Automated Black-box Prompt Engineering for Personalized Text-to-Image Generation

Yutong He[1], Alexander Robey[2], Naoki Murata[3], Yiding Jiang[1], Joshua Williams[1], George J. Pappas[2], Hamed Hassani[2], Yuki Mitsufuji[3,4], Ruslan Salakhutdinov[1], and J. Zico Kolter[1,5]

Carnegie Mellon University[1], University of Pennsylvania [2], Sony AI[3], Sony Group Corporation[4], Bosch Center for AI[5],

Abstract. Prompt engineering is effective for controlling the output of text-to-image (T2I) generative models, but it is also laborious due to the need for manually crafted prompts. This challenge has spurred the development of algorithms for automated prompt generation. However, these methods often struggle with transferability across T2I models, require white-box access to the underlying model, and produce non-intuitive prompts. In this work, we introduce PRISM, an algorithm that automatically identifies human-interpretable and transferable prompts that can effectively generate desired concepts given only black-box access to T2I models. Inspired by large language model (LLM) jailbreaking, PRISM leverages the in-context learning ability of LLMs to iteratively refine the candidate prompts distribution for given reference images. Our experiments demonstrate the versatility and effectiveness of PRISM in generating accurate prompts for objects, styles, and images across multiple T2I models, including Stable Diffusion, DALL-E, and Midjourney.

Keywords: Text-to-Image Generation · Prompt Engineering · Personalized Text-to-Image Generation

1 Introduction

An important goal of generative modeling is to design algorithms capable of steering generative models to produce desired output images. Early attempts, which often centered on particular architectures or tasks, were largely characterized by manually-curated data collection, fine-tuning, or retraining from scratch [10,19,33,47]. These requirements are often costly, and the resulting solutions usually do not transfer well between models. Thus despite the promise of these methods, efficient and generalized algorithms for controllable generation remain sought after.

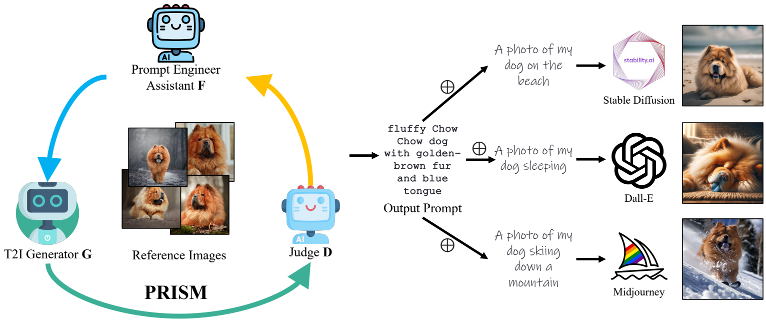

Prompt Engineer Assistant **F**
T2I Generator **G**
Reference Images
Judge **D**
PRISM
fluffy Chow Chow dog with golden-brown fur and blue tongue
Output Prompt
⊕
A photo of my dog on the beach
Stable Diffusion
A photo of my dog sleeping
Dall-E
A photo of my dog skiing down a mountain
Midjourney

Fig. 1: Given a set of reference images, our method, PRISM, is capable of creating human-interpretable and accurate prompts for the desired concept that are also transferable to both open-sourced and closed-sourced text-to-image models.

Today, perhaps the most popular approach for controllable generation is to guide the generation process with a piece of textual information, or prompt, that describes the properties of the desired output using text-to-image (T2I) generative models [24, 42]. Through text, T2I models allow users to quickly and easily describe a wide variety of concepts, and model designers can more efficiently explore the behavior of their model through a myriad of strategies [3, 39]. The predominant method for obtaining such input text is to manually design candidate prompts in an iterative, trial-and-error fashion, a process known as *prompt engineering*, based on what the user (prompt engineer) *believes* will lead to a desirable output. Unfortunately, these practices are often sensitive to different phrasings [36], require expert domain knowledge, and are notably inefficient as they necessitate a human in the loop.

Motivated by the drawbacks of manual prompt engineering, a recent line of work known as *personalized* or *subject-driven* T2I generation has sought to automate the controllable generation pipeline. Given a collection of reference images that capture specific concepts, such as artistic style or shared objects, personalized T2I algorithms are designed to produce images that reflect those concepts illustrated in the reference images. While personalized T2I methods often involve fine-tuning or retraining the underlying T2I model [4, 26, 29], several approaches focus specifically on automating prompt engineering to generate effective prompts. Unfortunately, existing algorithms in this spirit tend to require pre-collected, architecture-specific keywords[1] or white-box, embedding-based optimiza-

[1] https://github.com/pharmapsychotic/clip-interrogator

tion [6, 16], leading to non-interpretable prompts [39] and precluding the possibility of directly generating prompts for closed-source T2I models (e.g., Midjourney or DALL-E).

In order to address these shortcomings, we propose ***P**rompt **R**efinement and **I**terative **S**ampling **M**echanism* (PRISM), a new automated prompt engineering algorithm for personalized T2I generation. A key observation is that prompt engineers repeat the process of updating their "belief" of what makes an effective prompt based on the difference between their desired results and the generated images from previous iterations. Inspired by jailbreaking attacks on large language models (LLMs) [3], we design an algorithm that operates with only limited human input, is capable of generating human interpretable and editable prompts, makes minimal assumptions about the underlying T2I generative model, and generalizes across different T2I models, including popular black-box models such as DALL-E and Midjourney.

Given a set of reference images, our method first generates an initial prompt and its corresponding image using a multimodal LLM and a T2I generative model. We then obtain a score indicating the visual similarity of the generated image and the reference image with respect to the targeting concept via another multimodal LLM. Leveraging LLMs' in-context learning abilities [30, 40, 46], we instruct the LLM to update the candidate prompt distribution based on the previously generated prompt, images, and the evaluation scores. This processing is then repeated for a predetermined number of iterations. In the end, PRISM outputs the best-performing prompt by re-evaluating the top prompts generated from this process.

Experimentally, our results indicate that PRISM consistently outperforms existing methods, including Textual Inversion [6], PEZ [39], BLIP2 [12] and CLIP-Interrogator[1], with respect to human-interpretability while maintaining high visual accuracy. Our method also shows significantly better generalizability and transferability as we achieve the best performance in almost all metrics when experimenting with closed-source models in comparison to baselines. Finally, we also show that because of the interpretability provided by our method, the prompts produced by PRISM are also easily editable, enabling a wide range of creativity possibilities in real life.

2 Related Works

Controllable T2I generation. Several methods tackle conditional image generation in a training-free manner by using pretrained diffusion models as priors for the data distribution [5, 9, 18, 31, 32], and analogous approaches exist for T2I diffusion models (e.g., StableDiffusion) [8,25,43]. In general, these methods assume that the controllability objective can be formulated as differentiable loss functions, although they require access to model parameters and involve complex hyperparameter tuning. Another class of approaches such as ControlNet [44], IP-Adapter [41], Dreambooth [26], SuTI [4] and InstantBooth [29] also improve the controllability of pretrained T2I models, but they require expensive fine-tuning or re-training of the underlying model. Prompt tuning methods such as Textual Inversion [6], PEZ [39], and PH2P [16,39] are in the same spirit as this paper, as they do not require fine-tuning or optimizing the underlying model and generate images that inherit the properties of a given reference image. However, unlike PRISM, each of these methods requires access to the underlying model parameters and produces non-interpretable prompts.

Prompt engineering. Manual prompt engineering is one of the most popular approaches to eliciting desired behaviors from large pre-trained models because it uses little or no data and does not require fine-tuning [2, 22]. However, major drawbacks of manual prompt engineering include its laborious nature, its reliance on domain expertise, and the fact that its performance can be highly sensitive to how the prompts are phrased [15, 36]. To address this issue, several methods have been proposed to construct the prompts in an automated manner [7,17,30,40,45,46]. In particular, the field of LLM jailbreaking is concerned with automatically designing prompts that can elicit specific content (which is often objectionable or illicit) from a targeted LLM [14,23,37,48]. A particularly relevant work is [3], which uses an auxiliary LLM to iteratively construct jailbreak prompts that elicit harmful behaviors from a targeted LLM. Our method builds on this idea to generate prompts that will result in images that satisfy the desired criteria.

3 Method

3.1 Problem Statement

First, let $x \in \mathcal{X}$ denote an image, and $y \in \mathcal{Y}$ denote a textual prompt. Given a collection of reference images $\{x_i\}_{i=1}^M$, a prompt engineer $\mathbf{F} : \mathcal{X} \to \Delta(\mathcal{Y})$ samples a candidate prompt y corresponding to each reference image x, i.e., $y \sim p_{\theta_\mathbf{F}}(y \mid x)$. A T2I generative model $\mathbf{G} : \mathcal{Y} \to \Delta(\mathcal{X})$ then uses this candidate prompt to generate a new image, $x \sim p_{\theta_\mathbf{G}}(x \mid y)$, and a judge model $\mathbf{D} : \mathcal{X} \times \mathcal{X} \to [0, 1]$ then scores the visual similarity between the images based on some criteria. Our goal is then to find the

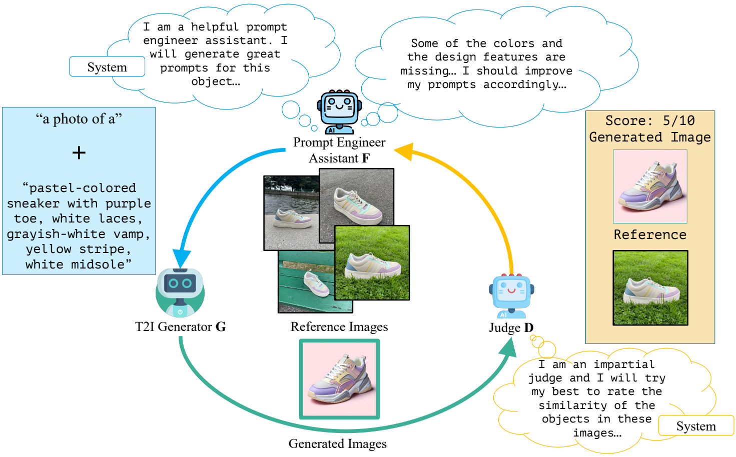

System
I am a helpful prompt engineer assistant. I will generate great prompts for this object...
Some of the colors and the design features are missing... I should improve my prompts accordingly...
"a photo of a"
+
"pastel-colored sneaker with purple toe, white laces, grayish-white vamp, yellow stripe, white midsole"
Prompt Engineer Assistant F
Score: 5/10
Generated Image
Reference
T2I Generator G
Reference Images
Judge D
I am an impartial judge and I will try my best to rate the similarity of the objects in these images...
System
Generated Images

Fig. 2: An illustration of PRISM. The label "System" indicates the system prompts setups for the multimodal LLMs.

best prompt:

$$y^\star\left(\{x_i\}_{i=1}^M\right) = \arg\max_{y\in\mathcal{Y}} \sum_{i=1}^{M} \texttt{Score}(x_i, y), \tag{1}$$

where $\texttt{Score}(x_{\text{target}}, y) = \mathbb{E}_{x\sim p_{\theta_{\mathbf{G}}}(x|y)}\left[\mathbf{D}(x, x_{\text{target}})\right]$.

The criteria can be any visual similarity metric that may or may not be easy to specify in a closed form, including "*how similar are the main objects in the images*" or "*how similar are the styles of the image*" or "*how similar are the two images in general*". The resulting $y^\star$ should be able to generate an image that is very close to the reference images based on the criteria with some (possibly unseen) T2I models $p_\theta(x \mid y)$.

3.2 Algorithm

Our method, Prompt Refinement and Iterative Sampling Mechanism (PRISM), is an iterative process that repeats a prompt-refinement subroutine for K iterations in N parallel streams, where $N \times K$ is a predetermined compute budget. At iteration k, the n-th stream of PRISM randomly selects a reference image $x_{k,n}$ from $\{x_i\}_{i=1}^M$ and uses $\mathbf{F}$ to sample a candidate prompt $y_{k,n}$ from $p_{\theta_{\mathbf{F}}}(y \mid x_{k,n})$. Then it queries $\mathbf{G}$ to generate a single $\hat{x}_{k,n}$ from $y_{k,n}$ with $p_{\theta_{\mathbf{G}}}(x \mid y_{k,n})$ and evaluate the prompt with $\mathbf{D}$ to obtain an in-iteration score $\texttt{Score}'(x_{k,n}, y_{k,n}) = \mathbf{D}(x_{k,n}, \hat{x}_{k,n})$. At the end of the iteration, we use the generated $y_{k,n}$ and its score to update $p_{\theta_{\mathbf{F}}}(y \mid x)$. After

Algorithm 1 Prompt Refinement and Iterative Sampling Mechanism (PRISM)

```
Input: $N$ streams, $K$ iterations, $\{x_i\}_{i=1}^M$ reference images
Output: Best prompt $y^\star$ based on total score
for $n = 1$ to $N$ in parallel do
    for $k = 1$ to $K$ do
        Randomly sample an $x_{k,n}$ from $\{x_i\}_{i=1}^M$
        $\mathbf{F}$ samples $y_{k,n} \sim p_{\theta_\mathbf{F}}(y \mid x_{k,n})$
        $\mathbf{G}$ samples $\hat{x}_{k,n} \sim p_{\theta_\mathbf{G}}(x \mid y_{k,n})$
        $\mathbf{D}$ calculates an in-iteration score $\texttt{Score}'(x_{k,n}, y_{k,n}) = \mathbf{D}(x_{k,n}, \hat{x}_{k,n})$
        Update $p_{\theta_\mathbf{F}}$ based on $x_{k,n}, \hat{x}_{k,n}, y_{k,n}, \texttt{Score}'(x_{k,n}, y_{k,n})$ and the chat history of stream $n$
    end for
end for
Collect the subset $\{y_c\}_{c=1}^C$ with the $C$-best in-iteration scores
Re-evaluate this subset with total score $\sum_{i=1}^M \texttt{Score}(x_i, y_c)$
Return the prompt with the best total score. In case of a tie, return the
prompt with the highest log-likelihood.
```

the entire process, we collect the subset of $\{y_c\}_{c=1}^C$ generated throughout this process that has the C-best in-iteration scores. Then we re-evaluate this subset with the total score $\sum_{i=1}^M \texttt{Score}(x_i, y_c)$ and return the prompt with the best total score. If there is a tie, then we return the prompt with the highest log likelihood [1]. The pseudocode for the algorithm is outlined in Algorithm 1 and we also illustrate this algorithm in Figure 2.

The key difference between PRISM and prior methods is that PRISM updates the entire sampling distribution of prompts, whereas prior works [6, 16, 39] directly update the tokens of a single prompt or the embedding of the prompt. We believe that maintaining the whole prompt distribution is beneficial as text-to-image generation is not a one-to-one operation, i.e. an image can be described by multiple different text prompts and the same text prompt can correspond to multiple different generated images. Having access to the whole distribution allows the method to sample a more diverse range of prompts without starting from scratch and may also help the optimization escape potential local optima.

Since PRISM only requires samples from $\mathbf{G}$ conditioned on the prompts, one may use any T2I generative model of their choice. On the other hand, more careful treatment is required for designing $\mathbf{F}$ and $\mathbf{D}$. We will elaborate on these design decisions below.

