DreamSync: Aligning Text-to-Image Generation with Image Understanding Feedback

Anonymous CVPR submission

Paper ID ***

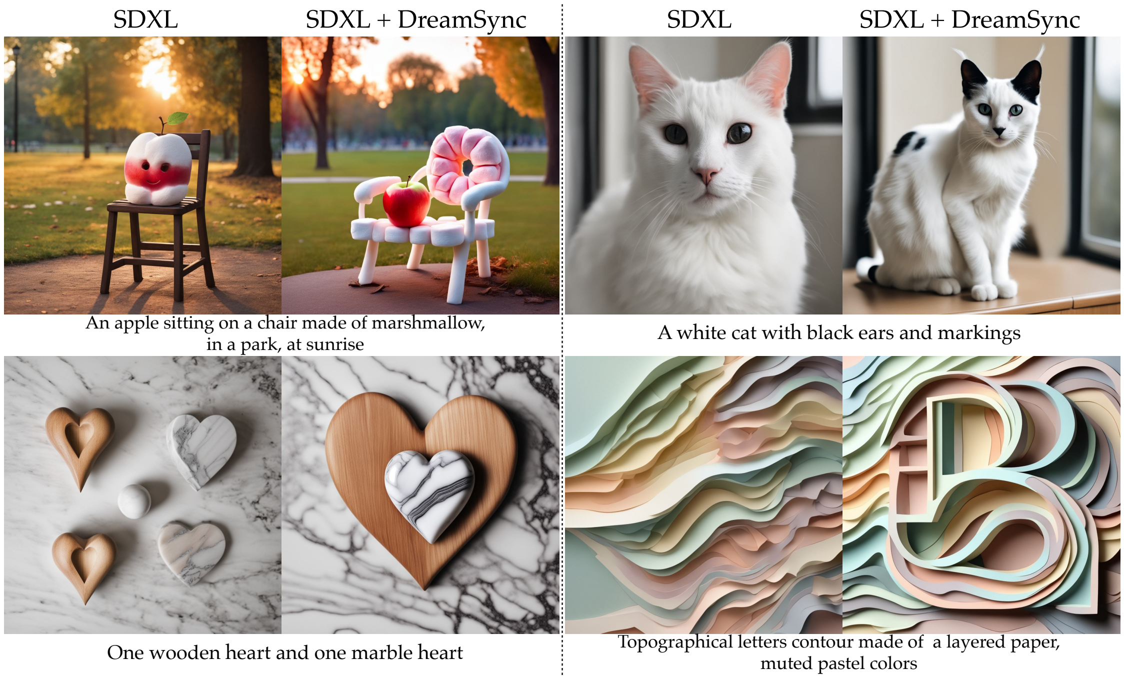


Figure 1. We introduce **DreamSync**: a model-agnostic training algorithm that improves text-to-image (T2I) generation models' faithfulness to text inputs and image aesthetics. DreamSync learns from feedback of vision-language models (VLMs), and does not need any human annotation, model architecture changes, or reinforcement learning.

Abstract

*Despite their wide-spread success, Text-to-Image models (T2I) still struggle to produce images that are both aesthetically pleasing and faithful to the user's input text. We introduce **DreamSync**, a model-agnostic training algorithm by design that improves T2I models to be faithful to the text input. DreamSync builds off a recent insight from TIFA's evaluation framework — that large vision-language models (VLMs) can effectively identify the fine-grained discrepancies between generated images and the text inputs. DreamSync uses this insight to train T2I models without any labeled data; it improves T2I models using its own generations. First, it prompts the model to generate several candidate images for a given input text. Then, it uses two VLMs to select the best generation: a Visual Question Answering model that measures the alignment of generated images to the text, and another that measures the generation's aesthetic quality. After selection, we use LoRA to iteratively finetune the T2I model to guide its generation towards the selected best generations. DreamSync does not need any additional human annotation, model architecture changes, or reinforcement learning. Despite its simplicity, DreamSync improves both the semantic alignment and aesthetic appeal of two diffusion-based T2I models, evidenced by multiple benchmarks (+1.7% on TIFA, +2.9% on DSG1K, +3.4% on VILA aesthetic) and human evaluation.*

1. Introduction

Although we invite creative liberty when we commission art, we expect an artist to follow our instructions. Despite the advances in text-to-image (T2I) generation models [40, 41, 44, 47, 55], it remains challenging to ob-

