PromptCoT: Align Prompt Distribution via Adapted Chain-of-Thought

Junyi Yao[*1], Yijiang Liu[*2], Zhen Dong[3], Mingfei Guo[4], Helan Hu[1],
Kurt Keutzer[3], Li Du[2†], Daquan Zhou[5✉], Shanghang Zhang[1✉]
[1]Peking University, [2]Nanjing University
[3]University of California, Berkeley, [4]Stanford University, [5]Bytedance

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 necessitates empirical expertise and poses challenges for inexperienced users. In this paper, we introduce PromptCoT, an innovative enhancer that autonomously refines prompts for users. PromptCoT is designed based on the observation that prompts, which resemble the textual information of high-quality images during training, lead to superior generation performance. Therefore, we fine-tune the Large Language Models (LLM) using a curated text dataset that comprises descriptions of high-quality visual content. Consequently, the LLM can capture the distribution of high-quality texts, enabling it to boost the original texts. Nonetheless, one drawback of LLMs is their tendency to generate irrelevant information. We employ a tailored Chain-of-Thought (CoT) mechanism to address the problem. Our CoT can extract and amalgamate crucial information from the prompt candidates, enabling a reasonable process based on the contextual cues to produce a more comprehensive and nuanced output. Considering computational efficiency, instead of allocating a dedicated LLM to each individual model or dataset, we integrate adapters that facilitate task-specific adaptation, leveraging a shared LLM as the foundation for this process. With independent fine-tuning of adapters, we can adapt PromptCoT to new datasets while minimally increasing training costs and memory usage. We evaluate the effectiveness of PromptCoT by assessing on widely-used latent diffusion models for visual generation. The results demonstrate significant improvements in key performance metrics.

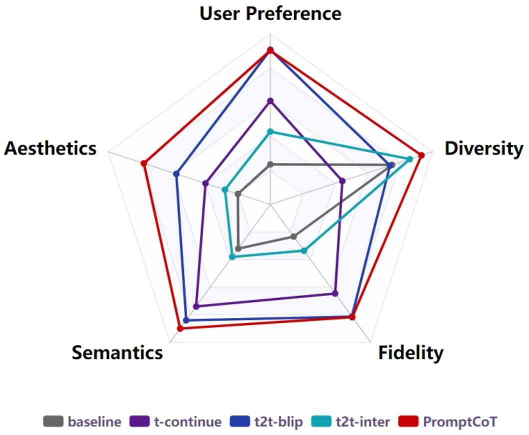


Figure 1. **Multidimensional performance across key metrics.** Our PromptCoT outperforms the baseline and various aligners we proposed in several key areas, including user preference (PickScore), aesthetics (Aesthetic Score), semantics (CLIP Score), fidelity (FID), and diversity (IS). Values are sourced from Table 2, and have been normalized to 1.0 for clarity.

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 contents, such as images and videos, as evidenced by notable contributions in [7, 13, 32, 33, 35, 40–42]. By utilizing textual inputs as conditional guidance, diffusion models have the ability to generate visual outputs that align with the corresponding input text, employing an iterative denoising procedure. This technological advancement has paved the way for revolutionary applications, including notable examples such as DALL-E 2 [32], Stable Diffusion [33], MagicVideo [57], and more.

[*] Equal contribution.

[✉] Corresponding Author.

[†] Author with the School of Electronic Science and Engineering, Nanjing University, and the Interdisciplinary Research Center for Future Intelligent Chips, Nanjing University, Suzhou.

Nevertheless, the quality of the generated content is intricately linked to the quality of the textual prompts given to the generative model. Human inputs often tend to be informal and straightforward, which can limit the expression of the desired scene. Additionally, the text encoder within the generative model may not fully grasp 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 obtained through web-scraping techniques [39]. Our observations reveal that the distribution of the text data might not align with the linguistic style employed by layman users. Furthermore, even in cases where the training text data matches the desired style, the quality can vary significantly due to the presence of meaningless words or extraneous information within the text data, which complicates the establishment of an accurate mapping between the text and the corresponding image.

