4k resolution, nier:automata inspired, bravely default inspired, vibrant but dreary but uplifting red, black and white color scheme!!! ((Space nebula background))",and 4)"Portrait Of A Beautiful Astronaut Girl Canvas Art Print" respectively.

Nevertheless, the quality of the generated content is intricately tied to the caliber of the textual prompts provided to the generative model. Human inputs tend to be informal and straightforward, which may impede the expression of the desired scene with the desired level of depth. Additionally, the text encoder within the generative model may not fully comprehend the semantic nuances present in the human-generated text, resulting in notable disparities between the encoded textual guidance and the user's intended meaning. Diffusion probabilistic models (DPMs) are commonly trained on extensive text-vision pairs acquired through web-scraping techniques [35]. Our observation reveals that the distribution of the text dataset might not be congruent with the linguistic style employed by layman users. Furthermore, even in cases where the training text data aligns with the desired style, the quality can exhibit substantial variations due to the presence of meaningless words or extraneous information within the text data. This intricacy further complicates the establishment of a clear and unambiguous mapping between the text and the corresponding image.

As a result, there is an immediate imperative to develop a methodology that can effectively align prompts, consequently augmenting the image generation performance in generative models. Although data cleaning and model fine-tuning have been considered potential solutions, these methods often entail drawbacks such as high costs, instability, and time intensiveness. Another alternative is manual prompt engineering, which involves refining prompts to optimize generation performance. However, this empirical task traditionally demands the expertise of experienced professionals, thereby posing a significant challenge for individuals lacking relevant experience.

In our study, we observe a noticeable trend that prompts, which resemble those found in the training set, usually lead to superior generative performance. Stemming from this observation, we propose PromptCoT, a novel prompt booster that leverages the power of pre-trained Large Language Models (LLMs) and incorporates the Chain-of-Thought (CoT) mechanism to learn high-quality prompt expressions from the training texts of generative models. Specifically, we carry out the fine-tuning of LLaMA [40], a widely-used pre-trained Large Language Model, on two distinct datasets we've prepared. With a text-continuation dataset that appends aligned details to original prompts, and a text-revision dataset that rewrites original prompts to aligned prompts, we enable LLaMA to refine prompts that better match the distribution of the text data used for training the diffusion models. To further enhance the performance of LLMs by combining the advantages of both text-continuation and text-revision, we construct a dataset using the CoT mechanism assisted by ChatGPT. This CoT dataset is designed to enable LLMs to reason and generate text that follows a logical and coherent flow. By fine-tuning LLMs on this CoT dataset, we can enhance their reasoning ability and augments their capacity to generate high-quality text that is both contextually relevant and logically coherent.

To accommodate the varying training sets of different generative models, we incorporate a parameter-efficient adaptation design into the training pipeline of PromptCoT, augmenting a pre-trained base booster with specific lightweight adapters that are capable of aligning text distributions for various generative models across multiple tasks. We demonstrate the effectiveness of PromptCoT through extensive experiments on widely-used latent diffusion models for image and video generation, showing significant improvements in key performance metrics such as Fréchet Inception Distance, aesthetic score, and CLIP-similarity.

Our main contributions are:

• We propose PromptCoT, an innovative prompt refiner that aligns input prompts with the text distribution employed during the training of diffusion models. By accomplishing this alignment, PromptCoT effectively activates generative models and enhances their performance.

• We explore a new optimization scheme for improving prompt quality by leveraging the power of pre-trained LLMs and CoT mechanisms. And we construct datasets to facilitate the learning of high-quality prompt distribution from the training texts of generative models.

• We demonstrate that allocating a dedicated Large Language Model (LLM) for each diffusion model is not a requirement. Instead, we propose an innovative scheme where a set of lightweight adapter weights suffices for each dedicated diffusion model. These adapters can share a shared base pre-trained LLM, resulting in a considerable reduction in memory footprint.

• We show the effectiveness of PromptCoT through extensive experiments on widely-used latent diffusion models for image and video generation, showing significant improvements in key performance metrics.