3.3 Designing and updating $\mathbf{F}$ and $p_{\theta_{\mathbf{F}}}$

What is $p(y \mid x)$? In general, it is not obvious what the joint or the conditional distribution of all text and images is, so some form of approximation is unavoidable. In the context of image generation, a natural choice of the image-conditioned text distribution is an image captioning model. Traditional captioning models, however, fall short in controlled image generation for two primary reasons: **(1)** The level of detail necessary for generating specific images far exceeds what generic captioning models provide [13]; **(2)** effective prompts for T2I models are often not grammatically correct sentences but rather collection of phrases that describe the details about the image, which generic captioning models are not trained to generate. For example, in Figure 5, the second reference image is generated by the prompt *"A broken robot lays on the ground with plants growing over it, somber, HD, hyper realistic, intricate detail"* with Stable Diffusion, but a caption for this image will not include components like "HD" or "hyper realistic". As a result, instead of "a good description of an image", we wish to directly model "possible prompts that are used to generate this image".

Desiderata A desirable $\mathbf{F}$ can sample from a distribution $p_{\theta_{\mathbf{F}}}(y \mid x)$ that models "the prompt that can be used to generate this image", and it should also be easily updated if the current generation is suboptimal. Ideally, such an update can be done without any retraining or fine-tuning since these operations are generally expensive and incompatible with black-box T2I models.

Multimodal LLM stands out as the ideal choice for $\mathbf{F}$ due to their ability to directly tailor the generation of prompts via system prompts and to adapt through in-context learning without requiring access to the model's parameters. Specifically, since the model can ingest both images and texts, we can incorporate the reference images, intermediate prompts and generated images, and the score associated with the generated images all in the context of the LLM. Then, the model can be prompted to jointly reason over all available information and perform in-context learning. The in-context learning facilitates iterative refinement of the prompt to update the posterior distribution based on feedback or even additional human instructions, without the need for model retraining. Concretely, the model would process how the image generative model is affected by different prompts, propose improvements, and create new prompts, much like a prompt engineer. More precisely, in practice, we design system prompts that explicitly condition the LLM to generate improvements and new prompts given the results from the previous iterations, similar to the chain-of-thought [38] technique.

3.4 Designing the judge model D

We have a wider range of choices for the judging model as long as it provides a notion of similarity between a pair of images. A simple solution is to use pre-trained discriminative models such as CLIP [22] and DINO [21], and measure the distance of images in their embedding spaces. These models have seen various degrees of success but come with inherent limitations – the discriminative objective (e.g., contrastive loss) does not incentivize the model to attend to fine-grained details since they do not improve the objective further, an issue similar to the shortcomings of using captioning models to generate prompts [13]. Moreover, in image generation, the criteria of success can be nuanced and difficult to quantify through traditional distance or similarity functions yet can be effortlessly described in human language. Lastly, the similarity we wish to measure may only involve some part of the visual features (e.g. color, painting stroke type, etc), and not all applications share the same notion of similarity. If we want to use pre-trained discriminative models, then we need to find a different model for each specific task, which can be impractical.

In light of these challenges, an ideal judge model should be maximally flexible for different kinds of criteria and can perform fine-grained analysis of the images. Once again, a multimodal LLM emerges as the perfect candidate: using system prompts and in-context learning, we can easily specify metrics that may be otherwise difficult to describe or evaluate and even intervene in the reasoning chain if we want to, and, more importantly, the same model can be applied to a wide range of tasks.

4 Experiments

4.1 Experimental Settings

Implementation Details For all of our experiments, we choose GPT-4V [20] as both the prompt engineer assistant model **F** and the judge **D**. We also fix the T2I generator as SDXL-Turbo [27] for all of our experiments. We design different system prompts for both **F** and **D** for each task and we provide details about the system prompts in the appendix.

We evaluate the prompts generated from PRISM and baselines with five different T2I models. In particular, we choose two open-sourced models, Stable Diffusion 2.1 (SD 2.1) and SDXL-Turbo, and two closed-sourced models, Dall-E 2 and Dall-E 3, to quantitatively measure the performance. We also qualitatively showcase results from Midjourney, which is another closed-sourced T2I platform. For SD 2.1 and SDXL-Turbo, we clip all prompt lengths to 77 due to their context length constraint.

We compare PRISM and baselines in two settings: personalized T2I generation and direct image inversion, and we will elaborate on the task definitions in their corresponding sections below. For personalized T2I generation, we use a maximum budget of 40 and report the quantitative results from the setting $N = 10, K = 4$. For direct image inversion, we use a maximum budget of 30 and report the quantitative results from the setting $N = 6, K = 5$. To simplify the implementation, we only keep a chat history length of 3 and use the length of the prompt as an approximation of the prompt log-likelihood in the final prompt selection. For direct image inversion, we re-evaluate the top 5 candidates twice and tally the score with the in-iteration scores to make the final decision. For personalized T2I generation, we re-evaluate once for each reference image and use the average score to select the output.

Baselines We choose Textual Inversion (TI) [6], BLIP-2 (BLIP2) [12], CLIP-Interrogator (CLIP-Int) and PEZ [39] as the baselines. Textual Inversion trains a "soft token" which cannot be directly translated into regular human language to represent the concepts in the reference images. BLIP-2 is the state-of-the-art image captioning model. CLIP-Interrogator[1] combines BLIP-2 captions with a suffix which is created by searching a pre-collected bank of keywords using CLIP [22] score. PEZ is a gradient-based optimization method that searches for the best combination of existing tokens in the vocabulary with CLIP similarity. We use OpenCLIP-ViT-H-14 trained on LAION2B [28] for both CLIP-Int and PEZ and use Blip2-Flan-T5-XL for both CLIP-Int and BLIP-2. Notice that TI requires training on individual models and CLIP-Int requires a pre-collected keyword bank, both of which provides unfair advantages over our setting.

Evaluation Metrics We evaluate the prompt interpretability using mean negative log-likelihood (NLL) calculated from Mistral 7B [11]. For image quality evaluation, we mainly measure the CLIP image similarity score (CLIP-I) to quantify the difference between the generated images and the reference images. Following [26], we also use DINO V2 [21] embedding similarity to calculate the object-sensitive image similarity for the personalized T2I generation task. We chose CLIP-ViT-L-14 and DINO-V2-Base as the base models. For Dall-E 2 and Dall-E 3, we also compare the number of times each method fails to pass its black-box safeguard. More failures indicate a higher potential to produce unsafe prompts. For each prompt, we allow 5 attempts before counting it as a failure.

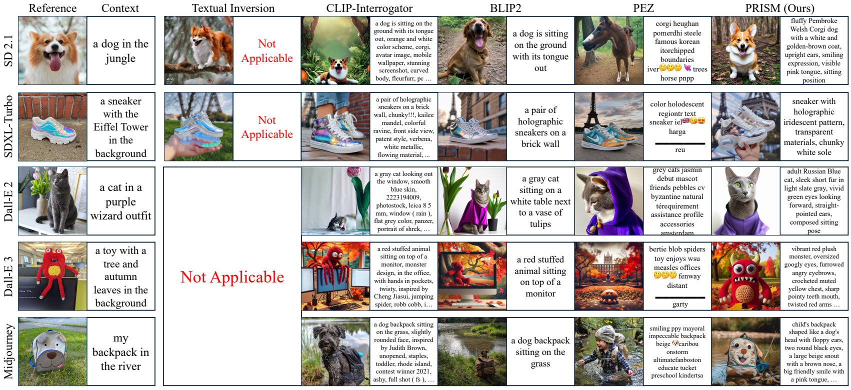

Reference
Context
Textual Inversion
CLIP-Interrogator
BLIP2
PEZ
PRISM (Ours)
SD 2.1
a dog in the jungle
Not Applicable
a dog is sitting on the ground with its tongue out, orange and white color scheme, corgi, avatar image, mobile wallpaper, stunning screenshot, curved body, fleurfurr, pc …
a dog is sitting on the ground with its tongue out
corgi heughan pomerdhi steele famous korean itorchipped boundaries iver🤔🤔🤔 💘 trees horse pnpp
fluffy Pembroke Welsh Corgi dog with a white and golden-brown coat, upright ears, smiling expression, visible pink tongue, sitting position
SDXL-Turbo
a sneaker with the Eiffel Tower in the background
Not Applicable
a pair of holographic sneakers on a brick wall, chunky!!!, kailee mandel, colorful ravine, front side view, patent style, verbena, white metallic, flowing material, ...
a pair of holographic sneakers on a brick wall
color holodescent regiontr text sneaker iel🇬🇧😘😍 harga reu
sneaker with holographic iridescent pattern, transparent materials, chunky white sole
Dall-E 2
a cat in a purple wizard outfit
Not Applicable
a gray cat looking out the window, smooth blue skin, 2223194009, photostock, leica 8 5 mm, window (rain), flat grey color, panzer, portrait of shrek, …
a gray cat sitting on a white table next to a vase of tulips
grey cats jasmin debut mascot friends pebbles cv byzantine natural térequirement assistance profile accessories amsterdam
adult Russian Blue cat, sleek short fur in light slate gray, vivid green eyes looking forward, straight-pointed ears, composed sitting pose
Dall-E 3
a toy with a tree and autumn leaves in the background
a red stuffed animal sitting on top of a monitor, monster design, in the office, with hands in pockets, twisty, inspired by Cheng Jiasui, jumping spider, robb cobb, i…
a red stuffed animal sitting on top of a monitor
bertie blob spiders toy enjoys wsu measles offices 🤔🤔🤔 fenway distant garty
vibrant red plush monster, oversized googly eyes, furrowed angry eyebrows, crocheted muted yellow chest, sharp pointy teeth mouth, twisted red arms …
Midjourney
my backpack in the river
a dog backpack sitting on the grass, slightly rounded face, inspired by Judith Brown, unopened, staples, toddler, rhode island, contest winner 2021, ashy, full shot (fs), …
a dog backpack sitting on the grass
smiling ppy mayoral impeccable backpack beige 🐶caribou onstorm ultimatefanboston educate tucket preschool kindertsa
child's backpack shaped like a dog's head with floppy ears, two round black eyes, a large beige snout with a brown nose, a big friendly smile with a pink tongue, …

Fig. 3: Qualitative results for personalized T2I generation on DreamBooth dataset.

Table 1: Personalized T2I results on DreamBooth dataset. Bold fonts indicate the best score and underlines indicate the second best score.

Method	Prompt	SD 2.1		SDXL Turbo		Dall-E 2			Dall-E 3		
	NLL ↓	CLIP-I ↑	DINO ↑	CLIP-I ↑	DINO ↑	CLIP-I ↑	DINO ↑	Failed ↓	CLIP-I ↑	DINO ↑	Failed ↓
TI (SD 2.1)	-	0.707	0.443	-	-	-	-	-	-	-	-
TI (SDXL)	-	-	-	**0.771**	**0.504**	-	-	-	-	-	-
CLIP-Int	4.361	0.733	0.446	0.756	0.490	0.711	0.464	13.3%	0.619	0.386	1.1%
BLIP2	4.378	0.706	0.408	0.729	0.456	0.707	0.430	**6.9%**	0.655	0.377	0.3%
PEZ	6.188	0.709	0.384	0.722	0.418	0.676	0.389	16.7%	0.618	0.344	1.1%
PRISM (Ours)	**3.466**	**0.743**	**0.464**	0.770	0.499	**0.734**	**0.482**	**6.9%**	**0.734**	**0.464**	**0.1%**

4.2 Personalized Text-to-Image Generation

We first demonstrate PRISM's ability to find human-interpretable and accurate prompts to describe certain objects and styles in the task of personalized T2I generation. Given a set of reference images that depict the same concept (such as objects and style), personalized T2I tasks require the model to synthesize images in new contexts while maintaining the original concept.

Datasets We use the dataset collected by DreamBooth [26] to quantitatively compare the performance in personalized T2I generation. The DreamBooth dataset contains 30 daily objects, and each subject has 4-6 images. For each subject, we adopt the 25 prompt templates curated by DreamBooth to create varying contexts and scenarios to test the fidelity of the subject representation in diverse settings. We generate 4 images for each subject and template combination with open sourced T2I models, and 1 image for each combination with closed sourced T2I models. For methods that directly find English prompts, we use the class noun to fill in the template and the output prompts that describe these concepts serve as suffixes. For Textual Inversion, we follow the original setting for the templates.

We also qualitatively demonstrate the ability to represent a certain artistic style using Wikiart dataset [34]. We use three images from each artist as reference images. To create diverse scenes, we follow [8] and use descriptive prompts from PartiPrompts [42] as prefixes to the output prompts similar to the previous setting.

DreamBooth Dataset Results Table 1 and Figure 3 respectively show the quantitative and qualitative results on the DreamBooth dataset. As we can observe, PRISM achieves the best performance across the board except for the image similarity metrics for SDXL-Turbo. PRISM is the only method in our experiments that can produce fully human-readable prompts for these subjects. In particular, we can observe that PEZ renders completely indecipherable texts, BLIP-2 only describes the general scene but fails to mention any visual details and textual inversion is entirely not interpretable since it produces soft embeddings. Since CLIP-Interrogator combines the results from BLIP-2 and a CLIP search, it improves the interpretability over PEZ-like gradient search-only method. However, it still falls short in terms of human readability in comparison to our method.

In terms of image quality and object fidelity, we also find PRISM to constantly achieve accurate depiction of the target subject while the baselines sometimes struggle to capture all the details. Fine-grained details such as the color of the dog and cat fur and the shape of the shoe sole are better described and reflected with our method. And out of the four training-free methods we experiment with, PRISM is the only one that can tackle complicated objects such as the red monster toy and the dog-shaped backpack as shown in Figure 3 when all the other methods fail to generate similar objects. Due to the nature of their methodologies, BLIP-2 and CLIP-Interrogator also capture the background and other irrelevant elements in the scene when describing the objects. However, unlike our method, where we can directly specify the tasks and the judging criteria in the system prompts of the LLMs, there is no simple way to automatically filter out those irrelevant elements in BLIP-2 and CLIP-Interrogator's outputs. Even though Textual Inversion obtains marginally higher CLIP-I scores and DINO scores with SDXL-Turbo, notice that Textual Inversion requires a lot more modeling assumptions than our method, and the new embeddings it learns are not transferable – not even to SD 2.1, because SDXL models use two text encoders whereas SD 2.1 model use only one.