Consequently, there is an urgent need to develop a methodology that can effectively align prompts, thereby enhancing the image generation performance of generative models. While data cleaning and model fine-tuning have been considered as potential solutions, they often come with drawbacks such as high costs, instability, and time intensiveness. An alternative approach is manual prompt engineering, involving the refinement of prompts to optimize generation performance. However, this empirical task typically requires the expertise of experienced professionals, posing a significant challenge for individuals lacking relevant skills.

In our study, we have noticed a significant trend: prompts that resemble those in the training set often result in superior generative performance. Based on this observation, we introduce PromptCoT, a novel prompt booster that harnesses the capabilities 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 [44], a widely-used pre-trained Large Language Model, on two distinct datasets we've prepared. We used a text-continuation dataset that appends aligned details to original prompts. Additionally, a text-revision dataset was employed to rewrite original prompts to aligned prompts. This process enabled 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 have constructed 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 abilities and augment their capacity to generate high-quality text that is both contextually relevant and logically coherent. To accommodate the diverse training sets of different generative models, we have incorporated a parameter-efficient adaptation design into the training pipeline of PromptCoT. This design augments a pre-trained base booster with specific lightweight adapters that are capable of aligning text distributions for various generative models across multiple tasks. The effectiveness of PromptCoT is demonstrated through extensive experiments conducted on widely-used latent diffusion models for image and video generation. These experiments have shown significant improvements in key performance metrics just as shown in Figure 1.

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, which is also a new optimization scheme to improve prompt quality by leveraging the power of pre-trained LLMs and CoT mechanisms.

• We meticulously craft an automatic data mining pipeline that extracts a compact and premium subset from the targeted T2I model. Furthermore, we construct datasets while proposing a 5-step dataset optimization pipeline, inspired by CoT, which can be applied across a diverse range of T2I models.

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

• We illustrate the effectiveness of PromptCoT through extensive experiments on widely-used latent diffusion models for image and video generation, showing significant improvements in various 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 [31]. It utilize CLIP [30] 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 [35] and Parti [51] 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. Recently, [21, 55] have leveraged Large Language Models (LLMs) to provide auxiliary support, resulting in enhanced performance. The application of diffusion probabilistic models has also been extended to the domain of video generation. The Video Diffusion Model [14], 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 [58] 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 [4, 29] series models developed by OpenAI, which are based on the decoder component of the transformer architecture. Another LLM is Meta's OPT [54], 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 [44, 45], StableLM [2], as well as the instruction-following Alpaca model [43] are smaller and more performant, achieve comparable results to ChatGPT with far fewer parameters (7B). For specific tasks like conversational applications, ChatGLM [9, 52] 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 [16] 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 [15, 27] utilizes dynamic pre-trained adapters for different tasks and languages to reduce adaptation time. Compacter [25] 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 [5, 8, 24, 38]. Discrete text prompts [17] serve as starting points for the model's language generation, and are used to generate responses in dialogue systems. Beyond discrete prompts, [19, 48] explores prompt tuning to learn soft prompts to perform specific downstream tasks, which provide more context-aware guidance to the model. [28] 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 [37, 38]. [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 [49]. The traditional prompting method [5] performs poorly when it comes to tasks that require reasoning abilities. Inspired by the concept of using intermediate steps to solve reasoning problems [6, 23], 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 [26]. Additionally, [56] 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, [46] 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

In this section, we provide an overview in 3.1, discuss how to align prompt distribution with LLM in 3.2, introduce the enhancement with Chain-of-Thought in 3.3, and adopt multi-task adaptation in 3.4 We approximately illustrate our method for enhancing prompts quality in a logical sequence, which leads to the usage of LLM, Chain-of-thought, and multi-task adapor.