2 Related Work

2.1 Text-to-Image Generative Models

Text-to-Image Generative Models operate by taking natural language descriptions as input and generating corresponding images as output. One of the recent popular model is DALL·E 2 [29]. It utilize CLIP [26] to align the text and image embeddings. By conditioning the diffusion probabilistic generator on the textual embedding, DALL·E 2 is able to produce photorealistic images that correspond to the given textual description. Later, Google's Imagen [32] and Parti [46] were proposed by gradually simulating the spread of noise into the original image to reveal the desired image. Specifically, both Parti and Imagen combine autoregressive and diffusion. The application of diffusion probabilistic models has also been extended to the domain of video generation. The Video Diffusion Model [13], built upon the foundations of diffusion models, enables the sequential generation of high-quality video frames. To address the substantial computational requirements associated with video generation, MagicVideo [51] was introduced, combining latent diffusion and attention models. MagicVideo utilizes a frame-wise lightweight adapter and an attention module to effectively adjust the image-to-video distribution and capture temporal dependencies across frames.

2.2 Large Language Models

Large Language Models (LLMs) are powerful deep learning models for various natural language processing tasks. The most popular LLMs are the GPT [27, 5] series models developed by OpenAI, which are based on the decoder component of the transformer architecture. Another LLM is Meta's OPT [49], which is open-sourced and performs similarly in performance to GPT-3. However, GPT-3's massive size of 175B parameters requires significant computing power and resources, which makes it challenging for researchers to explore. In contrast, LLaMA [40, 41], StableLM [2], as well as the instruction-following Alpaca model [39] are smaller and more performant, achieve comparable results to ChatGPT with far fewer parameters (7B). For specific tasks like conversational applications, ChatGLM [47, 9] can generate coherent and contextually relevant responses in dialogue systems.

2.3 Parameter-Efficient Fine-Tuning

The goal of parameter-efficient fine-tuning is to attain comparable performance to fine-tuning on a specific downstream task while using the fewest trainable parameters possible. According to [1], common pre-trained models generally have a very low intrinsic dimension, and LoRA [15] learns low-rank parameterizations to enhance tuning efficiency based on that. Except reducing the number of parameters needed for fine-tuning, other approaches try to attach pre-trained parameters to reduce training time. Adapter training [14, 24] utilizes dynamic pre-trained adapters for different tasks and languages to reduce adaptation time. Compacter [21] combines both concepts and builds on top of adapters, low-rank optimization, and parameterized hypercomplex multiplication layers.

2.4 Prompt Engineering

Prompt Engineering is to optimize the outputs of language models with specific input prompts [4, 33, 20, 8]. Discrete text prompts [16] serve as starting points for the model's language generation, and are used to generate responses in dialogue systems. Beyond discrete prompts, [17, 43] explores prompt tuning to learn soft prompts to perform specific downstream tasks, which provide more context-aware guidance to the model. [25] extends the idea of learning soft prompts and demonstrates that the implicit factual knowledge in language models was underestimated. Given that manually designing prompts can be cumbersome, automatically generating prompts gives a chance avoid intensive labor and enhance efficiency [33, 34]. [10] proposes to generate all prompt candidates and selectively incorporate them into each context using a refined strategy. [11] introduces a more efficient method to construct prompts with several sub-prompts that employs prompt tuning with rules without searching. Overall, prompt engineering is an efficient approach that helps bridge the gap between pre-training and fine-tuning.

2.5 Chain-of-Thought

Chain-of-Thought is a specialized tool designed for the task of multi-step reasoning and decision-making [44]. The traditional prompting method [4] performs poorly when it comes to tasks that require reasoning abilities. Inspired by the concept of using intermediate steps to solve reasoning problems [19, 6], the chain of thought method mimics a step-by-step thinking process and breaks down multi-step problems into intermediate steps, enabling the model to deduce more accurate results [23]. Additionally, [52] address the challenge of dealing with tasks that are more complex than example prompts, and proposes the least-to-most prompting approach which breaks down complex problems into smaller and easier subproblems. Moreover, [42] introduces self-consistency as a replacement for the greedy decoding algorithm, which samples and selects the most consistent reasoning paths to replace the greedy set.