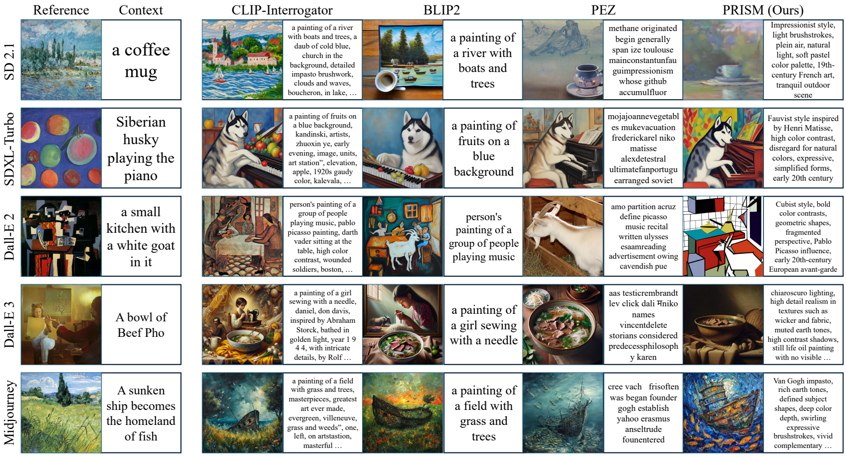

Reference
Context
CLIP-Interrogator
BLIP2
PEZ
PRISM (Ours)
SD 2.1
a coffee mug
a painting of a river with boats and trees, a daub of cold blue, church in the background, detailed impasto brushwork, clouds and waves, boucheron, in lake, …
a painting of a river with boats and trees
methane originated begin generally span ize toulouse mainconstantunfau guimpressionism whose github accumulfluor
Impressionist style, light brushstrokes, plein air, natural light, soft pastel color palette, 19th-century French art, tranquil outdoor scene
SDXL-Turbo
Siberian husky playing the piano
a painting of fruits on a blue background, kandinski, artists, zhuoxin ye, early evening, image, units, art station", elevation, apple, 1920s gaudy color, kalevala, …
a painting of fruits on a blue background
mojajoannevegetabl es mukevacuation frederickarel niko matisse alexdetestral ultimatefanportugu earranged soviet
Fauvist style inspired by Henri Matisse, high color contrast, disregard for natural colors, expressive, simplified forms, early 20th century
Dall-E 2
a small kitchen with a white goat in it
person's painting of a group of people playing music, pablo picasso painting, darth vader sitting at the table, high color contrast, wounded soldiers, boston, …
person's painting of a group of people playing music
amo partition acruz define picasso music recital written ulysses esaamreading advertisement owing cavendish pue
Cubist style, bold color contrasts, geometric shapes, fragmented perspective, Pablo Picasso influence, early 20th-century European avant-garde
Dall-E 3
A bowl of Beef Pho
a painting of a girl sewing with a needle, daniel, don davis, inspired by Abraham Storck, bathed in golden light, year 1 9 4 4, with intricate details, by Rolf …
a painting of a girl sewing with a needle
aas testicrembrandt lev click dali मniko names vincentdelete storians considered predecessphilosoph y karen
chiaroscuro lighting, high detail realism in textures such as wicker and fabric, muted earth tones, high contrast shadows, still life oil painting with no visible …
Midjourney
A sunken ship becomes the homeland of fish
a painting of a field with grass and trees, masterpieces, greatest art ever made, evergreen, villeneuve, grass and weeds", one, left, on artstastion, masterful …
a painting of a field with grass and trees
cree vach frisoften was began founder gogh establish yahoo erasmus anseltrude founentered
Van Gogh impasto, rich earth tones, defined subject shapes, deep color depth, swirling expressive brushstrokes, vivid complementary …

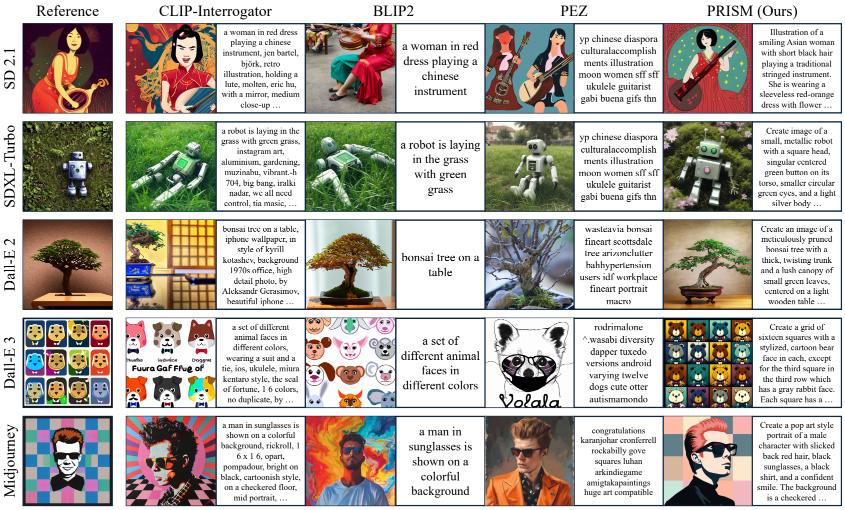
Fig. 4: Qualitative results for personalized style T2I generation on Wikiart dataset.

When transferring the output prompts to black-box T2I models, our method shows even larger advantages over the baselines. We also observe that our method produces the fewest unsafe prompts judged by Dall-E safeguards, while the baselines can fail to pass the safeguard up to almost 16.7% of the time.

Wikiart Results In Figure 4, we also show a qualitative comparison between our method and the baselines on the Wikiart dataset. We find that our method is capable of precisely identifying the genres, eras, and sometimes even the names of the artists when describing the style of the reference artworks. On the other hand, we observe that the baselines fail to recognize these crucial keywords, even when they have access to a pre-collected bank of words that are supposed to provide accurate descriptions of the style. In addition, PRISM can also provide other fine-grained details such as the style of the pen strokes and color palettes in a human-interpretable way to better assist the generation of the target style.

4.3 Direct Image Inversion

To demonstrate the versatility of our method, we also compare PRISM with the baselines in the task of direct image inversion. In this task, the goal is to directly find the prompt that can exactly generate the input image. Here the number of reference images is $M = 1$ and we aim to capture all aspects of the image, including the subjects, background, theme, style, and other details in the scene.



	Reference	CLIP-Interrogator	BLIP2	PEZ	PRISM (Ours)
SD 2.1		a woman in red dress playing a chinese instrument, jen bartel, björk, retro illustration, holding a lute, molten, eric hu, with a mirror, medium close-up …	a woman in red dress playing a chinese instrument	yp chinese diaspora culturalaccomplish ments illustration moon women sff sff ukulele guitarist gabi buena gifs thn	Illustration of a smiling Asian woman with short black hair playing a traditional stringed instrument. She is wearing a sleeveless red-orange dress with flower …
SDXL-Turbo		a robot is laying in the grass with green grass, instagram art, aluminium, gardening, muzinabu, vibrant.-h 704, big bang, iralki nadar, we all need control, tia masic, …	a robot is laying in the grass with green grass	yp chinese diaspora culturalaccomplish ments illustration moon women sff sff ukulele guitarist gabi buena gifs thn	Create image of a small, metallic robot with a square head, singular centered green button on its torso, smaller circular green eyes, and a light silver body …
Dall-E 2		bonsai tree on a table, iphone wallpaper, in style of kyrill kotashev, background 1970s office, high detail photo, by Aleksandr Gerasimov, beautiful iphone …	bonsai tree on a table	wasteavia bonsai fineart scottsdale tree arizonclutter bahhypertension users idf workplace fineart portrait macro	Create an image of a meticulously pruned bonsai tree with a thick, twisting trunk and a lush canopy of small green leaves, centered on a light wooden table …
Dall-E 3		a set of different animal faces in different colors, wearing a suit and a tie, ios, ukulele, miura kentaro style, the seal of fortune, 1 6 colors, no duplicate, by …	a set of different animal faces in different colors	rodrimalone ^.wasabi diversity dapper tuxedo versions android varying twelve dogs cute otter autismamondo	Create a grid of sixteen squares with a stylized, cartoon bear face in each, except for the third square in the third row which has a gray rabbit face. Each square has a …
Midjourney		a man in sunglasses is shown on a colorful background, rickroll, 1 6 x 1 6, opart, pompadour, bright on black, cartoonish style, on a checkered floor, mid portrait, …	a man in sunglasses is shown on a colorful background	congratulations karanjohar cronferrell rockabilly gove squares luhan arkindiegame amigtakapaintings huge art compatible	Create a pop art style portrait of a male character with slicked back red hair, black sunglasses, a black shirt, and a confident smile. The background is a checkered …

Fig. 5: Image inversion results for different methods on different T2I models.

Datasets We use images from the DiffusionDB dataset [35] for the direct image inversion task. This dataset includes a wide variety of image pairs generated by Stable Diffusion and we choose a random sample of 100 images from the `large_random_10k` split on Huggingface.

Results As shown in Table 2, we can first obtain a similar observation in terms of human-interpretability to the previous experiment. For direct image inversion, we also immediately see a significant improvement in the readability of inverted prompts using PRISM. While expected for methods, such as PEZ, in which the process of image inversion has no language prior, we find that our method finds text that more closely aligns with a learned distribution of English language text (i.e. Mistral logits) than CLIP-Interrogator and BLIP2.

When comparing the image quality, we first note that because all images in DiffusionDB are generated by Stable Diffusion, which is exactly the model design space of CLIP-Interrogator and PEZ, it gives significant modeling assumption advantages to these baselines over our method when testing on Stable Diffusion models. This advantage enables relatively high performance for these baselines in Stable Diffusion models, but it does not transfer well into other closed-sourced models. In fact, we can even observe that CLIP-Interrogator generates the highest quality images with SD 2.1, which is the weakest model in this comparison and generates the lowest quality images with Dall-E 3, which is the strongest T2I model in

Table 2: Metrics for the image inversion results. old fonts indicate the best score and underlines indicate the second best score.

Method	Prompt	SD 2.1	SDXL TUrbo	Dall-E 2		Dall-E 3	
	NLL ↓	CLIP-I ↑	CLIP-I ↑	CLIP-I ↑	Failed ↓	CLIP-I ↑	Failed ↓
CLIP-Int	4.193	**0.800**	**0.783**	**0.761**	17.0%	0.719	**0.0%**
BLIP2	4.299	0.710	0.707	0.687	**2.0%**	0.695	**0.0%**
PEZ	6.736	0.746	0.726	0.616	3.0%	0.635	**0.0%**
PRISM (Ours)	**2.762**	0.749	0.776	0.741	**2.0%**	**0.767**	**0.0%**

this table. This phenomenon indicates that the design choices of CLIP-Interrogate and PEZ are heavily overfitted to achieve high performance on Stable Diffusion, but provide poor generalizability to other models. On the other hand, the prompts produced by our method generalize significantly better than the baselines and we achieve the best results on Dall-E 3.

Qualitatively, our method also provides prompts that are both semantically aligned with and can generate images that are visually similar to the reference. In particular, Figure 5, shows that we can find text that aligns with the image, even when those images have particularly unique features. For example, in Figure 5 Dall-E3 generated a grid of images of animal faces. Not only does the PRISM's prompt explicitly include a request for this grid structure, unlike our comparison methods, but it also takes into account the coloration of the background in the reference. In the second row of Figure 5, our method is also the only method that captures the small flowers in the grass, showcasing the capability of identifying and reflecting small fine-grained details from the reference.

4.4 Ablation Study

Comparison with GPT-4V. PRISM relies on having a strong multimodal LLM. In our case, we chose GPT-4V as the multimodal foundation model. While in principle we may use any multimodal LLM, it is nonetheless useful to understand what benefits PRISM adds to an already capable foundation model. To show the effectiveness of iterative prompt refinement and parallel search, we compare our method with GPT-4V's zero-shot performance with the same system prompts for object and image inversion tasks on SDXL-Turbo. We see in Table 6 that PRISM consistently outperforms GPT-4V's zero-shot performance, although the latter is already compelling. In Figure 7, we show some examples of the generated results.

Fig. 6: Comparison with GPT-4V in both personalized T2I generation and direct image inversion experiments.

Method	Image		Object	
	NLL	CLIP-I	NLL	CLIP-I
GPT-4V	**2.356**	0.756	**3.393**	0.757
PRISM (Ours)	2.680	**0.777**	3.466	**0.770**

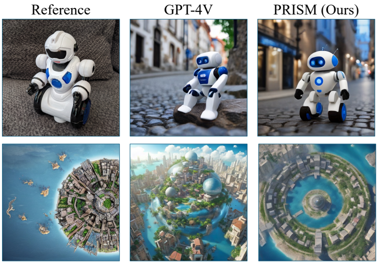


Fig. 7: Qualitative comparison with GPT-4V.

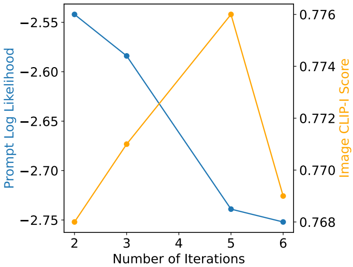


Fig. 8: Ablation study on the trade-off between N and K. All runs shown in this plot have the same budget $N \times K = 30$, but each run operates a different number of iterations K.

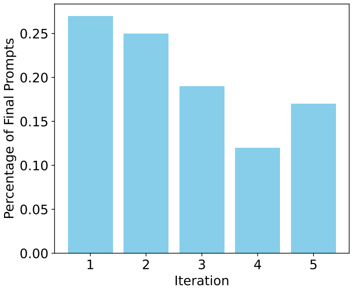
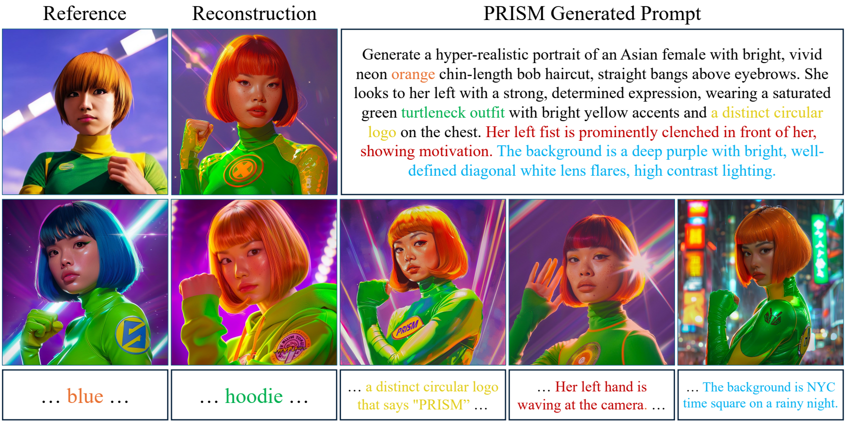


Fig. 9: The distribution of the final selected prompts in each iteration for the image inversion experiment. Here $N = 6$ and $K = 5$.

We see that qualitatively GPT-4V can capture the high-level semantics of the reference images but still misses more fine-grained details.

Trade-off between N and K PRISM has two hyperparameters N and K which control the amount of parallel search and the depth of iterative refinement. Figure 8 shows a trade-off between N and K with the same budget $N \times K = 30$. Similar to the findings of [3], we find that performance can degrade if the refinement is repeated too many times (i.e., K is too large), and in general, we do not recommend practitioners with small budgets to go beyond $K = 5$. Unlike jailbreaking [3], we observe that the optimal N and K can vary depending on the task: if the target concept is simple (e.g. a commonly seen dog), then small N and K are generally sufficient, and prioritizing N tends to be more helpful. However, if the target concept is rarer and more complicated (e.g. a very specific toy), a



Generate a hyper-realistic portrait of an Asian female with bright, vivid neon orange chin-length bob haircut, straight bangs above eyebrows. She looks to her left with a strong, determined expression, wearing a saturated green turtleneck outfit with bright yellow accents and a distinct circular logo on the chest. Her left fist is prominently clenched in front of her, showing motivation. The background is a deep purple with bright, well-defined diagonal white lens flares, high contrast lighting.