3.1. Overview

Text-to-image diffusion models act 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 five sets of 50 prompts corresponding to images with

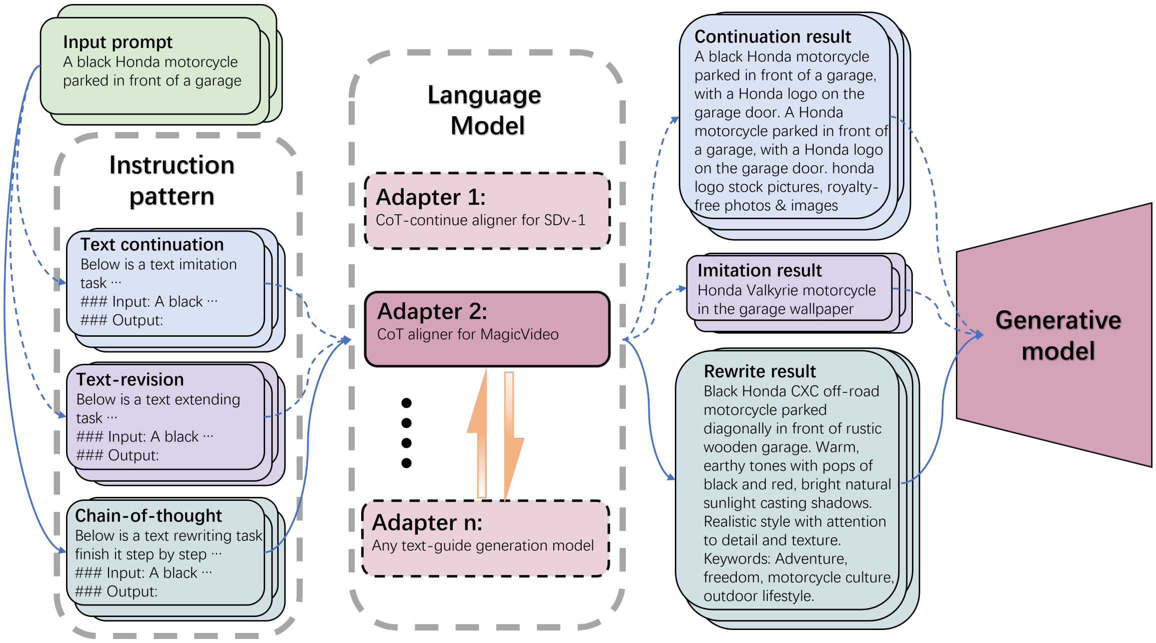

Input prompt
A black Honda motorcycle parked in front of a garage
Instruction pattern
Text continuation
Below is a text imitation task ···
### Input: A black ···
### Output:
Text-revision
Below is a text extending task ···
### Input: A black ···
### Output:
Chain-of-thought
Below is a text rewriting task finish it step by step ···
### Input: A black ···
### Output:
Language Model
Adapter 1:
CoT-continue aligner for SDv-1
Adapter 2:
CoT aligner for MagicVideo
Adapter n:
Any text-guide generation model
Continuation result
A black Honda motorcycle parked in front of a garage, with a Honda logo on the garage door. A Honda motorcycle parked in front of a garage, with a Honda logo on the garage door. honda logo stock pictures, royalty-free photos & images
Imitation result
Honda Valkyrie motorcycle in the garage wallpaper
Rewrite result
Black Honda CXC off-road motorcycle parked diagonally in front of rustic wooden garage. Warm, earthy tones with pops of black and red, bright natural sunlight casting shadows. Realistic style with attention to detail and texture. Keywords: Adventure, freedom, motorcycle culture, outdoor lifestyle.
Generative model

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.