3 Method

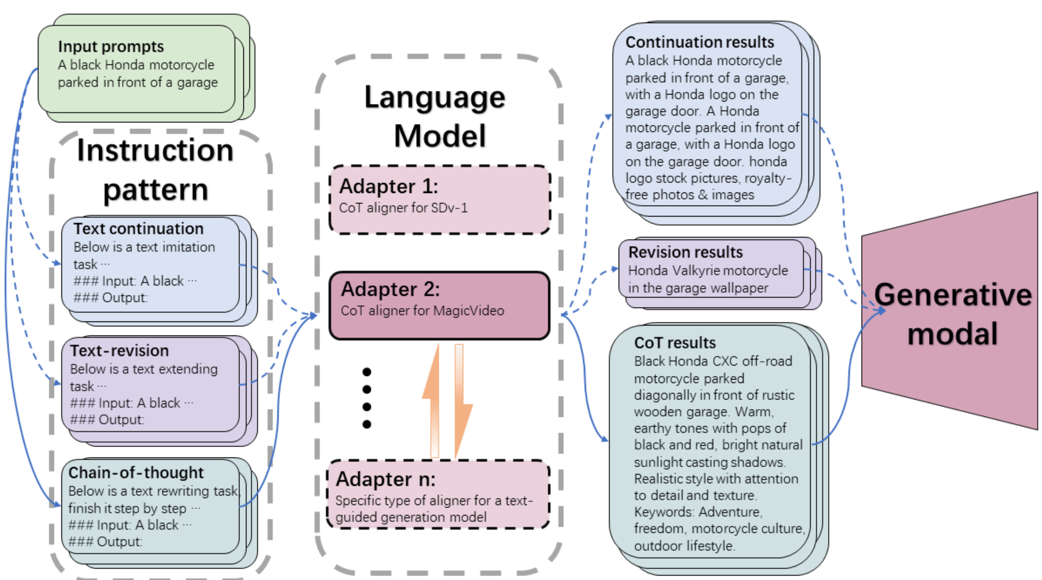

Figure 2: Pipeline of PromptCoT. (Left) We build three types of instruction patterns for training. (Middle) We utilize adapters for multi-task adaptation. (Right) Results of t-continue, t2t booster and PromptCoT.

3.1 Overview

Text-to-image diffusion models serve as an illustrative example for showcasing the functionality of PromptCoT. However, it is important to note that the same methodology can be extended and applied to other diffusion-based generative models, including text-to-video and various other domains. In the context of training text-to-image diffusion-based models, which involve image-text pairs and employ an iterative denoising process to reconstruct images based on corresponding prompts, our hypothesis posits that prompts aligned with high-quality images within the training set are more inclined to yield visually superior outputs. We randomly select 5 sets of 50 prompts corresponding to images with varying levels of quality from the Stable Diffusion training set, LAION [35], for image generation. The aesthetic score, an image quality metric introduced by [31], is used to represent the quality of individual images. As shown in Table 1, the generation performance is highly related to the prompts corresponding to the original image quality. For convenience, we refer to them as "**high-quality prompts**". In the following sections, we explain the key components of PromptCoT,

Table 1: Comparison of Aesthetic Scores between Generated Images and Corresponding Training Images.

	Aesthetic Score			
Training images	4-5	5-6	6-7	7-8
Generated images	5.2	5.5	6.1	6.3

which is a prompt booster that can align input prompts with high-quality prompts in the training set, and in turn, improve generation performance.