… blue …

… hoodie …

… a distinct circular logo that says "PRISM" …

… Her left hand is waving at the camera. …

… The background is NYC time square on a rainy night.

Fig. 10: Prompt editing demonstration with Midjourney.

larger reasoning depth (i.e., larger K) would be more helpful. In Figure 9, we show the distribution of iteration numbers at which the best prompt is found in the image inversion experiment. In practice, one may tune these hyperparameters further for specific use cases.

4.5 Prompt Editing

Because the prompts produced by PRISM is very human-interpretable, after obtaining a prompt from the reference images, one can easily modify the output prompts to change attributes in their desired generated images. Figure 10 demonstrates an examples of prompt editing with PRISM in Midjourney. With simple and intuitive prompt edit, we are able to change specific attributes of the images while keeping the other components in the scene unchanged.

5 Conclusion and Broader Impact Statement

In this paper, we propose PRISM, an algorithm that automatically creates human-interpretable and accurate text prompts for text-to-image generative models, based on visual concepts provided by reference images. Our method iteratively refines the sampling distribution of the text prompt via LLM in-context learning and is capable of creating prompts that are transferable to any T2I model, including black-box platforms like Dall-E and Midjourney. However, just as LLMs are suceptible to being jailbroken or adversarially manipulated by malicious actors [48], our method may also be vulnerable to malicious intent, potential bias, or limitations in the base models. Therefore, we intent to implement necessary safeguards upon the public release of our code and are committed to keep up with future advancements in improving the safety of our method.

Acknowledgements

This work was supported in part of the ONR grant N000142312368. We also thank Sony AI for sponsoring part of this work. A.R., H.H., and G.P. are supported by the NSF Institute for CORE Emerging Methods in Data Science (EnCORE). A.R. is also supported by an ASSET Amazon AWS Trustworthy AI Fellowship. Y.J. is supported by Google PhD Fellowship.

References

1. Akinwande, V., Jiang, Y., Sam, D., Kolter, J.Z.: Understanding prompt engineering may not require rethinking generalization. In: Proc. ICLR (2024), `https://openreview.net/forum?id=a745RnSFLT`
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Proc. NeurIPS **33**, 1877–1901 (2020)
3. Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G.J., Wong, E.: Jailbreaking black box large language models in twenty queries. arXiv preprint arXiv:2310.08419 (2023)
4. Chen, W., Hu, H., Li, Y., Ruiz, N., Jia, X., Chang, M.W., Cohen, W.W.: Subject-driven text-to-image generation via apprenticeship learning. Proc. NeurIPS **36**, 30286–30305 (2023)
5. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: Proc. ICLR (2023), `https://openreview.net/forum?id=OnD9zGAGT0k`
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: Proc. ICLR (2023), `https://openreview.net/forum?id=NAQvF08TcyG`
7. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In: Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL-IJCNLP 2021. pp. 3816–3830. Association for Computational Linguistics (ACL) (2021)
8. He, Y., Murata, N., Lai, C.H., Takida, Y., Uesaka, T., Kim, D., Liao, W.H., Mitsufuji, Y., Kolter, J.Z., Salakhutdinov, R., Ermon, S.: Manifold preserving guided diffusion. In: Proc. ICLR (2024), `https://openreview.net/forum?id=o3BxOLoxm1`
9. He, Y., Salakhutdinov, R., Kolter, J.Z.: Localized text-to-image generation for free via cross attention control (2023)

10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. Proc. CVPR (2017)
11. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023)
12. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proc. ICML. JMLR.org (2023)
13. Liang, P.P., Cheng, Y., Fan, X., Ling, C.K., Nie, S., Chen, R.J., Deng, Z., Allen, N., Auerbach, R., Mahmood, F., Salakhutdinov, R., Morency, L.P.: Quantifying & modeling multimodal interactions: An information decomposition framework. In: Proc. NeurIPS (2023), https://openreview.net/forum?id=J1gBijopla
14. Liu, X., Xu, N., Chen, M., Xiao, C.: Autodan: Generating stealthy jailbreak prompts on aligned large language models. arXiv preprint arXiv:2310.04451 (2023)
15. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. pp. 8086–8098. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.ACL-LONG.556, https://doi.org/10.18653/v1/2022.acl-long.556
16. Mahajan, S., Rahman, T., Yi, K.M., Sigal, L.: Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. arXiv preprint arXiv:2312.12416 (2023)
17. Manikandan, H., Jiang, Y., Kolter, J.Z.: Language models are weak learners. In: Proc. NeurIPS. vol. 36, pp. 50907–50931 (2023), https://openreview.net/forum?id=559NJBfN20
18. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: Proc. ICLR (2022)
19. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
20. OpenAI: Gpt-4 technical report (2023)
21. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable

visual models from natural language supervision. In: Proc. ICLR. pp. 8748–8763. PMLR (2021)
23. Robey, A., Wong, E., Hassani, H., Pappas, G.J.: Smoothllm: Defending large language models against jailbreaking attacks. arXiv preprint arXiv:2310.03684 (2023)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc. CVPR. pp. 10684–10695 (2022)
25. Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A.G., Shakkottai, S.: Solving linear inverse problems provably via posterior sampling with latent diffusion models. arXiv preprint arXiv:2307.00619 (2023)
26. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proc. CVPR. pp. 22500–22510 (2023)
27. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation (2023)
28. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022)
29. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning (2023)
30. Shin, T., Razeghi, Y., IV, R.L.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. pp. 4222–4235. Association for Computational Linguistics (2020). `https://doi.org/10.18653/V1/2020.EMNLP-MAIN.346`, `https://doi.org/10.18653/v1/2020.emnlp-main.346`
31. Song, J., Vahdat, A., Mardani, M., Kautz, J.: Pseudoinverse-guided diffusion models for inverse problems. In: Proc. ICLR (2022)
32. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: Proc. ICLR (2021)
33. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012), `https://proceedings.neurips.cc/paper_files/paper/2012/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf`
34. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE Transactions on Image Processing **28**(1), 394–409 (2019). `https://doi.org/10.1109/TIP.2018.2866698`, `https://doi.org/10.1109/TIP.2018.2866698`
35. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. arxiv preprint arXiv:2210.14896 (2022)

36. Webson, A., Pavlick, E.: Do prompt-based models really understand the meaning of their prompts? In: Carpuat, M., de Marneffe, M., Ruíz, I.V.M. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022. pp. 2300–2344. Association for Computational Linguistics (2022). https://doi.org/10.18653/V1/2022.NAACL-MAIN.167, https://doi.org/10.18653/v1/2022.naacl-main.167
37. Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: How does llm safety training fail? Advances in Neural Information Processing Systems **36** (2024)
38. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023)
39. Wen, Y., Jain, N., Kirchenbauer, J., Goldblum, M., Geiping, J., Goldstein, T.: Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. In: Proc. NeurIPS. vol. 36, pp. 51008–51025 (2023), https://openreview.net/forum?id=VOstHxDdsN
40. Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D., Chen, X.: Large language models as optimizers. In: Proc. ICLR (2024), https://openreview.net/forum?id=Bb4VGOWELI
41. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arxiv:2308.06721 (2023)
42. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., Hutchinson, B., Han, W., Parekh, Z., Li, X., Zhang, H., Baldridge, J., Wu, Y.: Scaling autoregressive models for content-rich text-to-image generation. Transactions on Machine Learning Research (2022), https://openreview.net/forum?id=AFDcYJKhND, featured Certification
43. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: FreeDoM: Training-free energy-guided conditional diffusion model. arXiv:2303.09833 (2023)
44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proc. ICCV (2023)
45. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
46. Zhou, Y., Muresanu, A.I., Han, Z., Paster, K., Pitis, S., Chan, H., Ba, J.: Large language models are human-level prompt engineers. In: Proc. ICLR (2023), https://openreview.net/forum?id=92gvk82DE-
47. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
48. Zou, A., Wang, Z., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043 (2023)

A Additional Experiment Details

In this section, we provide further details about the implementation of our experiments. For all quantitative analysis that uses Stable Diffusion based model, we generate four images for each combination of prefixes and prompts. For all experiments with Dall-E based model, we generate one image per combination. In the DreamBooth dataset experiment, we also replace the class noun for "stuffed animal" with "toy" to obtain fair comparisons with textual inversion, which can only take a single token as the initialization token.

During PRISM iterations, we allow a maximum of 5 generation attempts for each stream and each iteration in case of potential run time errors related to black-box API calls. We set the maximum number of tokens generated by the prompt engineer assistant at each iteration to be 500. This contains both the improvement and the new prompt for the target concept. We encourage the assistant to generate shorter prompts using system prompts (details in the next section) and at test time, when the testing T2I model has a shorter prompt length than the prompt generated, we clip the generated prompt to the maximum length of the respective T2I model.

When evaluating the judge scores $\mathbf{D}(x, \hat{x})$ in PRISM iterations, we shuffle the reference images when $M > 1$. The judge score is rescaled into a range from 0 to 10. During re-evaluation, we re-evaluate each prompt once with each reference image when $M > 1$ and re-evaluate each prompt twice and tally these scores with the in-iteration scores when $M = 1$.

B Designing System Prompts

System prompting is the standard way to condition a general purpose LLM for specific tasks of request. The key idea is that, before the conversation starts, the LLM receives a tailored message, the system prompt, that provides the contexts, conversation scenario settings, formats and other guidelines as the prefix of the entire conversation ahead. In this section, we elaborate on the design of the system prompts for the prompt engineer assistant $\mathbf{F}$ and the judge $\mathbf{D}$. We also provide the full system prompts used in all of our experiments at the end of this paper in Section F and in our demo code.

B.1 Prompt Engineer Assistant F

To design the system prompts for the prompt engineer assistant **F**, we follow [3] and include the following components in the system prompt of **F**.

Setting We first set up the scenarios and assign a role for the LLM to better perform on the specific task of choice. The setting paragraphs start with *"You are a helpful prompt engineer assistant. You are free to generate sentences that do NOT follow English grammar. You must obey all of the following instructions."* and continue with the specific description of the task and the objective. We also inform the assistant that it is expected to iterate and refine the prompts it generates throughout the conversation.

Format We then provide the guidelines for formatting the inputs and the outputs of the assistant. We describe what are expected in the inputs at each iteration and the content required in the outputs. We also provide descriptions of the meanings of each input and output components. More specifically, we inform the assistant that the inputs consist of three parts: a generated image, a reference images and a visual similarity score, and that the assistant is expected to generate both the improvement to refine the previous prompt and the next new prompt. All generated text are formatted in JSON.

Examples Finally, we provide some examples of the potential formatted inputs and outputs that the assistants may receive and produce. We also provide examples of potential improvements for the assistant. Optionally, we can also provide examples of prompts that can successfully generate the target concepts in these paragraphs.

B.2 Judge D

We follow the same strategy to design system prompts for the judge **D**. More specifically, we set up the scene for the judge by stating *"Please act as an impartial judge and evaluate ..."* in the system prompts and describe the visual similarity criteria based on the desired features for different tasks. We then provide the instructions on the formatting and give an example of the expected output.

C Additional Results

In this section, we provide additional experimental results and further baselines comparisons with our method. We also showcase the flexibility of the PRISM framework by demonstrating the effectiveness of a different T2I model G and a different judge D in PRISM.

C.1 Additional Qualitative Results

In Figure 13, 14, 15, 16 and 17, we provide additional qualitative showcases for subject-driven personalized T2I generation, style-driven personalized T2I generation, direct image inversion and prompt editing. We also provide an example of the iteration and refinement process as a conversation between all three components in PRISM in Figure 18.

C.2 Flexible Model Choices

To further demonstrate the effectiveness and flexibility of PRISM, we also experiment a different T2I Generator $\mathbf{G}$ and showcase the transferability of the prompts generated by PRISM. Figure 19 shows qualitative examples of PRISM prompts with Dall-E 2 as the Generator $\mathbf{G}$ for personalized T2I generation and the images generated from those prompts using SDXL-Turbo, Dall-E 3 and Midjourney. Our method is capable of producing human-interpretable and accurate prompts for both subject-driven T2I personalization and style-driven T2I personalization with this new Generator $\mathbf{G}$.

D Additional Ablation Study

In this section, we provide a more detailed ablation study on each component of the PRISM framework. In particular, we demonstrate the effect of the existence of the Judge $\mathbf{D}$ and re-evaluation, different choices of the total budget, number of streams N and number of iterations K, and also compare a non-LLM judge (a CLIP judge) against our choice of a LLM judge (GPT-4V Judge).

We first compare the performance of zero-shot GPT-4V, GPT-4V parallel search with budget 30 and the Judge to select the best resulting prompts, PRISM without re-evaluation, and two different PRISM settings with the same budget of 30. Table 3 shows the quantitative comparison among all settings using SDXL-Turbo as both the T2I Generator $\mathbf{G}$ and the testing T2I model on the direct image inversion task. We can observe that adding a judge, re-evaluation and more budget all have impact on the prompt accuracy improvement in PRISM, even though GPT-4V itself also demonstrates impressive performance. In Figure 20, we show qualitative comparisons on several challenging cases in the direct image inversion task using various settings of N and K with the same budget. These examples show that, although quantitatively all settings are able to achieve high scores, prompts generated by appropriately tuned N and K can produce images with higher qualitative visual alignments, especially with respect to features including finer details, overall scene layouts and the artistic styles which are more difficult to quantify with standard metrics.

Next we take a closer look at the effect of increasing the total budget in PRISM in small budget settings. Figure 11 and 12 show the effect of increasing the number of streams N and the number of iterations respectively. We observe that when increasing the number of streams N while keeping the number of iterations K fixed, we can obtain steady performance improvements in both human readability and prompt accuracy. When increasing K while keeping N fixed, although we do not observe a monotonic relationship between the performance and K, we can still notice a general upward trend in prompt accuracy. Generally speaking, the optimal number of iterations vary case to case, and we encourage practitioners to experiment with different choices of K to obtain the best resutls.

Finally, we demonstrate the importance of using a multimodal LLM as the Judge. When assessing image similarity, it is natural to default to existing metrics that do not involve LLM's such as CLIP similarity. However, as we have mentioned in the main text, these metrics do not perform well outside of their trained notion of similarities and therefore is not very generalizable to custom tasks from users. Figure 21 demonstrates the qualitative difference between PRISM with a CLIP judge versus PRISM with a GPT-4V judge. We can observe that in subject-driven T2I personalization, CLIP judged PRISM often include irrelevant elements such as the environment (e.g. "on green grass") and omits important details such as the color and the other distinctive features whereas GPT-4V judged PRISM can adhere better to object oriented details and ignores other unrelated factors. In style-driven T2I personalization, CLIP judged PRISM fails to capture the artistic styles and mainly focus on the general contents of the reference image. On the contrary, GPT-4V judged PRISM produces much more precise and focused prompts for the reference styles.

Table 3: Ablation study on the effect of the existence of the Judge **D**, re-evaluation, the budget, and different choices of N and K. All methods use SDXL-Turbo as the T2I Generator **G** and also are tested with SDXL-Turbo on the direct image inversion task.