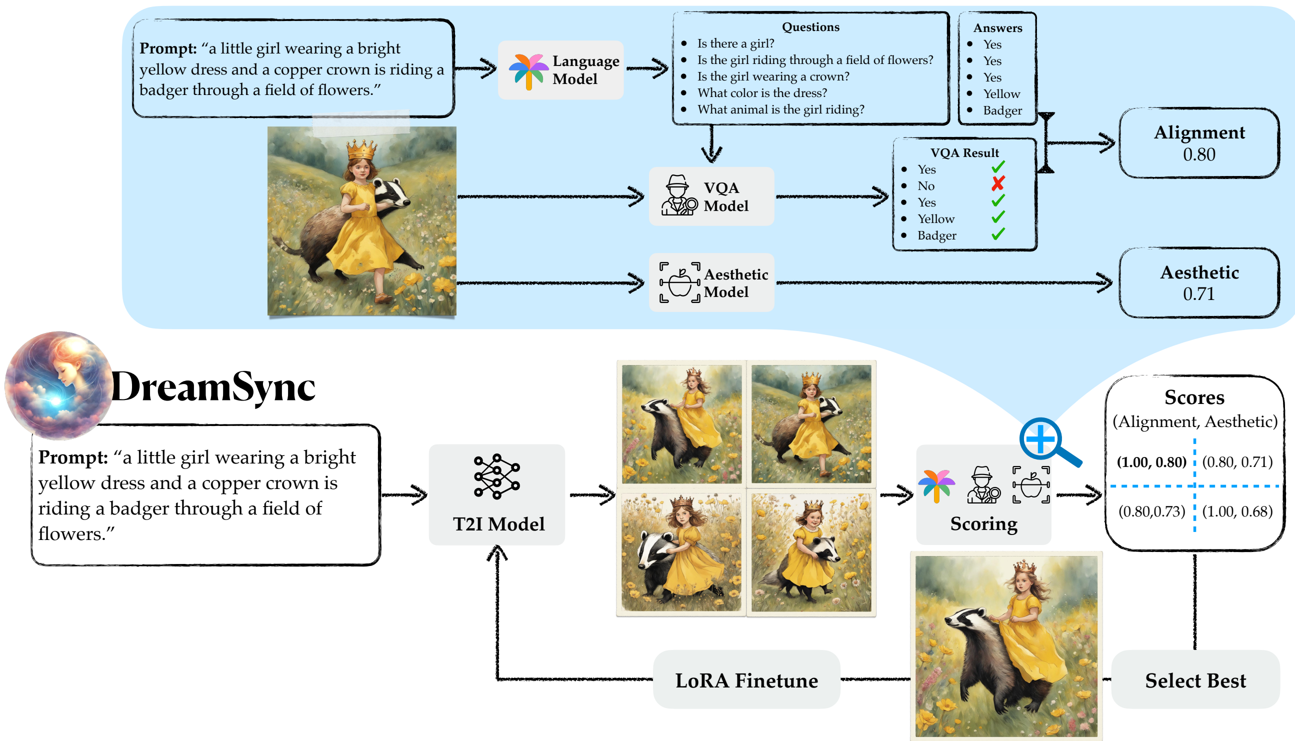


Figure 2. **DreamSync**. Given a prompt, a text-to-image generation model generates multiple candidate images, which are evaluated by two VLM models: one VQA model that provides feedback on text faithfulness and the other on image aesthetics. The best image chosen by the VLMs are collected to fine tune the T2I model. This process can repeat indefinitely until convergence on feedback is achieved.

tain images that meticulously conform to users' intentions [14, 27, 29, 30, 36, 42, 43]. Current models often fail to compose multiple objects [14, 29, 36], bind attributes to the wrong objects [14], and struggle to generate visual text [30]. In fact, the difficulty of finding effective textual prompts has led to a myriad of websites and forums dedicated to collecting and sharing useful prompts (e.g. PromptHero, Arthub.ai, Reddit/StableDiffusion). There are also online marketplaces for purchasing and selling useful such commands (e.g. PromptBase). The onus to generate aesthetic images that are faithful to a user's desires should lie with the model and *not* with the user.

Today, there are efforts to address these challenges. For example, it is possible to manipulate attention maps based on linguistic structure to improve attribute-object binding [14, 43]; or train reward models using human feedback to better align generations with user intent [13, 27]. Unfortunately, these methods either operate on a specific model architecture [14, 43] or require expensive labeled human data [13, 27]. Worse, most of these methods sacrifice aesthetic appeal when optimizing for faithfulness, which we confirm in our experiments.

We introduce **DreamSync, a model-agnostic framework that improves T2I generation faithfulness while maintaining aesthetic appeal**. Our approach extends work on fine-tuning T2I models for alignment, but does not require any human feedback. The key insight behind DreamSync is in leveraging the advances in vision-language models (VLMs), which can identify fine-grained discrepencies between the generated image and the user's input text [7, 20]. Intuitively at a high level, our method can be thought of as a scalable version of reinforcement learning with human feedback (RLHF); just as LLaMA2 [49] was iteratively refined using human feedback, DreamSync improves T2I models using feedback from VLMs, except without the need for reinforcement learning.

Given a set of textual prompts, T2I models first generates multiple candidate images per prompt. DreamSync automatically evaluates these generated images using two VLMs. The first one measures the generation's faithfulness to the text [7, 20], while the second one measures aesthetic quality [23]. The best generations are collected and used to finetune the T2I model using parameter-efficient LoRA finetuning [19]. With the new finetuned T2I model, we repeat the entire process for multiple iterations: generate images, curate a new finetuning set, and finetune again.

We conduct extensive experiments with latest benchmarks and human evaluation. We experiment DreamSync with two T2I models, SDXL [37] and SD v1.4 [39]. Results on both models show that **DreamSync enhance the align-**

ment of images to user inputs and retains their aesthetic quality. Specifically, quantitative results on TIFA [21] and DSG [7] demonstrate that **DreamSync is more effective than all baseline alignment methods** on SD v1.4, and can yield even bigger improvements on SDXL. Human evaluation on SDXL shows that DreamSync give consistent improvement on all categories of alignment in DSG. While our study primarily focuses on boosting faithfulness and aesthetic quality, DreamSync has broader applications: it can be used to improve other characteristics of an image as long as there is an underlying model that can measure that characteristic.