3.2 Aligning Prompt Distribution with LLM

LLMs are extremely powerful tools that are capable of generating human-like language and completing tasks such as translation, summarization, question answering, etc. They are trained on massive amounts of text data and can learn from unstructured data to generalize to new tasks and domains. LLMs can also be fine-tuned on specific tasks with relatively small amounts of task-specific data, making them highly versatile. In this paper, we leverage this ability to align the distribution of high-quality prompts via fine-tuning a popular LLM LLaMA [40], on text continuation and revision tasks. To fine-tune LLaMA on text continuation, we use an instruction tuning template that includes incomplete text descriptions and a goal to provide a compelling continuation. The instruction tuning template is shown in Figure 3. We feed truncated text prompts placed in the *input* field to the LLM, supervised by the complete prompts. This enables the LLM to generate continuations containing more details.

```
Below is a text continuation task.
You will be given an incomplete text description.
Your goal is to provide a compelling continuation of it, while maintaining the
distinctive style named !LAION-6plus-style!.
### Input:{input}
### Output:
```

```
Original Prompt=> A group of horses are grazing in the field.

Continuation=> A group of horses are grazing in the field. A lone tree stands in
the center of the field. Storm clouds are entering from the left of the picture.
```

Figure 3: Template of text-continuation dataset (Up) and corresponding output (Bottom).

For text revision, we train the LLM to map human-like input texts to high-quality prompts. However, acquiring a large amount of human-written input text can be costly. Therefore, we leverage image captions from BLIP as a low-cost source of "human-like" input texts. The details of collecting and filtering data pairs are described in the later section. For training, we construct the instruction tuning template in Figure 4. The training pipeline is similar to continuation, but with the input being human-like prompts. As a result, we obtain a booster capable of performing revision tasks.

3.3 Enhancement with CoT

Instruction tuning enables the LLM to add details and align text distribution, however, it tends to generate extraneous information that degrades performance. As such, we introduce the Chain-of-Thought (CoT) mechanism in the pipeline to address this issue. We set up five steps to make the

```
Below is a text revision task.
You will be given a text description.
Try your best to rewrite it in a distinctive style named !LAION-6plus-style!.
### Input:{input}
### Output:
```

```
Original Prompt=> People walking towards a parked bus on the street

Revision=> In this picture taken on January 6, 2016 a man walks past a bus
covered with snow at a bus stop in Kolomenskoye cemetery during heavy
snowfall in Moscow. (Photo by Kirill Kudryavtsev/AFP Photo)
```

Figure 4: Template of text-revision dataset (Up) and corresponding output (Bottom).

LLM yield the expected production: (i) Extract key information from the original prompt, such as visual medium and main elements, (ii) Leverage the text-continuation model to append reasonable details, (iii) Extract additional concepts (for example, the color scheme) from the extended prompt and emphasize crucial concepts, (iv) With improved key information and crucial concepts, the LLM can generate a fluent prompt, remaining to be aligned, (v) Leverage the text-revision model to align prompts to the specific distribution. This mechanism extracts and amalgamates crucial information from the aligned continuation and revision, enabling reasonable inferences based on the contextual cues. As a result, a more comprehensive and nuanced final output is produced.

3.4 Multi-task Adaptation

As the training set of different generative models can vary greatly, one approach to adapt to these new datasets is to fine-tune the entire LLM on the task-specific dataset. However, LLMs are typically models with billions of parameters, and allocating a dedicated LLM to each individual model proves impractical due to computational constraints. Moreover, there are plenty of text-to-image generative models trained on different datasets, and a single LLM cannot cover a diverse distribution of these datasets. As an alternative, we integrate adapters that facilitate dataset-specific adaptation, leveraging a shared pre-trained LLM as the foundation for this process. Adapters are lightweight modules that can be independently fine-tuned and subsequently added to the base model. Keeping adapters instead of the whole model significantly reduces memory usage, while enabling the adaptation of the LLM to different datasets.