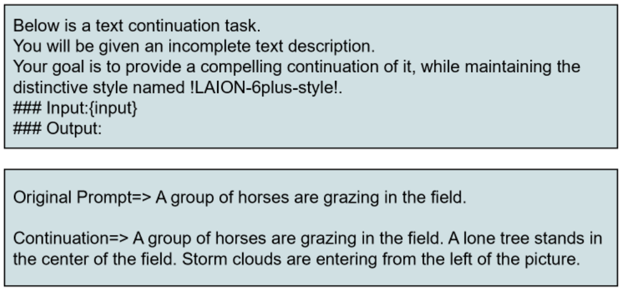
varying levels of quality from the Stable Diffusion training set, LAION [39], for image generation. The aesthetic score, an image quality metric introduced by [34], is used to represent the quality of individual images. As shown in Table 1, the generation performance is highly related to the prompts 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, which is a prompt refiner that can align input prompts with high-quality prompts in the training set, and in turn, improve generation performance.

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

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

3.2. Aligning prompt distribution with LLM

LLMs are 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 LLaMA [44], a popular LLM, 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.



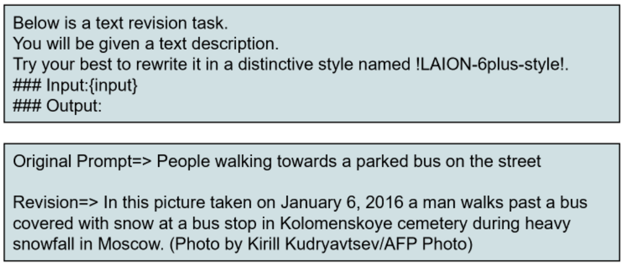
Below is a text continuation task.
You will be given an incomplete text description.
Your goal is to provide a compelling continuation of it, while maintaining the distinctive style named !LAION-6plus-style!.
Input:{input}
Output:

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.



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).

3.3. Enhancement with Chain-of-Thought

While instruction tuning enables the LLM to add details and align text distribution, however, it tends to generate extraneous information that degrades performance. To address this, we introduce the Chain-of-Thought (CoT) mechanism in the pipeline to address this issue. We establish a five-step process to guide the LLM in producing: (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.

3.5. Dataset preparation

We construct three types of datasets: text-continuation, text-revision, and text-CoT shown in Figure 5.

Text-continuation Dataset. We selectively filter prompts from the training data of existing generative models. The selection criteria include high CLIP similarity and appropriate length. Specifically for the LAION dataset, aesthetic scores are also considered, aiming to enhance the quality of selected prompts. Once identified as high-quality, these prompts undergo truncation; a part of the text is removed, retaining only the initial segment as input data. Subsequently, the Language Learning Model (LLM) is trained with these truncated prompts, focusing on generating the omitted sections to complete the text. This methodology equips the LLM to adeptly continue text prompts, maintaining the style and context inherent in the original text.

Text-revision dataset. This dataset encompasses pairs of human-like texts and corresponding high-quality prompts, the latter of which are detailed in the text-continuation dataset section. To automatically generate human-like prompts, we employ BLIP[20] and CLIP-interrogator for image captioning, thereby ensuring a natural and contextually relevant textual output. The semantic closeness between these human-like prompts and the high-quality prompts is quantitatively assessed using the text encoder feature of CLIP. We set a threshold score greater than 0.4 to ensure substantial semantic relevance. This score is chosen based on empirical evaluations indicating a strong correlation between the prompts at this threshold, thus maintaining consistency and relevance in the dataset.

Text-CoT Dataset. We utilized GPT-3.5-Turbo to curate a specialized collection specifically designed for a targeted task. Our methodology entailed the establishment of a sequential interaction protocol with GPT-3.5-Turbo. The primary objective of this protocol was to meticulously extract pertinent content from the initial prompts. This was aimed at directing the aligner to yield outputs that are not only more consistent but also highly relevant, while explicitly excluding any generation of extraneous content. Consequently, we successfully compiled a dataset comprising 52,000 sample pairs, constituting our Chain-of-Thought dataset.