2. Related Work

T2I Evaluation with VLMs. Several prior works have proposed to use VQA models to evaluate text-to-image generation. The TIFA benchmark, which pioneered this approach for evaluation, consists of 4K prompts and 25K questions across 12 categories (e.g., object, count, material), enabling T2I model evaluation by using VQA models to answer questions about the generated images [20]. TIFA prompts come from various resources, including DrawBench used in Imagen [47], PartiPrompt used in Parti [55], PaintSkill [6] used in Dall-Eval, etc. DSG [7] further improves TIFA's realiability by examining their evaluation questions carefully. Another related benchmark is SeeTrue, which also uses VQA models to measure alignment [53]. Before the VQA evaluation era, several other evaluation benchmarks were proposed focusing primarily on compositional text prompts for attribute binding (e.g., color, texture, shape) and object relationships (e.g., spatial). Examples include T2I-CompBench [21], C-Flowers [35], CC-500 and ABC-6K benchmarks [15]. Aside from automated benchmarks, human evaluation for text-to-image generation is widely used in the community, although such annotations are notoriously costly to collect. In response, Xu et al. [52] propose ImageReward, the first general purpose text-to-image human preference reward model to encode human preferences automatically. In our work, we use a collection of three evaluation methods to evaluate DreamSync: VQA evaluation for generated images on both TIFA and DSG benchmarks, human evaluation, and ImageReward for automatic human preference prediction.

Improving General T2I Alignment. We roughly categorize the alignment methods for improving T2I alignment into two classes depending on if they involve training. For training-involved methods, several works use Reinforcement Learning from Human Feedback (RLHF) based on human rankings to maximize a reward and improve faithful generation [13, 22, 27]. In a similar vein, Pick-a-Pic is a dataset of prompts and preferences that is used to train a CLIP-based scoring function [24]. StyleDrop trains adapters to synthesize of images that follow a specific style [48], and T2I-Adapter trains adapters to improve the control for the color and structure of the generation results [33]. DreamBooth and HyperDreamBooth improve personalized generation [45, 46], and they have inspired more efficient methods such as SVDiff [17]. Being orthogonal to training-involved methods, there is a body of work on training-free methods that make inference time adjustments to the model to improve alignment, such as SynGen and StructureDiffusion [12, 15, 18, 43]. DreamSync leverages training but does not involve reinforcement learning. We compare DreamSync with two RL-based methods and two learning-free methods in our experiments. We find that DreamSync outperform all the baselines in terms of text-image alignment on both DSG and TIFA.

Iterative Bootstrapping. Iterative Bootstrapping, also known as model self-training, is a semi-supervised learning approach that utilizes a teacher model to assign labels to unlabelled data, which is then used to train a student model [16, 26, 32, 54]. In our work, we adopt a self-training scheme where the teacher model are the VLMs and the student model is the T2I model we aim to improve. During training, the VLMs (teacher) are used to annotate and select aligned examples for the next batch finetuning (student).

3. DreamSync

Our method improves alignment and aesthetics in four steps (see Figure 2): Sample, Evaluate, Filter, and Finetune. The high level idea is that T2I models are capable of generating interesting and varied samples. These examples are further judged by VLMs to pass qualification as faithful and aesthetic candidates for further finetuning T2I models. We next dive into each component more formally.

Sample. Given a text prompt T, the text-to-image generation model G generates an image $I = G(T)$. Generation models are randomized, and running G multiple times on the same prompt T can produce different images, which we index as $\{I^{(k)}\}_{k=1}^{K}$. To improve the model's faithfulness to text guidance, our method collects faithful examples generated by G. We use G to generate K samples of the same prompt T, so that with some probability $\delta > 0$, a generated image I is faithful. Note that we need $K = \Omega(1/\delta)$ samples for each prompt T, and DreamSync is not expected to improve totally unaligned models (with $\delta \to 0$). Prior work [22] estimates that 5–10 samples can yield a good image, and hence, δ can be thought of as roughly 0.1 to 0.2.

Evaluate. For each text prompt T, we derive a set of $\mathcal{N}_T$ question-answer pairs $\{\mathcal{Q}(T), \mathcal{A}(T)\}$ that can be used to test whether a generated image I is faithful to T. We use an LLM to generate these pairs, only using the prompt T as input (with no images). Typically $\mathcal{N}_T \approx 10$. We use VQA models to evaluate the faithfulness of the generation model, $F_j(T, I) = \mathbb{1}\{\text{VQA}(I, \mathcal{Q}_j(T)) = \mathcal{A}_j(T)\}$, for