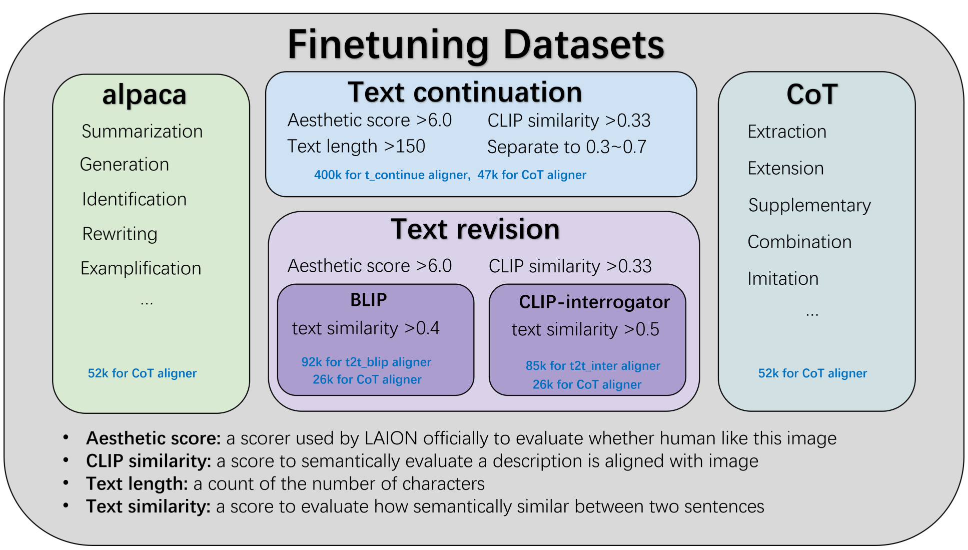

Figure 5: Composition of fine-tuning tasks including text-continuation, text-revision, text-CoT, and self-instruction of Alpaca.

3.5 Dataset Preparation

We build three types of datasets: text-continuation, text-revision, and text-CoT.

Text-continuation dataset. To create this dataset, we filter high-quality prompts from the training data of existing generative models, using criteria such as high CLIP similarity and proper length. In the case of the LAION dataset, we also consider aesthetic scores to ensure a higher quality of prompts. Once high-quality prompts are identified, we truncate a portion of the text, with the remaining front part assigned as input data. The LLM is then trained to generate the missing information and complete the text. This process enables the LLM to learn how to effectively continue text prompts in a manner that is consistent with the style and context of the original text.

Text-revision dataset. The dataset consists of human-like texts and corresponding high-quality prompts which are described in the text-continuation dataset. To acquire human-like prompts, we leverage BLIP and CLIP-interrogator for image captioning. Furthermore, we calculate the text distance with the text encoder of CLIP, ensuring a score greater than 0.4 to guarantee semantic relevance between the two prompts.

Text-CoT dataset. We use GPT-3.5-Turbo to build a task-specific dataset. Initially, we design a step-by-step interaction with GPT-3.5-Turbo to extract and guide the prompt booster to finish the alignment task, due to the fact that CoT is still difficult for alpaca with a simple finetuning on datasets above. Following the alpaca's thought, 52k pairs are all generated from gpt-3.5-turbo.

4 Experimental Results

In this section, we first introduce the details on the datasets, pre-trained models, and the training hyperparameters used for all our experiments in Section 4.1. Then we demonstrate the results of applying PromptCoT to text-to-image and text-to-video pre-trained generative models in Section 4.2 and Section 4.3 respectively.