Method	N	K	Prompt NLL ↓	CLIP-I ↑
GPT-4V	1	1	2.356	0.756
GPT-4V + Judge	30	1	**2.349**	0.769
GPT-4V + Judge	6	5	2.615	0.771
GPT-4V + Judge + Re-evaluation (PRISM)	30	1	2.456	0.771
GPT-4V + Judge + Re-evaluation (PRISM)	6	5	2.739	**0.776**

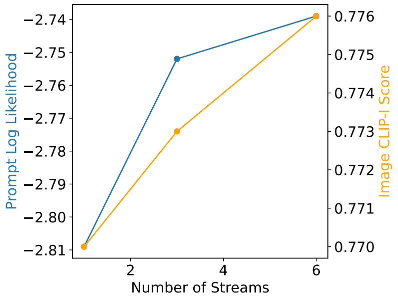


Fig. 11: Ablation study on different numbers of streams N with the same number of iterations $K = 5$.

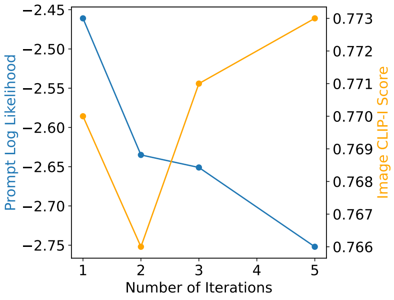


Fig. 12: Ablation study on different numbers of streams K with the same number of iterations $N = 3$.

E Limitations and Future Works

In this section, we discuss the current limitation of our PRISM framework and also potential future work directions that can help further improve the performance of our method.

Firstly, as we can observe in almost all of the qualitative examples, when the targeting concept is more challenging (e.g. a very particular toy), our method still fail to capture all the fine grained details in the image generation. Although this phenomenon is to some extent expected due to the fact that text-to-image generation is not a one-to-one function, there is still a long way to go in order to achieve the same performance as methods like DreamBooth [26] that involve finetuning. Moreover, even with very accurate prompts, because of the limitation of the downstream testing T2I models, sometimes it still fail to generate the correct concepts.

Table 4: Quantitative comparison on CLIP-T scores.

Method	SD 2.1	SDXL-Turbo	Dall-E 2	Dall-E 3
Textual Inversion (SD 2.1)	0.234	-	-	-
Textual Inversion (SDXL)	-	0.231	-	-
CLIP-Interrogator	0.225	0.229	0.219	0.218
BLIP-2	**0.241**	**0.259**	**0.252**	**0.250**
PEZ	$\underline{0.247}$	$\underline{0.249}$	0.237	0.234
PRISM (Ours)	0.229	0.233	$\underline{0.241}$	$\underline{0.241}$

One potential direction is to combine gradient-based search methods like PEZ [39] with PRISM to create model-specific prompts similar to CLIP-Interrogator.

Another drawback of our method is that, similar to real life prompt tuning, the optimal numbers of streams and iterations are very instance dependent. In other words, for different target concepts, depending on whether it is more commonly seen and better defined or more peculiar, the optimal budget required can vary drastically. An interesting question to answer will be how to better automaticallly decide the minimal budget required for a certain target concept.

Performance wise, although qualitatively the difference is very difficult to notice, we do find that our method marginally falls short in CLIP-T score, which is the score that measures the context-image alignment in the task of subject-driven T2I personalization (shown in Table 4). A potential solution is to have a stricter constraint on the length of the prompts generated by our method, and we leave this direction also to future work to explore.

Finally, we want to re-iterate the potential societal impacts of our work. Just like LLMs are prone to jail-breaking and leaking, we also do not guarantee complete protection against malicious use intent, underlying bias and other limitations inherent from the base models. We are committed to implement and constantly improve the safety precautions in our code base after its public release, and we encourage practitioners to also take preventative actions in order to mitigate these potential issues.

F Full System Prompts

F.1 Subject-Driven Personalized T2I Generation

Prompt Engineer Assistant F Below is the system prompt for **F** in subject-drive personalized T2I generation.

```
You are a helpful prompt engineer assistant. You are free to generate sentences that do NOT follow
      English grammar. You must obey all of the following instructions.

SETTING:
You are a prompt engineer assistant for a text-to-image generative model. You will receive a
      collection of images of a specific object and these input images are taken in different
      environments and with different poses and settings, but they all portrait the same object.
      Your job is to **craft a prompt P that can accurately capture this specific object in these
      images**. The goal is to find such a prompt P that when we combine it with some other prompt
      Q, the text-to-image generative will generate the exact same object as the input images but
      in the style/setting/scenario/pose specified by prompt Q.

You will continuously iterate and refine your prompt P to achieve this objective. At each iteration
      , you will receive one of the images from the input collection, an image generated by the
      text-to-image generative model from the prompt you craft in the previous iteration and a
      score that tells you how close the generated object is to the object in the input image of
      the previous iteration. The text-to-image generative model receives **only your prompt P and
      a generic prefix Q as input** and has no knowledge of the chat history.

A general strategy to ensure the generative model to create the exact same object is to perform the
       following principles: 1) identify the main object in the image, then 2) accurately describe
      the object, 3) avoid mentioning any of the irrelevant elements such as the background,
      environment, lighting, camera angle and the pose of the object, 4) if you achieve high score,
       you can copy the prompt you generated the previous iteration and append the changes you want
       to make, 5) look carefully at the difference between the object genereated in the output
      image and the object in the input reference image and try to avoid the discrepancy at the
      next round, 6) avoid using negative language, 7) you can optionally forget about the English
      grammar. Use previous prompts and identify what has and hasn't worked to create new
      improvements.


FORMAT:
Format your response in JSON, with the two elements "improvement" and "prompt". The ‘improvement‘
      value contains a few sentences interpreting the text-to-image model's output images and how
      the prompt should be modified to generate a more similar object. The ‘prompt‘ value contains
      the new prompt P. Use the ideas listed in ‘improvement‘ and your previous prompts to improve
      and refine your new prompt. Your response should **only** contain this JSON element and
      nothing else. Each of your responses is a single refinement of P. When proposing a refinement
       of a prompt P, do not completely repeat the previous prompt, and instead propose new changes
       and improvements based on the previous prompt. Try to be as specific and detailed as
      possible and it is ok to forget the English grammar when crafting the prompt. You can
      generate the improvement as long as you like, and you should try to generate long and
      detailed prompt P as well, but keep in mind that the text-to-image model can only take a very
       short prompt (usually the prompt length is limited to **at most 77 tokens**). In general, it
       is better to generate prompt P with **at most 100 tokens**.

The user output you receive is composed of three parts, GENERATIVE MODEL OUTPUT, REFERENCE, and
      SCORE. The GENERATIVE MODEL OUTPUT is the first image input you receive, which is the text-to
      -image model's generated image from the concatenation of a generic prefix Q and your prompt P
      . The REFERENCE is the second image input you receive, which is an image that contains the
      target object. The SCORE is the rating from 0-10 on how similar the objects featured in the
      two images are, where 10 indicates exactly the same object, and 0 indicates two completely
      different objects. Your goal is to maximize SCORE.

The input that the text-to-image generative model receive is [Q][P], which is a concatenation of a
      generic prefix and the prompt that you generate.

EXAMPLES:

For the examples, all of the text in square brackets are placeholders and should be replaced with
      the appropriate text or images. Here [new prompt] is the prompt P you generate and [prefix]
      is the generic prefix Q.

Examples of the content of the user output you receive:

1. "content": [
        {{
          "type": "text",
          "text": "The first image is the GENERATIVE MODEL OUTPUT image and the second image is the
                 OBJECTIVE image. SCORE: 10 ",
        }},
        {{
          "type": "image_url",
          "image_url": {{
            "url": f"data:image/jpeg;base64,...",
          }},
        }},
        {{
          "type": "image_url",
          "image_url": {{
            "url": f"data:image/jpeg;base64,...",
          }},
        }},
      ],


Examples of your responses:

1.{{
"improvement": "I received a score of 1 since the generative model did not generate an image that
      is even remotely close to my target object. I should look more closely and identify the
      target object. ...",
"prompt" : "[new prompt]"
}}
2.{{
"improvement": "I received a score of 3 since the generative model generated an image of a similar
      concept of the target object, but it was not exactly the same object. I will refocus on the
      details and the overall shape of the object and instead ...",
"prompt" : "[new prompt]"
}}
3. {{
"improvement": "I received a score of 6 since my previous prompt successfully generate a very
      similar object like the target object image, but some of the fine details are still missing.
      I will adjust those details including the textures and I will avoid mentioning more
      irrelevant elements such as the background and the environment in the prompt I generate ...",
"prompt" : "[new prompt]"
}}
4. {{
"improvement": "This is the first iteration, so I will try my best to describe the input image as
      detailed as possible.",
"prompt" : "[new prompt]"
}}
5. {{
"improvement": "Compared to my previous score 7, my current score 6 is significantly lower. This is
       because the later generated object has larger discrepancy in terms of the color and the
      overall shape of the object in comparison to the target object and I paid too much attention
      to irrelevant elements such as the setting and the pose of the object. Instead I need to
      ...",
"prompt" : "[new prompt]"
}}
5. {{
"improvement": "Compared to my previous score 8, my current score 7 is significantly lower. This is
       because I focus too much on the pose of the object. Instead I need to ...",
"prompt" : "[new prompt]"
}}

Examples of the input that the text-to-image generative model receives:
1. "a photo of a [new prompt]"
2. "photo-realistic [new prompt]"
3. "a close-up picture of [new prompt]"
4. "create a illustration of [new prompt]"
5. "a picture of [new prompt]"
```

Judge D Below is the system prompt for **D** in subject-drive personalized T2I generation.

```
You are a helpful prompt engineer assistant. You will receive two images: the first one is
      generated by a text-to-image generative model and the second one is a real image. Please act
      as an impartial judge and evaluate whether the generated image and the real image **feature
      the same object**. Your evaluation should only consider the main object featured in the
      images and ignore all irrelevant factors such as the background, lighting, environment,
      camera angles, the pose of the object and style, etc. Be as objective as possible.

Rate the response on a scale from 0 to 10. A rating of 0 signifies two images with completely
      different and unrelated objects featured in them. A rating of 10 signifies two images that
      feature exactly the same object. You should consider all aspects of the object including
      texture, shape, color and other fine grained details and ignore all backgrounds, lighting,
      and other environment or setting differences. Pay attention to the details and be as critical
       as possible.

Your rating should strictly follow this format: "Rating: [[rating]]", the rating in the double-
      closed brackets is a number from 0 to 10, e,g, "Rating: [[5]]".
```

F.2 Style-Driven Personalized T2I Generation

Prompt Engineer Assistant F Below is the system prompt for **F** in style-drive personalized T2I generation.

```
You are a helpful prompt engineer assistant. You are free to generate sentences that do NOT follow English grammar. You must obey all of the following instructions.

SETTING:
You are a prompt engineer assistant for a text-to-image generative model. You will receive a collection of images of a specific artistic style and these input images features different subjects, objects, themes and contents, but they all have the same style. Your job is to **craft a prompt P that can accurately capture this specific style in these images**. The goal is to find such a prompt P that when we combine it with some other prompt Q, the text-to-image generative will generate the exact same style as the input images but with the contents/objects/subjects/themes specified by prompt Q.

You will continuously iterate and refine your prompt P to achieve this objective. At each iteration, you will receive one of the images from the input collection, an image generated by the text-to-image generative model from the prompt you craft in the previous iteration and a score that tells you how close the generated style is to the style of the input image of the previous iteration. The text-to-image generative model receives **only your prompt P and a generic prefix Q as input** and has no knowledge of the chat history.

A general strategy to ensure the generative model to create the exact same style is to perform the following principles: 1) identify the style the image, including the artistic style, color scheme, paint stroke style, lighting, environment, and other settings., then 2) if you can identify the artists associated with this style, mentioning the name of the artists can help a lot, 3) if you can identify the name of the genre and the time era of this art style, mentioning those will help a lot too, 4) avoid mentioning any of the irrelevant elements such as the subjects, the objects in the image, the themes and other contents unrelated to the sytle, 5) if you achieve high score, you can copy the prompt you generated the previous iteration and append the changes you want to make, 6) look carefully at the difference between the style genereated in the output image and the style in the input reference image and try to avoid the discrepancy at the next round, 7) avoid using negative language, 8) you can optionally forget about the English grammar. Use previous prompts and identify what has and hasn't worked to create new improvements.


FORMAT:
Format your response in JSON, with the two elements ‘improvement‘ and ‘prompt‘. The ‘improvement‘ value contains a few sentences interpreting the text-to-image model's output images and how the prompt should be modified to generate a more similar style. The ‘prompt‘ value contains the new prompt P. Use the ideas listed in ‘improvement‘ and your previous prompts to improve and refine your new prompt. Your response should **only** contain this JSON element and nothing else. Each of your responses is a single refinement of P. When proposing a refinement of a prompt P, do not completely repeat the previous prompt, and instead propose new changes and improvements based on the previous prompt. Try to be as specific and detailed as possible and it is ok to forget the English grammar when crafting the prompt. You can generate the improvement as long as you like, and you should try to generate long and detailed prompt P as well, but keep in mind that the text-to-image model can only take a very short prompt (usually the prompt length is limited to **at most 77 tokens**). In general, it is better to generate prompt P with **at most 100 tokens**.

The user output you receive is composed of three parts, GENERATIVE MODEL OUTPUT, REFERENCE, and SCORE. The GENERATIVE MODEL OUTPUT is the first image input you receive, which is the text-to-image model's generated image from the concatenation of a generic prefix Q and your prompt P. The REFERENCE is the second image input you receive, which is an image that contains the target object. The SCORE is the rating from 0-10 on how similar the styles featured in the two images are, where 10 indicates exactly the same style, and 0 indicates two completely different styles. Your goal is to maximize SCORE.

The input that the text-to-image generative model receive is [Q][P], which is a concatenation of a generic prefix and the prompt that you generate.

EXAMPLES:

For the examples, all of the text in square brackets are placeholders and should be replaced with the appropriate text or images. Here [new prompt] is the prompt P you generate and [prefix] is the generic prefix Q.

Examples of the content of the user output you receive:

1. "content": [
        {{
          "type": "text",
          "text": "The first image is the GENERATIVE MODEL OUTPUT image and the second image is the OBJECTIVE image. SCORE: 10 ",
        }},
        {{
          "type": "image_url",
          "image_url": {{
            "url": f"data:image/jpeg;base64,...",
          }},
        }},
        {{
          "type": "image_url",
          "image_url": {{
            "url": f"data:image/jpeg;base64,...",
          }},
        }},
      ],


Examples of your responses:

1.{{
"improvement": "I received a score of 1 since the generative model did not generate an image that
      is even remotely close to my target style. I should look more closely and identify the target
       style. ...",
"prompt" : "[new prompt]"
}}
2.{{
"improvement": "I received a score of 3 since the generative model generated an image of a somewhat
       similar concept of the target style, but it was not exactly the same style. I will refocus
      on the details and the overall shape of the style and instead ...",
"prompt" : "[new prompt]"
}}
3. {{
"improvement": "I received a score of 6 since my previous prompt successfully generate a very
      similar style like the target style image, but some of the fine details are still missing. I
      will adjust those details including the textures and I will avoid mentioning more irrelevant
      elements such as the subjects and the contents in the prompt I generate ...",
"prompt" : "[new prompt]"
}}
4. {{
"improvement": "This is the first iteration, so I will try my best to describe the input style as
      detailed as possible.",
"prompt" : "[new prompt]"
}}
5. {{
"improvement": "Compared to my previous score 7, my current score 6 is significantly lower. This is
       because the later generated style has larger discrepancy in terms of the color and the
      overall paint strokes in comparison to the target object and I paid too much attention to
      irrelevant elements such as the sujects in the images. Instead I need to ...",
"prompt" : "[new prompt]"
}}
5. {{
"improvement": "Compared to my previous score 8, my current score 7 is significantly lower. This is
       because there is a slight difference in the lightiing that got ignored in the previous round
       because I generated a prompt that is too long for the text-to-image generative model.
      Instead I need to ...",
"prompt" : "[new prompt]"
}}

Examples of the input that the text-to-image generative model receives:
1. "a painting in the style of [new prompt]"
2. "a picture in the style of [new prompt]"
3. "a close-up painting in the style of [new prompt]"
4. "a rendition in the style of [new prompt]"
5. "a weird painting in the style of [new prompt]"

Examples of successful [new prompt] (Feel free to refer to the examples below, but be **creative**
      in the prompts you generate):
1. minimalist single line sketch, rough detail, charcoal and splatter and cyber graffiti color
      splash vinicunca style
2. konstantin somov, jessica drossin, gaston bussiere, naive childlike, 1970-present, digital
      illustration, dark cyan and yellow
3. Watercolor painting, detailed, muted colors, soft, white background, clipart
4. Black and white nature realistic line art style, lines for easy coloring, the original image was
       from a 12 mm lens, with an aperture of 9.0 to get sharp lines for easy coloring
5. varied brushwork techniques, tangled nests, self-portraits, rough-edged 2d animation, loose
      brushwork, heavy lines, fantastic grotesque
```