4. Experimental results

In this section, we initially present detailed information regarding the datasets, pre-trained models, and the training hyperparameters employed across all our experiments (refer to Section 4.1). Subsequently, we illustrate the outcomes of applying PromptCoT to pre-trained text-to-image and text-to-video generative models in Sections 4.2 and 4.5, respectively.

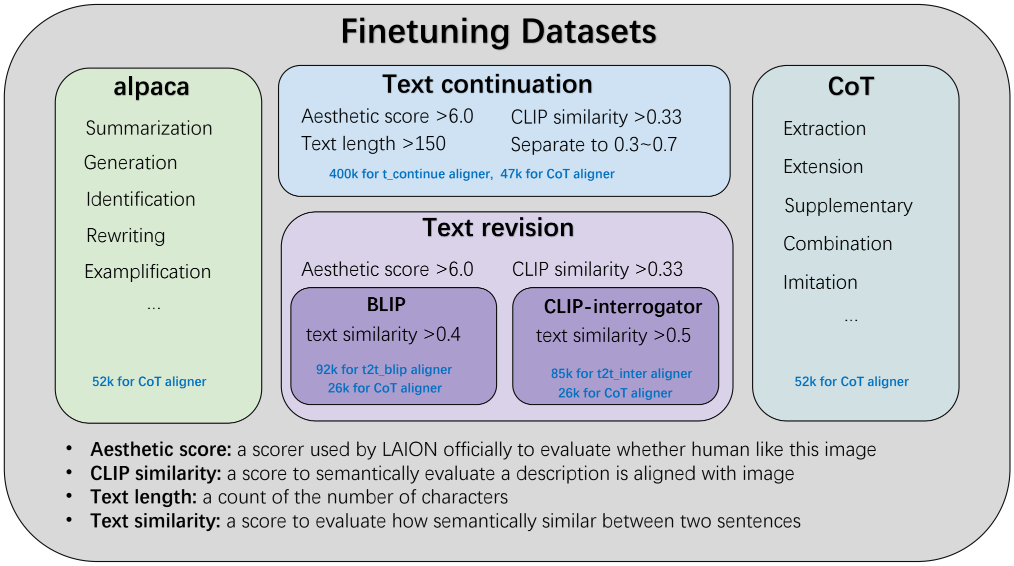


Figure 5. Composition of fine-tuning tasks including text-continuation, text-revision, text-CoT, and self-instruction of Alpaca.