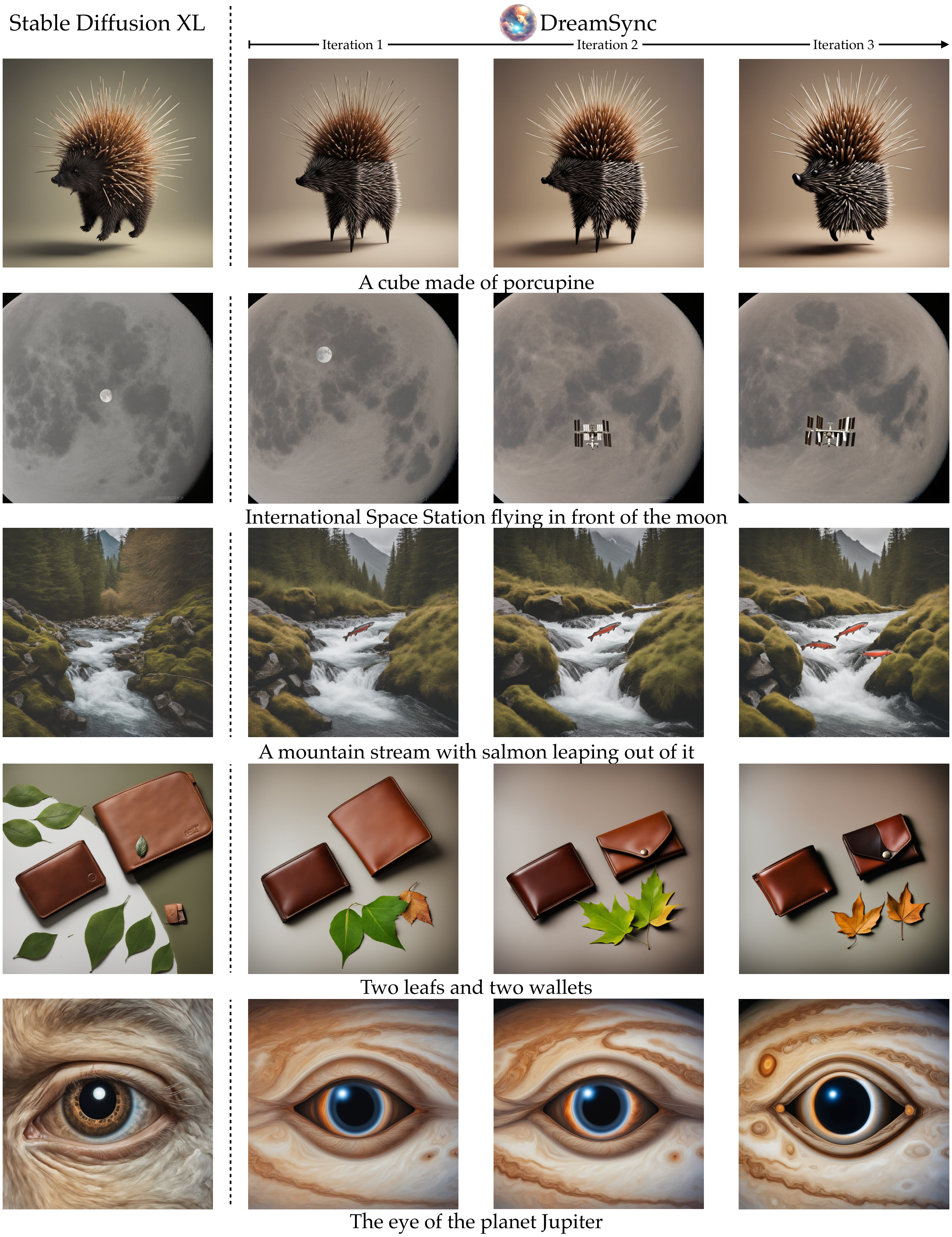


Figure 3. Qualitative examples of DreamSync improving image-text alignment after each iteration. LoRA fine-tuning on generated and filtered prompt-image pairs can steer the model to gradually capture more components of the text inputs.

$j \in \{1, \dots, \mathcal{N}_T\}$. We measure the faithfulness of a caption-image pair (T, I) given all questions and answers, using two metrics. Intuitively, we can average the number of correct answers, or we can be more strict, and only count an image as a success if all the answers are correct. Formally, the *Mean* score is the expected success rate

$$\mathcal{S}_{\mathrm{M}}(T, I) = \frac{1}{\mathcal{N}_T} \sum_{j=1}^{\mathcal{N}_T} F_j(T, I),$$

and the *Absolute* score is the absolute success rate

$$\mathcal{S}_{\mathrm{A}}(T, I) = \prod_{j=1}^{\mathcal{N}_T} F_j(T, I).$$

Filter. We combine text faithfulness and visual appeal (given by $\mathcal{V}(\cdot)$) as rewards for filtering. For a text prompt T and its corresponding synthetic image set $\{I_k\}_{k=1}^{K}$, we select samples that pass both VQA and aesthetic filters:

$$C(T) = \{(T, I_k) : \mathcal{S}_{\mathrm{M}}(T, I_k) \geq \theta_{\text{Faithful}}, \\ \mathcal{V}(I_k) \geq \theta_{\text{Aesthetic}}\}.$$

To avoid an imbalanced distribution where easy prompts have more samples, which could cause adversely affected image quality, we select one representative image (denoted as $\hat{I}_T$) having the highest visual appeal for each T:

$$(T, \hat{I}_T) = \underset{\mathcal{V}(I_k)}{\operatorname{argmax}}\, C(T).$$

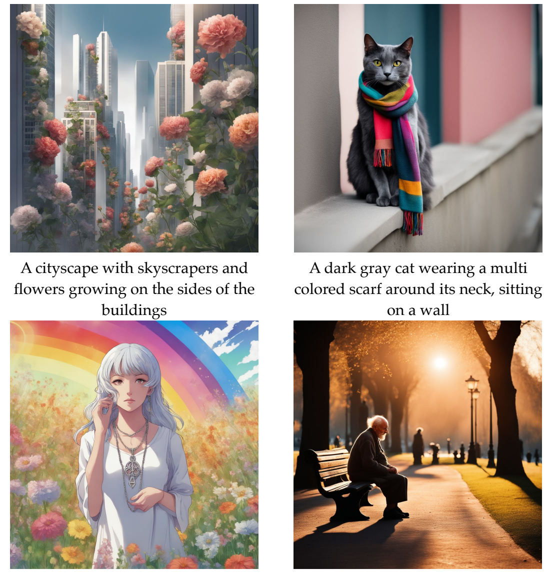
We apply this procedure to all text prompts in our finetuning prompt set $\{T_i\}_{i=1}^{N}$ with $T_i \sim \mathcal{D}$, where $\mathcal{D}$ is a prompt distribution. After filtering, we collect a subset of examples, $D(G) := \bigcup_{i \in \{j | C(T_j) \neq \varnothing\}} \{(T_i, \hat{I}_{T_i})\}$, that meet our aesthetic and faithfulness criteria. Note that it is possible for $C(T_i)$ to be empty, and we empirically show what fraction of the training data is selected in Figure 5. We ablate other aspects of the selection procedure in § 5.3.