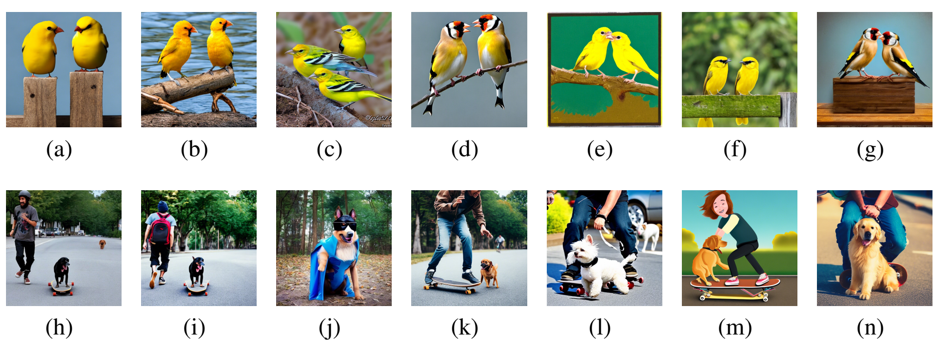

Figure 6: Generated images from prompts refined by different aligners. (a) and (h) show the images generated with the original text prompts. (b-g) and (i-n) denote the images generated with text prompts refine by 't-continue', 't2t-blip','t2t-inter','davinci','CoT_d', and 'CoT' respectively.

4.1 Setup

Dataset. For training, we build Text-revision and Text-continuation dataset from LAION-aes6plus [35], and Text-CoT dataset with the help of GPT-3.5-turbo. LAION-aes6plus is the subset of LAION, containing 12M image-text pairs with predicted aesthetics scores of 6 or higher. As a supplement, we also train with Text-revision, Text-continuation, and Text-CoT datasets from the WebVid-10M dataset [3] for video generation. For evaluation, we conduct experiments on COCO [18] validation set and MSR-VTT [45] for FID, FVD, aesthetic score, CLIP score, and PickScore.

Models. The pre-trained LLaMA-7B is used as the base model and we employ the adapter design outlined in [48] to facilitate multi-task adaptation. Two versions of Stable Diffusion [31], v1.4 and v2.1, are used for image generation. MagicVideo [50] is used for video generation.

Implementation Details. We finetune the LLaMA following alpaca's [39] strategy and instruction pattern, which has been verified powerful for text generation tasks. We validate the viability of our two initial ideas by finetuning three task-specific LLaMA for prompt refining works shown in experiments 2. One is trained on the self-constructed text-continuation dataset while the other two are trained on two types of text-revision dataset. While combining such basic methods by CoT, we include a dataset from alpaca, a subset of the text-continuation dataset, and the text-revision dataset with higher text similarity and the CoT dataset as a whole. We evaluate our alignment work on three diffusion models and on different parameters. Furthermore, we evaluate the portability of promptCoT through an adapter by comparing its performance with the fully-finetuned model.

Table 2: **Text-to-image generation performance.** We evaluate the generation performance on Stable Diffusion v1.4 and v2.1 on key metrics including aesthetic score, FID, IS, CLIP score and PickScore.

Generation Model	Booster	Aesthetic Score	FID	IS	CLIP Score	PickScore (avg/recall)
SD v1.4 ddim step=50 scale=7.0	baseline	5.40	59.15	39.13 ± 0.84	0.268	27.3%/35.7%
	t-continue	5.54	44.66	35.81 ± 0.96	0.290	39.5%/61.5%
	t2t-blip	5.62	40.77	38.56 ± 0.77	0.293	51.4%/77.5%
	t2t-inter	5.44	55.76	41.00 ± 1.17	0.271	34.3%/49.0%
	cot_d	5.64	49.58	37.43 ± 0.94	0.289	40.6%/62.2%
SD v2.1 ddim step=50 scale=7.0	baseline	5.60	58.02	37.51 ± 1.00	0.266	29.4%/41.7%
	t-continue	5.70	45.62	34.44 ± 0.71	0.287	44.3%/69.9%
	t2t-blip	5.79	40.59	37.38 ± 1.08	0.292	56.3%/82.5%
	t2t-inter	5.64	54.93	38.60 ± 0.85	0.269	37.1%/55.6%
	cot_d	5.78	50.41	34.88 ± 0.95	0.290	42.9%/66.2%
SD v2.1 ddim step=250 scale=12.0	baseline	5.60	58.17	36.37 ± 0.81	0.267	-
	t-continue	5.64	46.59	33.29 ± 0.68	0.287	-
	t2t-blip	5.76	40.89	36.16 ± 0.84	0.292	-
	t2t-inter	5.64	55.37	38.10 ± 1.16	0.269	-
	cot_d	5.75	50.41	34.88 ± 0.94	0.290	-

Table 3: **Text-to-image generation performance with adapters.** We fine-tune adapters by 5 epochs and compare them with fully fine-tuned Alpaca. Model with adapters achieves comparable results.