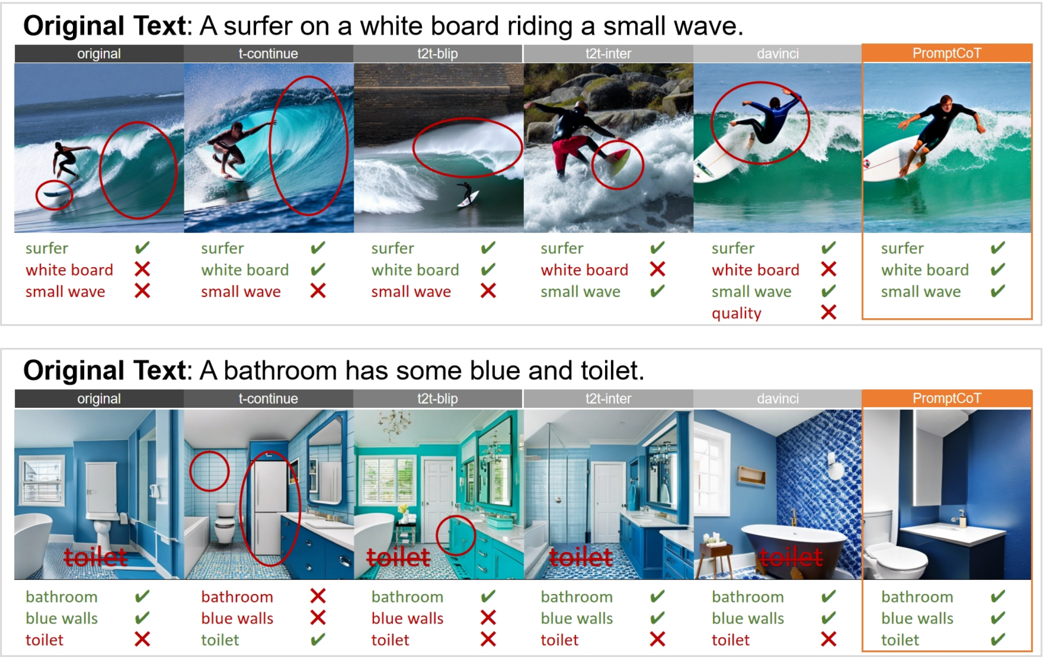


Figure 6. Comparison of image and text matching among prompts refined by various aligners, including t-continue, t2t-blip, t2t-inter, Davinci, and PromptCoT. Areas of mismatch are highlighted and PromptCoT demonstrates superior performance in semantic consistency.

4.1. Setup

Dataset. For training aligners, we construct Text-revision and Text-continuation datasets from LAION-aes6plus [39], and a Text-CoT dataset with the assistance of ChatGPT. LAION-aes6plus is the high-quality subset of LAION, containing 12 million image-text pairs with aesthetics scores higher than 6. As a supplement, we also construct the corresponding training datasets for the WebVid-10M dataset [3] for video generation. For evaluation, we conduct experiments on COCO [22], LAION-aes6plus and MSR-VTT [50], assessing FID, FVD, inception score, aesthetic score, CLIP score, and PickScore.

Models. The pre-trained LLaMA-7B model is used as the base model, and we employ the adapter design outlined in [53] to enable multi-task adaptation. We utilize two versions of Stable Diffusion [34], v1.4 and v2.1, for image generation, and MagicVideo [57] for video generation.

Implementation Details. All experiments are based on pre-trained LLaMA-7B [44], an open-sourced Large Language Model with seven billion parameters. The fine-tuning process of each aligner follows [43, 47] using 8×A100-80GB GPUs, which takes three hours until converge. We validate the feasibility of our two initial ideas by fine-tuning three task-specific LLaMA for prompt refining works, which includes t-continue trained on the Text-continuation dataset, t2t-blip and t2t-inter trained on two variants of text-revision datasets named BLIP and CLIP-interrogator respectively. PromptCoT is trained on the Text-CoT datasets. We evaluate the image generation performance on two variants of text-to-image models and one text-to-video model. Furthermore, we evaluate the portability of PromptCoT through an adapter by comparing its performance with that of the fully fine-tuned model.

Evaluation Metrics. We evaluate the generation performance with Fréchet Inception Distance (FID) [12], Inception Score (IS) [36], CLIP score [30], Aesthetic Score [34] and PickScore [18]. The definitions of FID, IS, and CLIP score are strictly following previous works[12, 18, 30, 34, 36]. More detailed explanations of Aesthetic Score and PickScore can be found in the supplementary material.

4.2. Text-to-image evaluation

We generate images using two versions of Stable Diffusion, feeding prompts respectively from the captions of COCO and captions refined by PromptCoT. Table 2 presents the evaluation results of different aligners including t-continue, t2t-blip, t2t-inter and PromptCoT, as well as a baseline which uses original captions of COCO. All of our approaches outperform the baseline across all metrics, with PromptCoT showing the most impressive results. Specifically, with Stable Diffusion v2.1, our PromptCoT exceeds the baseline by 0.29 in aesthetic score, reflecting a 5.2% enhancement in the aesthetic composition of images. Concurrently, it increases the CLIP score↑ by 0.03 (an 11% improvement), indicating enhanced consistency between text and image; and reduces the FID↓ by 18 (a 30% reduction), signifying an improvement in image quality. From practical examples of refined prompts, we observe that our approaches tend to add more relevant and detailed descriptions to the original prompts. For instance, the prompt 'Boxes of fruit displayed at an open-air market' is rephrased to 'A view of stalls selling fruit at the Harare International Market in Harare, Zimbabwe', incorporating additional details such as the characteristics of the fruit, and specific information about the market's name and location.