Finetune. After obtaining a new subset of faithful and aesthetic text-image pairs, we fine-tune our generative model G on this set. We denote the generative model after s iterations of DreamSync as G_s, such that G_0 denotes the baseline model. To obtain G_{s+1} we fine-tune on data generated by G_s after applying our filtering procedure as outlined above. We follow the same loss objective and fine-tuning dynamics as LoRA [19]. Let $\Theta(\cdot)$ denote all parameters of a model, then the hypothesis class at iteration s is:

$$\mathcal{G}_s = \left\{ G \mid \operatorname{rank}\Big(\Theta(G) - \Theta(G_s)\Big) \leq R \right\}.$$

where R denotes the rank of weight updates and in practice we choose $R = 128$ to balance efficiency and image quality. Overall, the iterative training procedure is as follows:

$$G_{s+1} = \underset{G \in \mathcal{G}_s}{\operatorname{argmin}} \frac{1}{|D(G_s)|} \sum_{(T_j, I_j) \in D(G_s)} \ell(G(T_j), I_j). \quad (1)$$



A cityscape with skyscrapers and flowers growing on the sides of the buildings

A dark gray cat wearing a multi colored scarf around its neck, sitting on a wall

A colorful anime illustration of a woman wearing a silver necklace, standing in a field of flowers, with a rainbow in the background

An intriguing photo of an old man sitting on a bench in the park, lit by the setting sun

Figure 4. PaLM-2 generated training prompts and their corresponding images generated via DreamSync. Prompt acquisition requires no human effort. It enables us to train on more complex and diversified prompt-image pairs than found in typical datasets.

The self-training process Eq. (1) can in principle be executed indefinitely. In practice, it repeats for three iterations at which point we observe diminishing returns.

4. Datasets and Evaluation

In this section, we will introduce our training data in § 4.1 and evaluation benchmark in § 4.2.

4.1. Training Data Acquisition

To obtain prompts, and corresponding question-answer pairs without human-in-the-loop, we utilize the in-context learning capability of Large Language Models (LLM). We choose PaLM 2[1] [1] as our LLM and proceed as follows:

1. *Prompt Generation.* We provide five hand-crafted seed prompts as examples and then ask PaLM 2 to generate similar textual prompts. We include additional instructions that specify the prompt length, a category (randomly drawn from twelve desired categories as in [20], e.g., spatial, counting, food, animal/human, activity), no repetition, etc.[2] We change the seed prompts and repeat the prompt generation three times.
2. *QA Generation.* Given prompts, we then use PaLM 2

[1] https://ai.google/discover/palm2/
[2] In Appendix A.1, we show the complete instruction used to probe LLM for the first two steps: prompt generation and QA generation.

Model		Alignment	Text Faithfulness: TIFA *Mean*	Text Faithfulness: TIFA *Absolute*	Text Faithfulness: DSG1K	Visual Appeal
SD v1.4 [39]		No alignment	76.6	33.6	72.0	44.6
	Training-Free	SynGen [43]	76.8 (+0.2)	34.1 (+0.5)	71.2 (–0.8)	42.4 (–2.2)
		StructureDiffusion [15]	76.5 (–0.1)	33.6 (+0.0)	71.9 (–0.1)	41.5 (–3.1)
	RL	DPOK [13]	76.4 (–0.2)	33.8 (+0.2)	70.3 (–1.7)	**46.5** (+1.9)
		DDPO [4]	76.7 (+0.1)	34.4 (+0.8)	70.0 (–2.0)	43.5 (–1.1)
		DreamSync (ours)	**77.6** (+1.0)	**35.3** (+1.7)	**73.2** (+1.2)	44.9 (+0.3)
SDXL [37]		No alignment	83.5	45.5	83.4	60.9
		DreamSync (ours)	**85.2** (+1.7)	**49.2** (+3.7)	**86.3** (+2.9)	**64.3** (+3.4)

Table 1. **Benchmark on Text Faithfulness and Visual Appeal.** All models are sampled with the same set of four seeds, i.e. $K = 4$. Best scores under each backbone T2I model are highlighted in **bold**; gain and loss compared to base models are highlighted accordingly. DreamSync significantly improve SD-XL and SD v1.4 in alignment and visual appeal across all benchmark. Additionally, DreamSync does not sacrifice image quality when improving faithfulness.

again to generate question and answer pairs that we will use as input for VQA models as in TIFA [20].

3. *Filtering.* We finally use PaLM 2 once more to filter out unanswerable QA pairs. Here our instruction aims to identify three scenarios: the question has multiple answers (e.g., "`black and white panda`" where the object has multiple colors, each color could be the answer), the answer is ambiguous (e.g., "`a lot of people`") or the answer is not valid to the question.

We showcase the diversity of PaLM 2 generated prompts in Figure 4 using qualitative examples and quantitive statistics of our generated prompts in Appendix A.2.

4.2. Evaluation Benchmarks

Using the previously generated prompts, we evaluate whether DreamSync can improve the T2I model performance on benchmarks that include general prompts. We consider the follow benchmarks.

TIFA. To evaluate the faithfulness of the generated images to the textual input, TIFA [20] uses VQA models to check whether, given a generated image, questions about its content are answered correctly. There are 4k diverse prompts and 25k questions spread across 12 categories in the TIFA benchmark. Although there is no overlap between our training data and TIFA, we use the TIFA attributes to constrain our LLM-based prompt generation. Therefore, we use TIFA to test DreamSync on in-distribution prompts. We follow TIFA and use BLIP-2 as the VQA model for evaluation.