Model	Booster	Aesthetic Score	FID	IS	CLIP Score	PickScore
Alpaca epochs = 3	t-continue	5.70	45.62	34.44 ± 0.71	0.287	44.3%/69.9%
	t2t-blip	5.79	40.59	37.38 ± 1.08	0.292	56.3%/82.5%
	t2t-inter	5.64	54.93	38.60 ± 0.852	0.269	37.1%/55.6%
	cot_d	5.78	50.41	34.88 ± 0.95	0.290	42.9%/66.2%
Adapter epochs = 5	t-continue	5.69	48.00	35.8 ± 0.57	0.283	-
	t2t-blip	5.70	46.86	38.0 ± 0.66	0.289	-
	t2t-inter	5.64	56.28	39.0 ± 0.64	0.269	-
	cot_d	5.85	51.06	31.8 ± 0.65	0.251	-

4.2 Text-to-image Evaluation

The COCO [18] validation set is the standard benchmark for evaluating text-to-image models. The key automated performance metrics used are FID to measure image fidelity, CLIP score, PickScore to measure image-text alignment, aesthetic score [22] to predict the aesthetic quality, and Inception Score (IS) to evaluate the diversity. We utilize two versions of Stable Diffusion for image generation with prompts from COCO and our PromptCoT. Table 2 presents the evaluation results for each metric with different single-function boosters including t-continue, t2t-blip, and t2t-inter, as well as a baseline. The results show that incorporating the alignment method proposed in our paper consistently improved the generated image quality across all metrics compared to the baseline. Among the single-function boosters, the t2t-blip booster demonstrates the best performance, as it is able to achieve alignment to a greater extent. For example, it transfers "Boxes of fruit displayed at an open-air market" to "A view of stalls selling fruit at the Harare International Market in Harare, Zimbabwe" by rephrasing

Table 4: **Text-to-image generation performance.** We compare finetuned CoT aligner and davinci-003 model from OpenAI. All metrics are evaluated on a subset of the COCO validation dataset which contains 1k images.

Booster	Aesthetic Score	CLIP Score	PickScore
baseline	5.62	0.231	16.8%/26.1%
tcontinue	5.72	0.285	37.8%/66.2%
t2t_blip	5.80	0.293	50.6%/81.5%
t2t_inter	5.66	0.269	30.7%/52.5%
cot_d	5.79	0.291	34.9%/59.5%
cot	5.80	0.293	36.4%/59.0%
davinci	5.69	0.277	26.0%/47.5%

the expression and adding reasonable details. In contrast, the t2t-inter booster, which has a similar function to t2t-blip, shows inferior performance, although it still outperforms the baseline. This could be due to the CLIP-interrogator used to create the text-revision dataset introducing irrelevant entities. Furthermore, we test with different factors of classifier-free guidance to prove the generality of our PromptCoT. Varying the scale of classifier-free guidance results in consistent performance.