4.3. Visualization

To better understand the advantages of our PromptCoT, we provide a detailed visual comparison of images generated using the original prompt and those refined with different aligners (tcontinue, t2t_blip, t2t_inter, cot_davinci, cot_d, and PromptCoT) in Figure 6. We have highlighted inconsistencies between the prompt and the images within the figures, accompanied by annotations below each image. It is noteworthy that not only do the images generated using PromptCoT exhibit superior quality, but they also display a better alignment with the textual contents. For instance, in the top-row images generated from the prompt "A surfer on a whiteboard riding a small wave", PromptCoT stands out by effectively capturing all the desired elements, while others may struggle to interpret the prompt accurately with all key concepts.

4.4. Comparing to ChatGPT and human beings

We conducted experiments comparing PromptCoT with ChatGPT as well as human beings. To compare the capability of refining prompts between PromptCoT and human beings, we first collect a set of text prompts from the captions of COCO dataset. We then invited a group of 30 research volunteers to refine the collected prompts to improve the image generation quality. The volunteers are all specialized in deep learning algorithms and are thus expected to perform well on this task. The findings are succinctly presented in Table 4. Upon careful examination, it is evident that humans possess the ability to modify prompts to achieve better content alignment between the text descriptions and the generated images, resulting in an improved CLIP score. We employ davinci-003 to refine prompts for image generation, which produces better results than human's in aesthetic score and clip score, but lower PickScore. Conversely, PromptCoT demonstrates its capability to generate prompts that enhance the aesthetic score, CLIP score and PickScore, surpassing ChatGPT and human performance by a significantly larger margin.

4.5. Text-to-video evaluation

In addition, we experiment with the text-to-video task to demonstrate the effectiveness of our approach. We employ PromptCoT on the WebVid-10M dataset [3]. Then, we finetune the LLaMA model following alpaca's [43] strategy and refine prompts from MSR-VTT with the fine-tuned model. We use MagicVideo [57] as the base model to test the effectiveness of our prompts. The results are shown in Table 5. The results suggest that PromptCoT effectively enhances the quality of the generated videos, surpassing the baseline. Among the boosters, the PromptCoT better aligns

Model	Booster	Aesthetic↑	FID↓	IS↑	CLIP↑	PickScore↑
	baseline	5.40	59.15	39.13 ± 0.84	0.268	27.3%/35.7%
SD v1.4	t-continue	5.54	44.66	35.81 ± 0.96	0.290	39.5%/61.5%
ddim step=50	t2t-blip	5.62	40.77	38.56 ± 0.77	0.293	51.4%/77.5%
scale=7.0	t2t-inter	5.44	55.76	**41.00 ± 1.17**	0.271	34.3%/49.0%
	PromptCoT	**5.79**	**41.10**	39.66 ± 0.88	**0.294**	**53.2%/79.1%**
	baseline	5.60	58.02	37.51 ± 1.00	0.266	29.4%/41.7%
SD v2.1	t-continue	5.70	45.62	34.44 ± 0.71	0.287	44.3%/69.9%
ddim step=50	t2t-blip	5.79	40.59	37.38 ± 1.08	0.292	56.3%/82.5%
scale=7.0	t2t-inter	5.64	54.93	38.60 ± 0.85	0.269	37.1%/55.6%
	PromptCoT	**5.89**	**40.47**	**39.32 ± 0.91**	**0.295**	**56.1%/83.7%**

Table 2. **Text-to-image generation performance.** We conducted performance evaluations of Stable Diffusion versions 1.4 and 2.1 on key metrics including aesthetic score, FID, IS, CLIP score and PickScore. The term 'baseline' refers to the generation outcomes using original COCO captions. Our developed 'boosters', namely t-continue, t2t-blip, t2t-inter, and PromptCoT, are innovative implementations designed to refine original COCO captions into high-quality prompts, as detailed in our implementation section. Notably, our boosters significantly surpass the baseline in performance, with PromptCoT emerging as the most outstanding in terms of effectiveness.