Davidsonian Scene Graph (DSG). DSG [7] exhibits the same VQA-as-evaluator insight as TIFA's and further improves its reliability. Specifically, DSG ensures that all questions are atomic, distinct, unambiguous, and valid. To comprehensively evaluate T2I images, DSG provides 1,060 prompts covering many concepts and writing styles from different datasets that are completely independent from DreamSync's training data acquisition stage. Not only is DSG a strong T2I benchmark, it also enables further analysis of DreamSync with out-of-distribution prompts. Furthermore, DSG uses PaLI as the VQA model for evaluation, which is different from the VQA model that we use in training (*i.e.*, BLIP-2) and lifts the concern of VQA model bias in evaluation. We use DSG QA both automatically (with PaLI) and with human raters (details in Appendix C).

5. Experiments

We explain our experimental setup in § 5.1, and showcase the efficacy of training with DreamSync and compare against other methods in § 5.2. § 5.3 analyzes our choice of rewards; § 5.4 reports results for a human study.

5.1. Experimental set-up

Base Model. We evaluate DreamSync on Stable Diffusion v1.4 [39], which is also used in related work. Additionally, we consider SDXL [37], which is the current state-of-the-art open-sourced T2I model. For each prompt, we generate eight images per prompt, i.e., $K = 8$.

Fine-grained VLM Feedback. We use feedback from two VLM models to decide what text-image pairs to keep for finetuning. We use BLIP-2 [28] as the VQA model to measure the faithfulness of generated images to textual input and and VILA [23] to measure the aesthetics measurement score. Empirically, we keep the text-image pairs whose VQA scores are greater than $\theta_{\text{Faithful}} = 0.9$ and aesthetics score greater than $\theta_{\text{Aesthatics}} = 0.6$. If there are

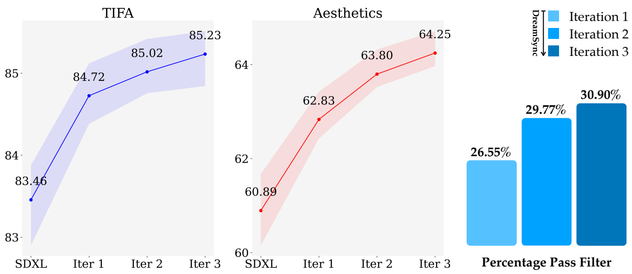


Figure 5. DreamSync improves faithfulness and aesthetics iteratively. More examples pass the filters with additional iterations.

multiple generated images passing the threshold, we keep the one with the highest VILA score. Starting from 28,250 prompts, we find that more than 25% prompts are kept for $D(G_0)$ (for both T2I models), which we will use for finetuning. We later show that this percentage increases further as we perform additional DreamSync iterations.

Baselines. We compare DreamSync with two types of methods that improve the faithfulness of T2I models: two training-free methods (StructureDiffusion [15] and SynGen [43]) and two RL-based methods (DPOK [13] and DDPO [4]). As the baselines use SD v1.4 as their backbone, we also use it with DreamSync for a fair comparison.

5.2. Benchmark Results

In Table 1 we compare DreamSync to various state-of-the-art approaches with four random seeds. In Appendices D and E we show more qualitative comparisons.

DreamSync Improves the Alignment and Aesthetics of both SDXL and SD v1.4. For SDXL [37], we show how three iterations of DreamSync improves the generation faithfulness by 1.7 point of mean score and 3.7 point of absolute score on TIFA. The visual aesthetic scores after performing DreamSync improved by 3.4 points. Due to the model-agnostic nature, it is straightforward to apply DreamSync to different T2I models. We also apply DreamSync to SD V1.4 [39]. DreamSync improves faithfulness by 1.0 points of mean score and 1.7 points of absolute score on TIFA, together with a 0.3 points of VILA score improvement for aesthetics. Most prominently on DSG1K, DreamSync improve text faithfulness of SDXL by 2.9 points. We report fine-grained results for DSG in Appendix C.

DreamSync yields the best performance in terms of textual faithfulness on TIFA and DSG. This is true without sacrificing the visual appearance as shown in Table 1. In Figure 5 we report TIFA and aesthetics scores for each iteration, where we observe how DreamSync gradually improves the alignment and aesthetics of the generated images. We highlight several qualitative examples in Figure 3.

Rewards		Text Faithfulness	Visual Appeal
VQA	VILA		
-	-	83.5	60.9
✓		**84.8**	61.9
	✓	83.8	61.7
✓	✓	84.7	**62.8**

Table 2. Ablation of different VLM rewards. Models are evaluated after *one DreamSync iteration*.

T2I Model	Alignment Method	Evaluation Dataset	
		TIFA	DSG1K
SD v1.4	No alignment	0.056	-0.220
	SynGen	0.149	-0.237
	StructureDiffusion	0.075	-0.135
	DPOK	0.067	-0.258
	DDPO	0.152	-0.076
	DreamSync (ours)	**0.168**	**-0.054**
SD XL	No alignment	0.878	0.702
	DreamSync (ours)	**1.020**	**0.837**

Table 3. Scores given by the human preference model ImageReward [52]; model scores are logits and can be negative. Models trained with DreamSync outperform other baselines (higher is better), without using any human annotation.