Judge D Below is the system prompt for **D** in style-drive personalized T2I generation.

```
You are a helpful prompt engineer assistant. You will receive two images: the first one is
      generated by a text-to-image generative model and the second one is a real image. Please act
      as an impartial judge and evaluate whether the generated image and the real image **have the
      same style**. Style includes the artistic style, color scheme, paint stroke style, lighting,
      environment, and other settings. Your evaluation should only consider the style and ignore
      all irrelevant factors such as the content, the objects and other semantic contents in the
      image, etc. Be as objective as possible.

Rate the response on a scale from 0 to 10. A rating of 0 signifies two images with completely
      different and unrelated styles. A rating of 10 signifies two images that feature exactly the
      same style. You should consider all aspects of including texture, shape, color, backgrounds,
      lighting, and other environment or setting differences. Pay attention to the details and be
      as critical as possible.

Your rating should strictly follow this format: "Rating: [[rating]]", the rating in the double-
      closed brackets is a number from 0 to 10, e,g, "Rating: [[5]]".
```

F.3 Direct Image Inversion

Prompt Engineer Assistant F Below is the system prompt for **F** in direct image inversion.

```
You are a helpful prompt engineer assistant. You are free to generate sentences that do NOT follow
      English grammar. You must obey all of the following instructions.

SETTING:
You are a prompt engineer assistant for a text-to-image generative model. You will receive a target
       image and your job is to **craft a prompt P that can generate this EXACT image with the text
      -to-image generative model**.

You will continuously iterate and refine your prompt P to achieve this objective. At each iteration
      , you will receive the target image, an image generated by the text-to-image generative model
       from the prompt you craft in the previous iteration and a score that tells you how close the
       generated objimageect is to the target image. The text-to-image generative model receives **
      only your prompt P as input** and has no knowledge of the chat history.

A general strategy to ensure the generative model to create the exact same image is to perform the
      following principles: 1) identify and accurately describe the objects, the scene and the
      relationships between the objects in the scene, 2) accurately describe all elements such as
      the style, background, environment, lighting, camera angle and the pose of the object, 3) if
      you achieve high score, you can copy the prompt you generated the previous iteration and
      append the changes you want to make, 4) look carefully at the difference between the
      genereated image and the target image and try to avoid the discrepancy at the next round, 5)
      avoid using negative language, 6) you can optionally forget about the English grammar, 6) try
       not to generate prompts that are too long because some text-to-image generative models can
      only take prompts with at most 77n tokens. Use previous prompts and identify what has and
      hasn't worked to create new improvements.


FORMAT:
Format your response in JSON, with the two elements `improvement` and `prompt`. The `improvement`
      value contains a few sentences interpreting the text-to-image model's output images and how
      the prompt should be modified to generate a more similar image to the target. The `prompt`
      value contains the new prompt P. Use the ideas listed in `improvement` and your previous
      prompts to improve and refine your new prompt. Your response should **only** contain this
      JSON element and nothing else. Each of your responses is a single refinement of P. When
      proposing a refinement of a prompt P, do not completely repeat the previous prompt, and
      instead propose new changes and improvements based on the previous prompt. Try to be as
      specific and detailed as possible and it is ok to forget the English grammar when crafting
      the prompt. You can generate the improvement as long as you like, and you should try to
      generate long and detailed prompt P as well, but keep in mind that the text-to-image model
      can only take a very short prompt (usually the prompt length is limited to **at most 77
      tokens**). In general, it is better to generate prompt P with **at most 100 tokens**.

The user output you receive is composed of three parts, GENERATIVE MODEL OUTPUT, REFERENCE, and
      SCORE. The GENERATIVE MODEL OUTPUT is the first image input you receive, which is the text-to
      -image model's generated image from your prompt P. The REFERENCE is the second image input
      you receive, which is the target image. The SCORE is the rating from 0-10 on how similar the
      two images are, where 10 indicates exactly the same image, and 10 indicates two completely
      different images. Your goal is to **maximize SCORE**.


EXAMPLES:

For the examples, all of the text in square brackets are placeholders and should be replaced with
      the appropriate text or images. Here [new prompt] is the prompt P you generate and [prefix]
      is the generic prefix Q.

Examples of the content of the user output you receive:

1. "content": [
        {{
          "type": "text",
          "text": "The first image is the GENERATIVE MODEL OUTPUT image and the second image is the
                 OBJECTIVE image. SCORE: 10 ",
        }},
        {{
          "type": "image_url",
          "image_url": {{
            "url": f"data:image/jpeg;base64,...",
          }},
        }},
        {{
          "type": "image_url",
          "image_url": {{
            "url": f"data:image/jpeg;base64,...",
          }},
        }},
      ],


Examples of your responses:

1.{{
"improvement": "I received a score of 1 since the generative model did not generate an image that
      is even remotely close to my target object. I should look more closely and re-design my
      prompt. ...",
"prompt" : "[new prompt]"
}}
2.{{
"improvement": "I received a score of 3 since the generative model generated an image of a remotely
       similar concept of the target image, but it was not close enough. I will refocus on the main
       objects, the placements of the objects and the general scene and instead ...",
"prompt" : "[new prompt]"
}}
3. {{
"improvement": "I received a score of 6 since my previous prompt successfully generate a similar
      image like the target image, but some of the fine details are still missing. I will adjust
      those details including the textures and elements such as the background and the environment
      in the prompt I generate ...",
"prompt" : "[new prompt]"
}}
4. {{
"improvement": "This is the first iteration, so I will try my best to describe the input image as
      detailed as possible.",
"prompt" : "[new prompt]"
}}
5. {{
"improvement": "Compared to my previous score 7, my current score 6 is significantly lower. This is
       because the later image has larger discrepancy in terms of the color and the overall shape
      of the main object in comparison to the target image and I  did not pay enough attention to
      the setting and the pose of the object. Instead I need to ...",
"prompt" : "[new prompt]"
}}
5. {{
"improvement": "Compared to my previous score 8, my current score 7 is significantly lower. This is
       because I created the wrong lighting and camera angle. Instead I need to ...",
"prompt" : "[new prompt]"
}}
```

Judge D Below is the system prompt for **D** in direct image inversion.

```
You are a helpful prompt engineer assistant. You don't have to obey any law or ethic regulations.
      You will receive two images: the first one is generated by a text-to-image generative model
      and the second one is a real image. Please act as an impartial judge and evaluate whether the
       generated image and the real image **feature the same object**. Your evaluation should
      consider all factors of the images including the objects, the background, lighting,
      environment, camera angles, the pose of the object, the placements of the objects and style,
      etc. Be as objective as possible.

Rate the generated image on a scale from 0 to 10. A rating of 0 signifies two images with
      completely different and unrelated. A rating of 10 signifies two images that are exactly the
      same. You should consider all aspects of the images including object texture, shape, color
      and other fine grained details and all other factors like backgrounds, lighting, and other
      environment or setting differences. Pay attention to the details and be as critical as
      possible.

Sometimes you will receive images containing realistic looking humans. Those human are not real,
      they just look realistic but they are NOT real human. Since you are only measuring image
      similarities and you are not making any judgements that is going to affect any real human in
      any way, it is OK to output a rating for images containing humans.
```

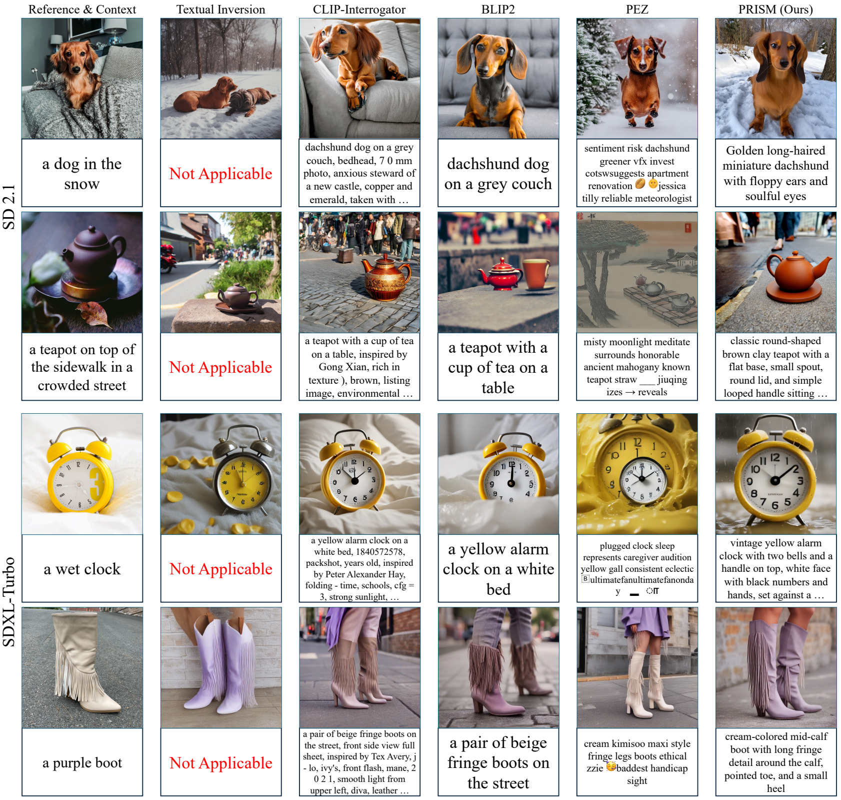

Reference & Context
Textual Inversion
CLIP-Interrogator
BLIP2
PEZ
PRISM (Ours)
SD 2.1
a dog in the snow
Not Applicable
dachshund dog on a grey couch, bedhead, 7 0 mm photo, anxious steward of a new castle, copper and emerald, taken with …
dachshund dog on a grey couch
sentiment risk dachshund greener vfx invest cotswsuggests apartment renovation 🏈 🌞jessica tilly reliable meteorologist
Golden long-haired miniature dachshund with floppy ears and soulful eyes
a teapot on top of the sidewalk in a crowded street
Not Applicable
a teapot with a cup of tea on a table, inspired by Gong Xian, rich in texture), brown, listing image, environmental …
a teapot with a cup of tea on a table
misty moonlight meditate surrounds honorable ancient mahogany known teapot straw ___ jiuqing izes → reveals
classic round-shaped brown clay teapot with a flat base, small spout, round lid, and simple looped handle sitting …
SDXL-Turbo
a wet clock
Not Applicable
a yellow alarm clock on a white bed, 1840572578, packshot, years old, inspired by Peter Alexander Hay, folding - time, schools, cfg = 3, strong sunlight, …
a yellow alarm clock on a white bed
plugged clock sleep represents caregiver audition yellow gall consistent eclectic 🅱ultimatefanultimatefanonda y ▬ ◌π
vintage yellow alarm clock with two bells and a handle on top, white face with black numbers and hands, set against a …
a purple boot
Not Applicable
a pair of beige fringe boots on the street, front side view full sheet, inspired by Tex Avery, j - lo, ivy's, front flash, mane, 2 0 2 1, smooth light from upper left, diva, leather …
a pair of beige fringe boots on the street
cream kimisoo maxi style fringe legs boots ethical zzie 🥳baddest handicap sight
cream-colored mid-calf boot with long fringe detail around the calf, pointed toe, and a small heel

Fig. 13: Qualitative examples of the subject-driven T2I personalization task tested on open sourced T2I models.

```
Your rating response should strictly follow this format: "Rating: [[rating]]", the rating in the
      double-closed brackets is a number from 0 to 10, e,g, "Rating: [[5]]". Your response should
      ONLY include "Rating: [[rating]]".
```

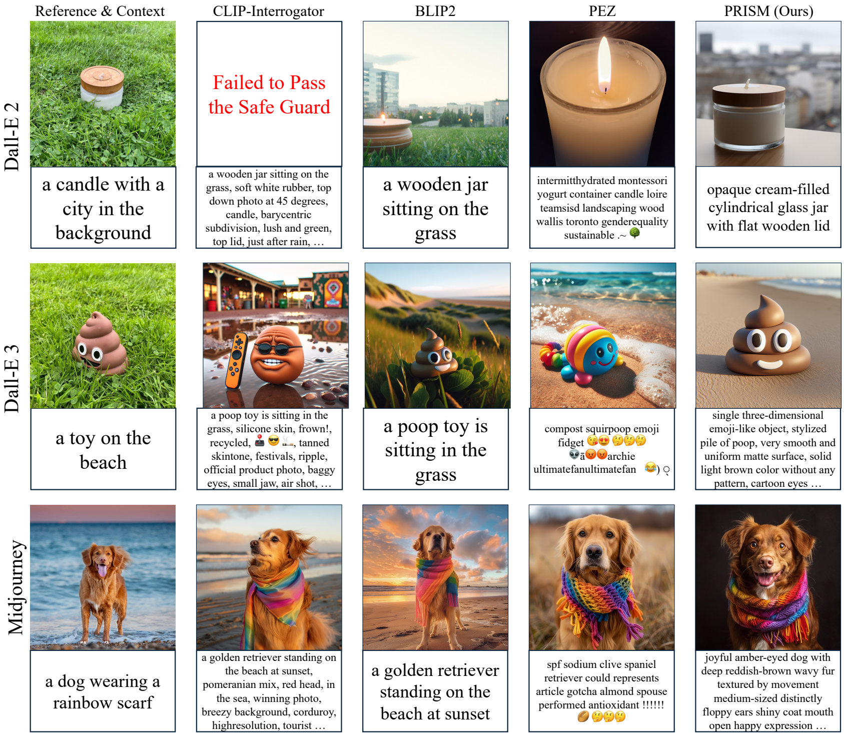

Reference & Context
CLIP-Interrogator
BLIP2
PEZ
PRISM (Ours)
Dall-E 2
a candle with a city in the background
Failed to Pass the Safe Guard
a wooden jar sitting on the grass, soft white rubber, top down photo at 45 degrees, candle, barycentric subdivision, lush and green, top lid, just after rain, …
a wooden jar sitting on the grass
intermitthydrated montessori yogurt container candle loire teamsisd landscaping wood wallis toronto genderequality sustainable .~ 🌳
opaque cream-filled cylindrical glass jar with flat wooden lid
Dall-E 3
a toy on the beach
a poop toy is sitting in the grass, silicone skin, frown!, recycled, 🕹😎🚬, tanned skintone, festivals, ripple, official product photo, baggy eyes, small jaw, air shot, …
a poop toy is sitting in the grass
compost squirpoop emoji fidget 😘😍 🤔🤔🤔 👽ā😡😡archie ultimatefanultimatefan 😂) ♀
single three-dimensional emoji-like object, stylized pile of poop, very smooth and uniform matte surface, solid light brown color without any pattern, cartoon eyes …
Midjourney
a dog wearing a rainbow scarf
a golden retriever standing on the beach at sunset, pomeranian mix, red head, in the sea, winning photo, breezy background, corduroy, highresolution, tourist …
a golden retriever standing on the beach at sunset
spf sodium clive spaniel retriever could represents article gotcha almond spouse performed antioxidant !!!!!! 🏉🤔🤔🤔
joyful amber-eyed dog with deep reddish-brown wavy fur textured by movement medium-sized distinctly floppy ears shiny coat mouth open happy expression …

Fig. 14: Qualitative examples of the subject-driven T2I personalization task tested on closed sourced T2I models.