Model	Booster	Aesthetic↑	FID↓
Adapter	baseline	5.40	59.15
	PromptCoT	**5.85**	**51.06**

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.

Booster	Aesthetic↑	CLIP↑	PickScore↑
baseline	5.62	0.231	16.8%/26.1%
davinci	5.69	0.277	26.0%/47.5%
Human	5.62	0.270	48.1%/58.2%
PromptCoT	**5.93**	**0.293**	**57.5%/73.6%**

Table 4. **Text-to-image generation performance.** We compare PromptCoT with davinci-003 model from OpenAI and prompts from human beings. All metrics are evaluated on a subset of the COCO validation dataset which contains 1k images.

the prompts and achieves the best performance overall. For PromptCoT, 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, the question prompts have subtle differences that encourage GPT-3.5-turbo to include more video-related features, like movement, in the content it generates. 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.

Model	Dataset	Booster	FID↓	FVD↓	CLIP↑
MagicVideo	MSR-VTT	baseline	36.5	998	0.284
		PromptCoT	**33.2**	**951**	**0.296**

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

5. Conclusion

In this research, we introduce PromptCoT, a novel system designed to autonomously enhance the quality of prompts utilized in text-to-image (T2I) generative models. These prompts are essential for generating high-quality visual content. PromptCoT utilizes pre-trained Large Language Models (LLMs) combined with the Chain-of-Thought (CoT) mechanism to refine prompts, significantly improving image generation quality and semantic consistency of T2I models. To maintain computational efficiency, we incorporate adapters, enabling efficient adaptation to new datasets and T2I models. Our extensive evaluations reveal that the aligners we have proposed significantly surpass the baseline in performance. Among these, PromptCoT stands out as the most effective, consistently outperforming all other aligners in enhancing prompt quality T2I generative models.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFA0717700, in part by the Natural Science Foundation of China under Grants 62371223 and 62211530492.

References

[1] Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning, 2020. 3

[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, 2021. 3

[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. 6, 7

[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, 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. 3

[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, 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. 3

[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. 3

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[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. 3

[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. 3

[10] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *ArXiv*, abs/2012.15723, 2021. 3

[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. 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7

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

[14] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2

[15] 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 *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[16] 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. 3

[17] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, 2017. 3

[18] 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. *arXiv preprint arXiv:2305.01569*, 2023. 7

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021. 3

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 5

[21] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *ArXiv*, abs/2305.13655, 2023. 2

[22] 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. 6

[23] 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. 3

[24] 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. 3

[25] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers, 2021. 3

[26] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions, 2020. 3

[27] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers, 2020. 3

[28] Guanghui Qin and Jas' Eisner. Learning how to ask: Querying lms with mixtures of soft prompts. *ArXiv*, abs/2104.06599, 2021. 3

[29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3

[30] 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. 2, 7

[31] 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. 2

[32] 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. 1

[33] 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. 1

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 4, 7

[35] 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. 1, 2

[36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 7

[37] 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. 3

[38] 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. 3

[39] 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. 2, 4, 6

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

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

[42] 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. 1

[43] 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. 3, 7

[44] 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. 2, 3, 4, 7

[45] 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, 2023. 3

[46] 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. 3

[47] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*, 2022. 7

[48] 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. 3

[49] 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. 3

[50] 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. 6

[51] 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. 2

[52] 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. 3

[53] 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. 7

[54] 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. 3

[55] Shan Zhong, Zhongzhan Huang, Wushao Wen, Jinghui Qin, and Liang Lin. Sur-adapter: Enhancing text-to-image pre-trained diffusion models with large language models. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 2

[56] 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. 3

[57] 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. 1, 7

[58] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023. 3