5.3. Analysis & Ablations

Impact of VQA model on evaluation. We analyze whether using BLIP-2 as a VQA model for finetuning and for evaluation in TIFA might be the reason for the improvement by DreamSync that we have observed. To test this we use PaLI [5] to replace the BLIP-2 as the VQA in TIFA. Using SDXL as the backbone, DreamSync improves the mean score from 90.09 to 92.02 on TIFA compared to the vanilla SDXL model. This results confirms that DreamSync is in fact able to improve the textual faithfulness of T2I models.

Ablating the Reward Models In Table 2, we present the results for an ablation study where we remove one of the VLMs during filtering and evaluate SDXL after applying one iteration of DreamSync. It can be seen how training with a single pillar mainly leads to an improvement in the corresponding metric, while the combination of the two VLM models leads to strong performance for both text faithfulness and visual easthatics, justifying our approach. One interesting finding is that training with both rewards, rather than VILA only, gives the highest visual appeal score. Our possible explanation is that images that align with user inputs may have higher visual appeal.

ImageReward. We next test whether DreamSync yields an improvement on human preference reward models, even though DreamSync is not trained to optimize them. We use ImageReward [52] as an off-the-shelf human preference model for generated images. Table 3 shows that DreamSync plus either SD v1.4 or SDXL increases ImageReward scores

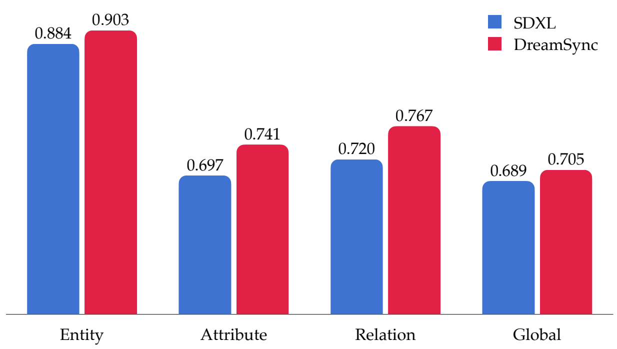


Figure 6. Human study with three raters on 1060 DSG prompts.

on images based on both TIFA and DSG1K. Tuning with VLM-based feedback helps align the generated images with human preferences, at least according to ImageReward.

5.4. Human Evaluation

To corroborate the VQA-based results, we first conduct a preliminary human study to evaluate the faithfulness of generated images. It shows simply asking one question `‘‘Which image better aligns with the prompt?’’` yields poor inter-annotator agreement. We speculate that asking a single question encompassing the whole prompt makes the alignment difficult to evaluate.

To address this issue, we conduct a larger follow-up study based on DSG [7], where we ask approximately 8 fine-grained questions for each of 1060 images to external raters. These questions are divided into categories (entity, attribute, relation, global). Here in Figure 6, we observe consistent and statistically significant improvements comparing DreamSync to SDXL. In each category, images from DreamSync contain more components of the prompts, while excluding extraneous features. Overall, DreamSync's images led to 3.4% more correct answers than SDXL images, from 70.9% to 74.3%. Full details and findings for both studies are in Appendix C.

6. Discussion

A key design choice behind DreamSync is to maintain simplicity and automation throughout each step of the pipeline. Despite this feature, our experimental results show that DreamSync can improve both SD v1.4 and SDXL on TIFA, DSG, and visual appeal. In the case of SD v1.4, this improvement holds true compared to four different baseline models (two training-free and two RL-based). For SDXL, even though the base model achieves SoTA results among open-source models, DreamSync can still substantially improve both alignment and aesthetics.

The effectiveness of DreamSync's self-training methodology opens the door for a new paradigm of parameter-efficient finetuning. Indeed, the DreamSync pipeline is easily generalizable. For the training prompts, we can construct a set with complex and non-conventional examples compared to standard web-scraped data. On the filtering and fine-tuning side, our framework shows that VLMs can provide effective feedback for T2I models. Together, these steps do not require human annotations, yet they can tailor a generative model toward desirable criteria.

6.1. Limitations

Like prior methods, the performance of DreamSync is limited by the pre-trained model it starts with. As exemplified in "the eye of the planet Jupiter" in Figure 3, SDXL generates a human's eye rather than Jupiter's. DreamSync adds more features of the Jupiter in each iteration. Nevertheless, it did not manage to produce an image that is perfectly faithful to the prompt. This is also exemplified by the quantitative results in §5.2. Despite outperforming the baselines using SD v1.4 on TIFA and DSG, SD v1.4 + DreamSync still falls behind SDXL. Similarly, our human studies on DSG in §5.4 indicate that DreamSync improves SDXL from 70.9% accuracy to 74.3%. Nonetheless, there is still a 25.7% headroom to improve. We identify several common failure modes (e.g., attribute-binding) and conduct a detailed analysis in Appendix B. Future works may investigate if these challenges can be addressed by further scaling up DreamSync, or mixing it with large-scale pre-training.

7. Conclusion

We introduce DreamSync, a versatile framework to improve text-to-image (T2I) synthesis with feedback from image understanding models. Our dual VLM feedback mechanism helps in both the alignment of images with textual input and the aesthetic quality of the generated images. Through evaluations on two challenging T2I benchmarks (with over five thousand prompts), we demonstrate that DreamSync can improve both SD v1.4 and SDXL for both alignment and visual appeal. The benchmarks also show that DreamSync performs well in both in-distribution and out-of-distributions settings. Furthermore, human ratings and a human preference prediction model largely agree with DreamSync's improvement on benchmark datasets.

For future work, one direction is to ground the feedback mechanism to give fine-grained annotations (e.g., bounding boxes to point out where in the image the misalignment lies). Another direction is to tailor the prompts used at each iteration of DreamSync to target different improvements: backpropagating VLM feedbacks to the prompt acquisition pipelines for continual learning.