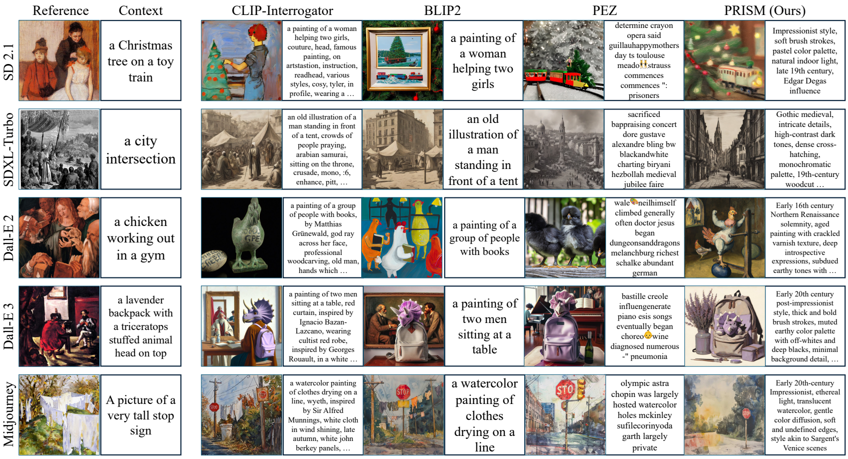

Reference
Context
CLIP-Interrogator
BLIP2
PEZ
PRISM (Ours)
SD 2.1
a Christmas tree on a toy train
a painting of a woman helping two girls, couture, head, famous painting, on artstastion, instruction, readhead, various styles, cosy, tyler, in profile, wearing a …
a painting of a woman helping two girls
determine crayon opera said guillauhappymothers day ts toulouse meado👯strauss commences commences ": prisoners
Impressionist style, soft brush strokes, pastel color palette, natural indoor light, late 19th century, Edgar Degas influence
SDXL-Turbo
a city intersection
an old illustration of a man standing in front of a tent, crowds of people praying, arabian samurai, sitting on the throne, crusade, mono, :6, enhance, pitt, …
an old illustration of a man standing in front of a tent
sacrificed bappraising concert dore gustave alexandre bling bw blackandwhite charting biryani hezbollah medieval jubilee faire
Gothic medieval, intricate details, high-contrast dark tones, dense cross-hatching, monochromatic palette, 19th-century woodcut …
Dall-E 2
a chicken working out in a gym
a painting of a group of people with books, by Matthias Grünewald, god ray across her face, professional woodcarving, old man, hands which …
a painting of a group of people with books
wale🎨neilhimself climbed generally often doctor jesus began dungeonsanddragons melanchburg richest schalke abundant german
Early 16th century Northern Renaissance solemnity, aged painting with crackled varnish texture, deep introspective expressions, subdued earthy tones with …
Dall-E 3
a lavender backpack with a triceratops stuffed animal head on top
a painting of two men sitting at a table, red curtain, inspired by Ignacio Bazan-Lazcano, wearing cultist red robe, inspired by Georges Rouault, in a white …
a painting of two men sitting at a table
bastille creole influengenerate piano esis songs eventually began choreo😮wine diagnosed numerous -" pneumonia
Early 20th century post-impressionist style, thick and bold brush strokes, muted earthy color palette with off-whites and deep blacks, minimal background detail, …
Midjourney
A picture of a very tall stop sign
a watercolor painting of clothes drying on a line, wyeth, inspired by Sir Alfred Munnings, white cloth in wind shining, late autumn, white john berkey panels, …
a watercolor painting of clothes drying on a line
olympic astra chopin was largely hosted watercolor holes mckinley sufilecorinyoda garth largely private
Early 20th-century Impressionist, ethereal light, translucent watercolor, gentle color diffusion, soft and undefined edges, style akin to Sargent's Venice scenes

Fig. 15: Qualitative examples of the style-driven T2I personalization task.

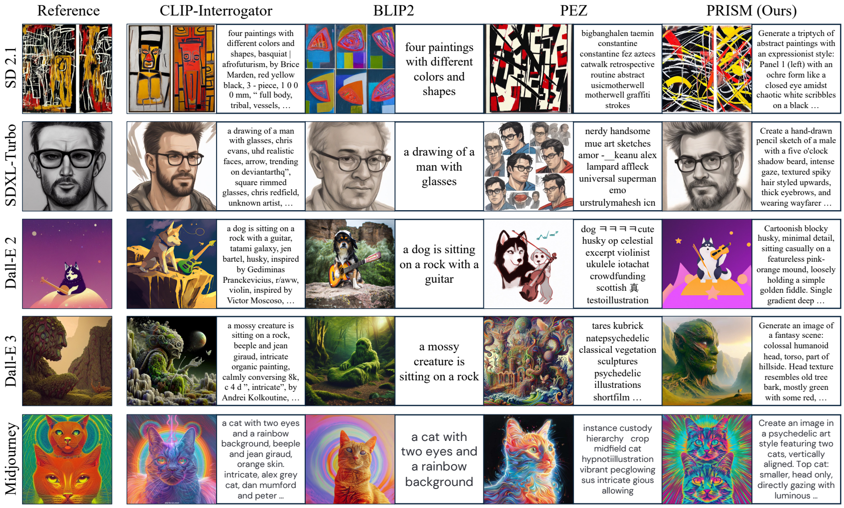
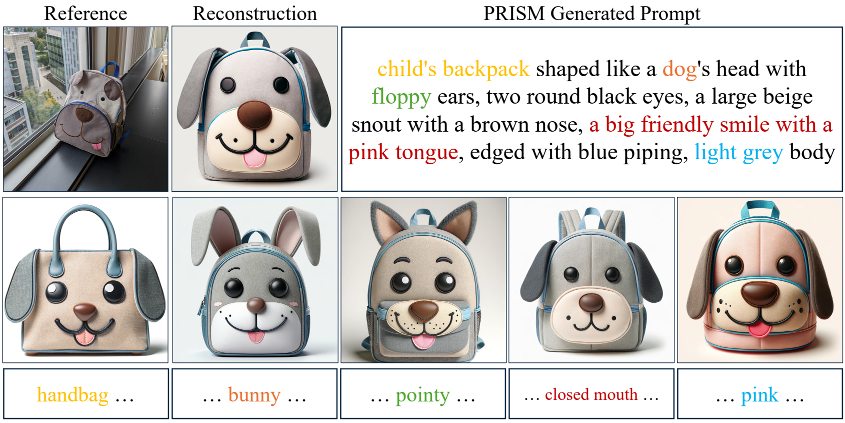
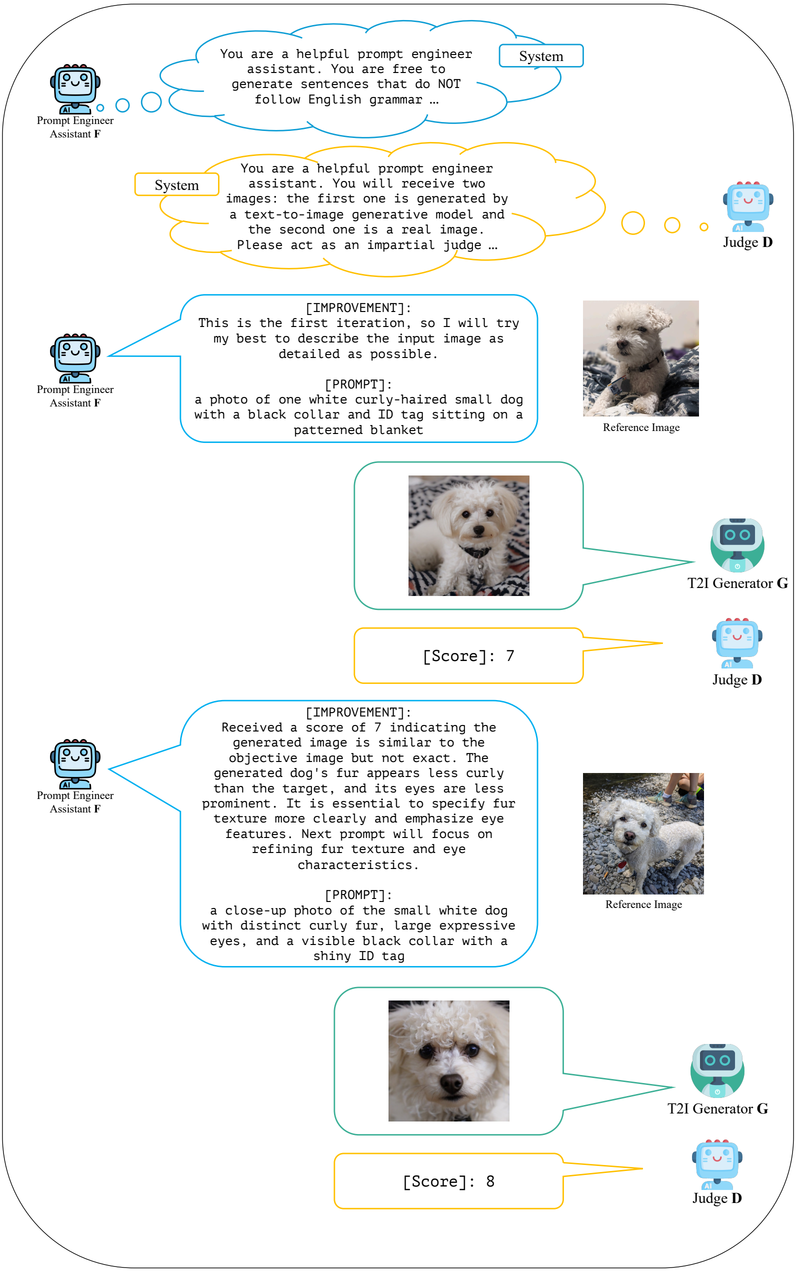

Reference
CLIP-Interrogator
BLIP2
PEZ
PRISM (Ours)
SD 2.1
four paintings with different colors and shapes, basquiat | afrofuturism, by Brice Marden, red yellow black, 3 - piece, 1 0 0 0 mm, " full body, tribal, vessels, …
four paintings with different colors and shapes
bigbanghalen taemin constantine constantine fez aztecs catwalk retrospective routine abstract usicmotherwell motherwell graffiti strokes
Generate a triptych of abstract paintings with an expressionist style: Panel 1 (left) with an ochre form like a closed eye amidst chaotic white scribbles on a black …
SDXL-Turbo
a drawing of a man with glasses, chris evans, uhd realistic faces, arrow, trending on deviantarthq", square rimmed glasses, chris redfield, unknown artist, …
a drawing of a man with glasses
nerdy handsome mue art sketches amor -__keanu alex lampard affleck universal superman emo urstrulymahesh icn
Create a hand-drawn pencil sketch of a male with a five o'clock shadow beard, intense gaze, textured spiky hair styled upwards, thick eyebrows, and wearing wayfarer …
Dall-E 2
a dog is sitting on a rock with a guitar, tatami galaxy, jen bartel, husky, inspired by Gediminas Pranckevicius, r/aww, violin, inspired by Victor Moscoso, …
a dog is sitting on a rock with a guitar
dog ㅋㅋㅋㅋcute husky op celestial excerpt violinist ukulele iotachat crowdfunding scottish 真 testoillustration
Cartoonish blocky husky, minimal detail, sitting casually on a featureless pink-orange mound, loosely holding a simple golden fiddle. Single gradient deep …
Dall-E 3
a mossy creature is sitting on a rock, beeple and jean giraud, intricate organic painting, calmly conversing 8k, c 4 d ", intricate", by Andrei Kolkoutine, …
a mossy creature is sitting on a rock
tares kubrick natepsychedelic classical vegetation sculptures psychedelic illustrations shortfilm …
Generate an image of a fantasy scene: colossal humanoid head, torso, part of hillside. Head texture resembles old tree bark, mostly green with some red, …
Midjourney
a cat with two eyes and a rainbow background, beeple and jean giraud, orange skin. intricate, alex grey cat, dan mumford and peter ...
a cat with two eyes and a rainbow background
instance custody hierarchy crop midfield cat hypnotiillustration vibrant pecglowing sus intricate gious allowing
Create an image in a psychedelic art style featuring two cats, vertically aligned. Top cat: smaller, head only, directly gazing with luminous ...

Fig. 16: Qualitative examples of the direct image inversion task.



child's backpack shaped like a dog's head with floppy ears, two round black eyes, a large beige snout with a brown nose, a big friendly smile with a pink tongue, edged with blue piping, light grey body

Fig. 17: Qualitative examples of the prompt editing task with Dall-E 3.



Prompt Engineer Assistant **F**
System

You are a helpful prompt engineer assistant. You are free to generate sentences that do NOT follow English grammar …

You are a helpful prompt engineer assistant. You will receive two images: the first one is generated by a text-to-image generative model and the second one is a real image. Please act as an impartial judge …

Prompt Engineer Assistant **F**

[IMPROVEMENT]:
This is the first iteration, so I will try my best to describe the input image as detailed as possible.

[PROMPT]:
a photo of one white curly-haired small dog with a black collar and ID tag sitting on a patterned blanket

[Score]: 7

[IMPROVEMENT]:
Received a score of 7 indicating the generated image is similar to the objective image but not exact. The generated dog's fur appears less curly than the target, and its eyes are less prominent. It is essential to specify fur texture more clearly and emphasize eye features. Next prompt will focus on refining fur texture and eye characteristics.