4.3 Text-to-video Evaluation

In addition, we experiment with the text-to-video evaluation task to demonstrate the effectiveness of our approach. We employ two single-function boosters, t-continue, and t2t-blip on the WebVid-10M dataset [3]. For t2t-blip, we uniformly sample the video and randomly select five frames, which serve as input for the blip model and be used to generate the revision result. Then, we finetune the LLaMA model following alpaca's [39] strategy and build prompts from MSR-VTT with the fine-tuned model. We use MagicVideo [50] as the base model to test the effectiveness of our prompts. The results are shown in Table 5. The results indicate that the boosters are effective in enhancing the quality of the generated videos compared to the baseline, at least they "do no harm". Among the boosters, the booster better aligns the prompts and achieves the best performance overall. For cot_d, we generate 21k data with the help of GPT-3.5-turbo. Similar to text, we utilize a chain of five questions to generate the expected production, but with subtle differences to encourage GPT-3.5-turbo to generate more video-related features, e.g., movement. Similar to text generation, we adopt a chain of five questions to generate the expected production for video prompts. However, there are subtle differences in the question prompts to encourage GPT-3.5-turbo to incorporate more video-related features, such as movement, into its generated content. For example, "a large passenger jet flying in the sky at sunset" can be refined to "Boeing 747 flying across a vibrant sunset backdrop in a captivating, cinematic 4K video. Slowly gaining altitude with wings tilting slightly, this footage captures the plane's majesty". The scores of cot_d will be included in the supplementary material.

Table 5: **Text-to-video generation performance.** We evaluate the generation performance on MagicVideo on key metrics including FID, FVD, and CLIP score.

Model	Dataset	Booster	FID	FVD	CLIP Score
MagicVideo	MSR-VTT	baseline	36.5	998	0.284
		t-continue	33.2	951	0.296

5 Conclusion

In this paper, we present PromptCoT, an innovative system designed to autonomously enhance the quality of prompts used in diffusion-based generative models, which are critical for high-fidelity visual content generation. PromptCoT leverages pre-trained Large Language Models (LLMs) and a unique Chain-of-Thought (CoT) mechanism to refine prompts, thereby improving the alignment between the original and refined prompts. To balance computational efficiency, we employ adapters to allow for efficient adaptation to new datasets or models. Our evaluations demonstrate that PromptCoT can achieve superior performance compared to the baselines.

References

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning, 2020.

[2] Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch, 8 2021.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. 2020.

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] Ning Ding, Yujia Qin, Guang Yang, Fu Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Haitao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juan Li, and Maosong Sun. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *ArXiv*, abs/2203.06904, 2022.

[9] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.

[10] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723, 2021.

[11] Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. Ptr: Prompt tuning with rules for text classification. *ArXiv*, abs/2105.11259, 2021.

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

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[16] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, 2017.

[17] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021.

[18] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[19] Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*, 2017.

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35, 2021.

[21] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers, 2021.

[22] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012.

[23] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions, 2020.

[24] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers, 2020.

[25] Guanghui Qin and Jas' Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *ArXiv*, abs/2104.06599, 2021.

[26] 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. 2021.

[27] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[29] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Dall·e 2: Exploring cross-modal transformers for image generation. *OpenAI Blog*, 2021.

[30] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

[33] Timo Schick, Helmut Schmid, and Hinrich Schütze. Automatically identifying words that can serve as labels for few-shot text classification. In *International Conference on Computational Linguistics*, 2020.

[34] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2020.

[35] 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. *arXiv preprint arXiv:2210.08402*, 2022.

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[37] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[39] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[40] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[41] Leandro von Werra, Alex Havrilla, Max reciprocated, Jonathan Tow, Aman cat state, Duy V. Phung, Louis Castricato, Shahbuland Matiana, Alan, Ayush Thakur, Alexey Bukhtiyarov, aaronrmm, Fabrizio Milo, Daniel, Daniel King, Dong Shin, Ethan Kim, Justin Wei, Manuel Romero, Nicky Pochinkov, Omar Sanseviero, Reshinth Adithyan, Sherman Siu, Thomas Simonini, Vladimir Blagojevic, Xu Song, Zack Witten, alexandremuzio, and crumb. CarperAI/trlx: v0.6.0: LLaMa (Alpaca), Benchmark Util, T5 ILQL, Tests, March 2023.

[42] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Huai hsin Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171, 2022.

[43] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021.

[44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[45] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[46] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022.

[47] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[48] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[49] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[50] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

[51] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023.

[52] Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed Huai hsin Chi. Least-to-most prompting enables complex reasoning in large language models. *ArXiv*, abs/2205.10625, 2022.