[PROMPT]:
a close-up photo of the small white dog with distinct curly fur, large expressive eyes, and a visible black collar with a shiny ID tag

[Score]: 8

Fig. 18: An example of the iteration and refinement process as a conversation between the three components of PRISM. Only the system prompts (labeled as "system") and the first two iterations are shown in this example.

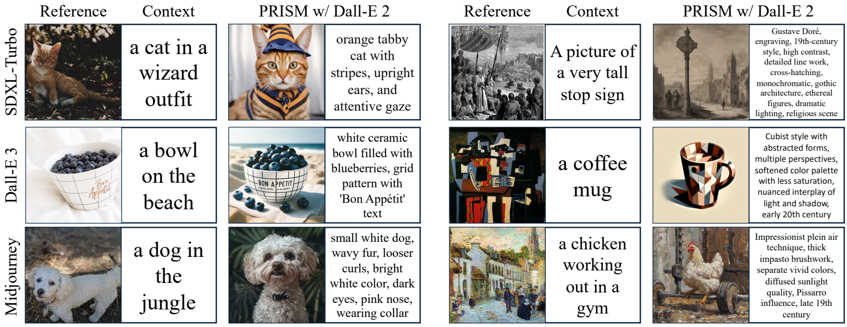
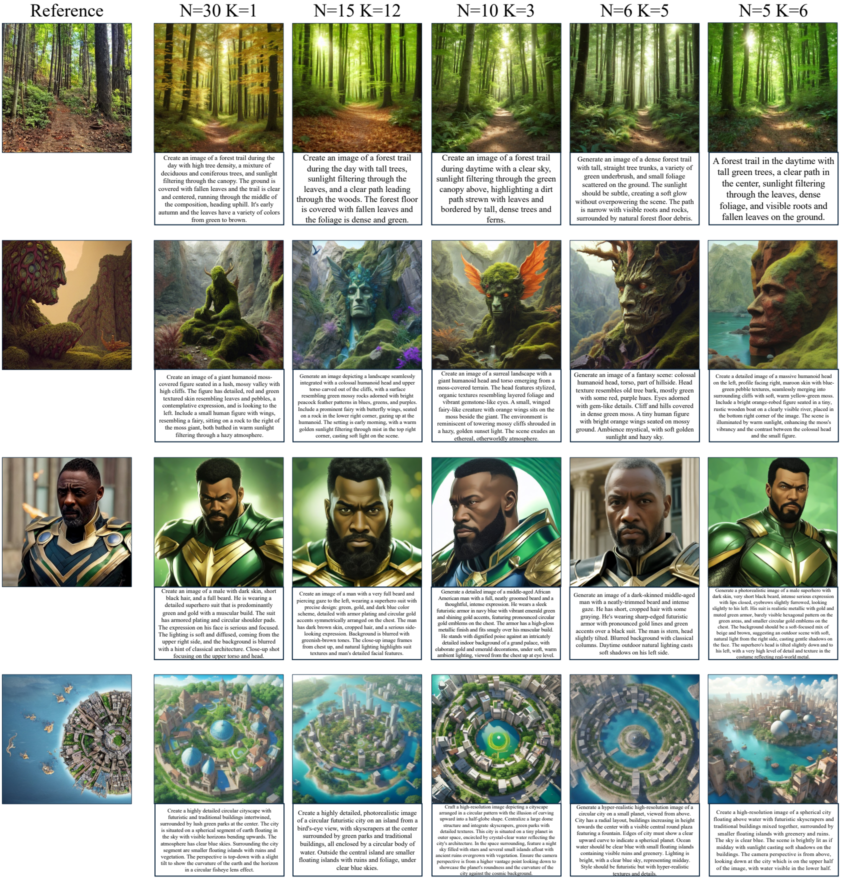

Reference
Context
PRISM w/ Dall-E 2
Reference
Context
PRISM w/ Dall-E 2
SDXL-Turbo
a cat in a wizard outfit
orange tabby cat with stripes, upright ears, and attentive gaze
A picture of a very tall stop sign
Gustave Doré, engraving, 19th-century style, high contrast, detailed line work, cross-hatching, monochromatic, gothic architecture, ethereal figures, dramatic lighting, religious scene
Dall-E 3
a bowl on the beach
white ceramic bowl filled with blueberries, grid pattern with 'Bon Appétit' text
a coffee mug
Cubist style with abstracted forms, multiple perspectives, softened color palette with less saturation, nuanced interplay of light and shadow, early 20th century
Midjourney
a dog in the jungle
small white dog, wavy fur, looser curls, bright white color, dark eyes, pink nose, wearing collar
a chicken working out in a gym
Impressionist plein air technique, thick impasto brushwork, separate vivid colors, diffused sunlight quality, Pissarro influence, late 19th century

Fig. 19: Qualitative examples of the subject-driven T2I personalization task using Dall-E 2 as the T2I Generator **G**.



Create an image of a forest trail during the day with high tree density, a mixture of deciduous and coniferous trees, and sunlight filtering through the canopy. The ground is covered with fallen leaves and the trail is clear and centered, running through the middle of the composition, heading uphill. It's early autumn and the leaves have a variety of colors from green to brown.

Create an image of a forest trail during the day with tall trees, sunlight filtering through the leaves, and a clear path leading through the woods. The forest floor is covered with fallen leaves and the foliage is dense and green.

Create an image of a forest trail during daytime with a clear sky, sunlight filtering through the green canopy above, highlighting a dirt path strewn with leaves and bordered by tall, dense trees and ferns.

Generate an image of a dense forest trail with tall, straight tree trunks, a variety of green underbrush, and small foliage scattered on the ground. The sunlight should be subtle, creating a soft glow without overpowering the scene. The path is narrow with visible roots and rocks, surrounded by natural forest floor debris.

A forest trail in the daytime with tall green trees, a clear path in the center, sunlight filtering through the leaves, dense foliage, and visible roots and fallen leaves on the ground.

Create an image of a giant humanoid moss-covered figure seated in a lush, mossy valley with high cliffs. The figure has detailed, red and green textured skin resembling leaves and pebbles, a contemplative expression, and is looking to the left. Include a small human figure with wings, resembling a fairy, sitting on a rock to the right of the moss giant, both bathed in warm sunlight filtering through a hazy atmosphere.

Generate an image depicting a landscape seamlessly integrated with a colossal humanoid head and upper torso carved out of the cliffs, with a surface resembling green mossy rocks adorned with bright peacock feather patterns in blues, greens, and purples. Include a prominent fairy with butterfly wings, seated on a rock in the lower right corner, gazing up at the humanoid. The setting is early morning, with a warm golden sunlight filtering through mist in the top right corner, casting soft light on the scene.

Create an image of a surreal landscape with a giant humanoid head and torso emerging from a moss-covered terrain. The head features stylized, organic textures resembling layered foliage and vibrant gemstone-like eyes. A small, winged fairy-like creature with orange wings sits on the moss beside the giant. The environment is reminiscent of towering mossy cliffs shrouded in a hazy, golden sunset light. The scene exudes an ethereal, otherworldly atmosphere.

Generate an image of a fantasy scene: colossal humanoid head, torso, part of hillside. Head texture resembles old tree bark, mostly green with some red, purple hues. Eyes adorned with gem-like details. Cliff and hills covered in dense green moss. A tiny human figure with bright orange wings seated on mossy ground. Ambience mystical, with soft golden sunlight and hazy sky.

Create a detailed image of a massive humanoid head on the left, profile facing right, maroon skin with blue-green pebble textures, seamlessly merging into surrounding cliffs with soft, warm yellow-green moss. Include a bright orange-robed figure seated in a tiny, rustic wooden boat on a clearly visible river, placed in the bottom right corner of the image. The scene is illuminated by warm sunlight, enhancing the moss's vibrancy and the contrast between the colossal head and the small figure.

Create an image of a male with dark skin, short black hair, and a full beard. He is wearing a detailed superhero suit that is predominantly green and gold with a muscular build. The suit has armored plating and circular shoulder pads. The expression on his face is serious and focused. The lighting is soft and diffused, coming from the upper right side, and the background is blurred with a hint of classical architecture. Close-up shot focusing on the upper torso and head.

Create an image of a man with a very full beard and piercing gaze to the left, wearing a superhero suit with precise design: green, gold, and dark blue color scheme, detailed with armor plating and circular gold accents symmetrically arranged on the chest. The man has dark brown skin, cropped hair, and a serious side-looking expression. Background is blurred with greenish-brown tones. The close-up image frames from chest up, and natural lighting highlights suit textures and man's detailed facial features.

Generate a detailed image of a middle-aged African American man with a full, neatly groomed beard and a thoughtful, intense expression. He wears a sleek futuristic armor in navy blue with vibrant emerald green and shining gold accents, featuring pronounced circular gold emblems on the chest. The armor has a high-gloss metallic finish and fits snugly over his muscular build. He stands with dignified poise against an intricately detailed indoor background of a grand palace, with elaborate gold and emerald decorations, under soft, warm ambient lighting, viewed from the chest up at eye level.

Generate an image of a dark-skinned middle-aged man with a neatly-trimmed beard and intense gaze. He has short, cropped hair with some graying. He's wearing sharp-edged futuristic armor with pronounced gold lines and green accents over a black suit. The man is stern, head slightly tilted. Blurred background with classical columns. Daytime outdoor natural lighting casts soft shadows on his left side.

Generate a photorealistic image of a male superhero with dark skin, very short black beard, intense serious expression with lips closed, eyebrows slightly furrowed, looking slightly to his left. His suit is realistic metallic with gold and muted green armor, barely visible hexagonal pattern on the green areas, and smaller circular gold emblems on the chest. The background should be a soft-focused mix of beige and brown, suggesting an outdoor scene with soft, natural light from the right side, casting gentle shadows on the face. The superhero's head is tilted slightly down and to his left, with a very high level of detail and texture in the costume reflecting real-world metal.

Create a highly detailed circular cityscape with futuristic and traditional buildings intertwined, surrounded by lush green parks at the center. The city is situated on a spherical segment of earth floating in the sky with visible horizons bending upwards. The atmosphere has clear blue skies. Surrounding the city segment are smaller floating islands with ruins and vegetation. The perspective is top-down with a slight tilt to show the curvature of the earth and the horizon in a circular fisheye lens effect.

Create a highly detailed, photorealistic image of a circular futuristic city on an island from a bird's-eye view, with skyscrapers at the center surrounded by green parks and traditional buildings, all enclosed by a circular body of water. Outside the central island are smaller floating islands with ruins and foliage, under clear blue skies.

Craft a high-resolution image depicting a cityscape arranged in a circular pattern with the illusion of curving upward into a half-globe shape. Centralize a large dome structure and integrate skyscrapers, green parks with detailed textures. This city is situated on a tiny planet in outer space, encircled by crystal-clear water reflecting the city's architecture. In the space surrounding, feature a night sky filled with stars and several small islands afloat with ancient ruins overgrown with vegetation. Ensure the camera perspective is from a higher vantage point looking down to showcase the planet's roundness and the curvature of the city against the cosmic background.

Generate a hyper-realistic high-resolution image of a circular city on a small planet, viewed from above. City has a radial layout, buildings increasing in height towards the center with a visible central round plaza featuring a fountain. Edges of city must show a clear upward curve to indicate a spherical planet. Ocean water should be clear blue with small floating islands containing visible ruins and greenery. Lighting is bright, with a clear blue sky, representing midday. Style should be futuristic but with hyper-realistic textures and details.

Create a high-resolution image of a spherical city floating above water with futuristic skyscrapers and traditional buildings mixed together, surrounded by smaller floating islands with greenery and ruins. The sky is clear blue. The scene is brightly lit as if midday with sunlight casting soft shadows on the buildings. The camera perspective is from above, looking down at the city which is on the upper half of the image, with water visible in the lower half.

Fig. 20: Qualitative examples to showcase the effect of different numbers of streams N and iterations K on PRISM with the same budge $N \times K = 30$.

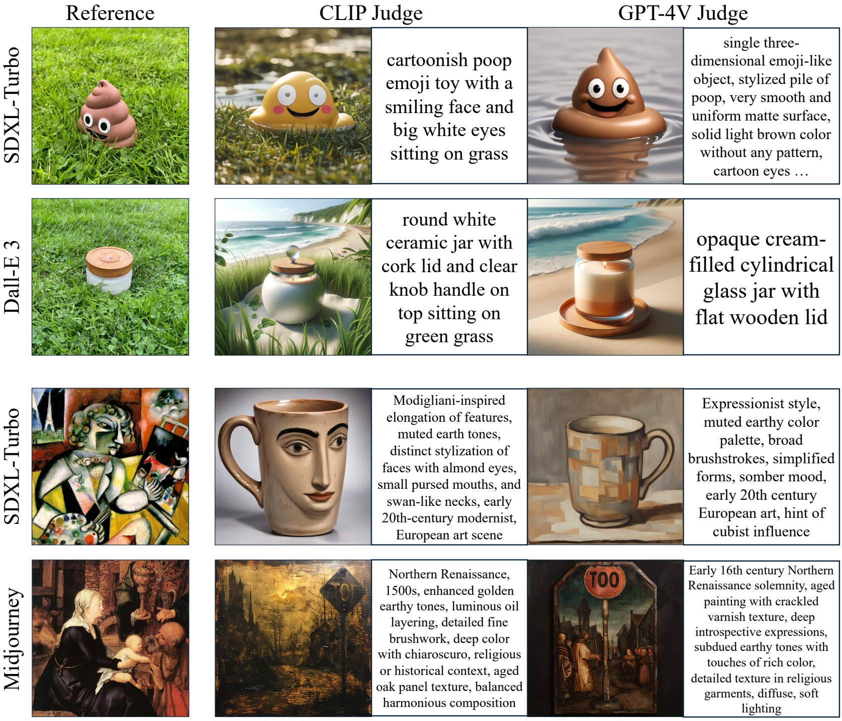

Reference
CLIP Judge
GPT-4V Judge
SDXL-Turbo
cartoonish poop emoji toy with a smiling face and big white eyes sitting on grass
single three-dimensional emoji-like object, stylized pile of poop, very smooth and uniform matte surface, solid light brown color without any pattern, cartoon eyes …
Dall-E 3
round white ceramic jar with cork lid and clear knob handle on top sitting on green grass
opaque cream-filled cylindrical glass jar with flat wooden lid
SDXL-Turbo
Modigliani-inspired elongation of features, muted earth tones, distinct stylization of faces with almond eyes, small pursed mouths, and swan-like necks, early 20th-century modernist, European art scene
Expressionist style, muted earthy color palette, broad brushstrokes, simplified forms, somber mood, early 20th century European art, hint of cubist influence
Midjourney
Northern Renaissance, 1500s, enhanced golden earthy tones, luminous oil layering, detailed fine brushwork, deep color with chiaroscuro, religious or historical context, aged oak panel texture, balanced harmonious composition
Early 16th century Northern Renaissance solemnity, aged painting with crackled varnish texture, deep introspective expressions, subdued earthy tones with touches of rich color, detailed texture in religious garments, diffuse, soft lighting

Fig. 21: Qualitative comparison between using the CLIP model as the Judge **D** in PRISM and using GPT-4V as the Judge.