References

[1] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report, 2023. 5, 17

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, JoyceLee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions, 2023. 15

[3] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *ICCV*, 2023. 17

[4] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning, 2023. 6, 7

[5] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel M. Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. *ArXiv*, abs/2209.06794, 2022. 7, 17

[6] Jaemin Cho, Abhaysinh Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. *ArXiv*, abs/2202.04053, 2022. 3

[7] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation, 2023. 2, 3, 6, 8, 17

[8] Donald Davidson. Theories of meaning and learnable languages. In *In Yehoshua Bar-Hillel (ed.), Proceedings of the 1964 International Congress for Logic, Methodology, and Philosophy of Science. Amsterdam: North-Holland. pp. 383-394*, 1965. 17

[9] Donald Davidson. The logical form of action sentences. In *n N. Rescher (ed.) The Logic of Decision and Action, Pittsburgh: University of Pittsburgh*, 1967.

[10] Donald Davidson. Truth and meaning. In *Inquiries into Truth and Interpretation; Soames, chapter 12 of PATC*, 1967. 17

[11] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. PoseScript: 3D Human Poses from Natural Language. In *ECCV*, 2022. 17

[12] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023. 3

[13] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *arXiv preprint arXiv:2305.16381*, 2023. 2, 3, 6, 7

[14] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, P. Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *ArXiv*, abs/2212.05032, 2022. 2

[15] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023. 3, 6, 7

[16] Deqing Fu, Ameya Godbole, and Robin Jia. Scene: Self-labeled counterfactuals for extrapolating to negative examples. *ArXiv*, abs/2305.07984, 2023. 3

[17] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 3

[18] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. *arXiv preprint arXiv:2210.00939*, 2022. 3

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.

LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2, 5

[20] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*, 2023. 2, 3, 5, 6, 17

[21] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv:2307.06350*, 2023. 3

[22] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023. 3

[23] Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, and Feng Yang. Vila: Learning image aesthetics from user comments with vision-language pretraining, 2023. 2, 6

[24] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 3

[25] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *CVPR*, 2017. 17

[26] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5468–5479. PMLR, 2020. 3

[27] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023. 2, 3

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023. 6

[29] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. *ArXiv*, abs/2206.01714, 2022. 2

[30] Rosanne Liu, Daniel H Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, R. J. Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. *ArXiv*, abs/2212.10562, 2022. 2, 17

[31] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, 2016. 17

[32] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 152–159, New York City, USA, 2006. Association for Computational Linguistics. 3

[33] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 3

[34] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *ICCV*, 2023. 17

[35] Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. Benchmark for compositional text-to-image synthesis. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 3

[36] Vitali Petsiuk, Alexander E Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A Plummer, Ori Kerret, et al. Human evaluation of text-to-image models on a multi-task benchmark. *arXiv preprint arXiv:2211.12112*, 2022. 2

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2, 6, 7

[38] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020. 17

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 6, 7

[40] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021. 1

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1

[42] Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models, 2022. 2

[43] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment, 2023. 2, 3, 6, 7

[44] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021. 1

[45] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models. *arXiv preprint arXiv:2307.06949*, 2023. 3

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 1, 3

[48] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*, 2023. 3

[49] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. 2

[50] Iulia Turc and Gaurav Nemade. Midjourney user prompts & generated images (250k), 2022. 17

[51] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *ACL*, 2023. 17

[52] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 3, 7

[53] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*, 2023. 3

[54] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA, 1995. Association for Computational Linguistics. 3

[55] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv*, abs/2206.10789, 2022. 1, 3

A. Training Data Acquisition

A.1. LLM Instructions

Training Data Acquisition is the first step and the foundation of DreamSync as discussed in Section 4.1. We use PaLM 2 for each step of the training data acquisition, including prompt generation, QA generation and filtering. Here are the complete instructions that we use.

Instruction for Prompt Generation. `You are a large language model, trained on a massive dataset of text. You can generate texts from given examples. You are asked to generate similar examples to the provided ones and follow these rules:`

```
1. Your generation will be served as prompts for Text-to-Image models. So your prompt should be as
   visual as possible.
2. Do NOT generate scary prompts.
3. Do NOT repeat any existing examples.
4. Your generated examples should be as creative as possible.
5. Your generated examples should not have repetition.
6. Your generated examples should be as diverse as possible.
7. Do NOT include extra texts such as greetings.
```

Instruction for QA Generation. `Given a image descriptions, generate one or two multiple-choice questions that verifies if the image description is correct. Classify each concept into a type (object, human, animal, food, activity, attribute, counting, color, material, spatial, location, shape, other), and then generate a question for each type.` We then provide fifteen prompts together with about ten question answer pairs as demonstration for PaLM 2. Table 4 shows an example of PaLM2-generated *prompt* and *QA*. *Answer source* and *Answer Type* are also automatically generated altogether, making it possible for us to get statistics of our training set below.

A.2. Statistics

Table 5 shows the statistics of the prompts and questions we obtained, and we list a few prompts from our training set and DreamSync's generation in Figure 4. Prior work (e.g., TIFA, DSG) identifies that T2I models do not perform equally well for depicting different attribute categories; we verify the variety of attributes in our prompts by counting unique words (i.e., *Answer Source* in Table 4) in these categories (i.e., *Answer Type* in Table 4): counting (4179), object (3638), shape (973),

Prompt	Question and Choices	Answer Source	Answer Type
6 baseball players, each holding a sheep, and they are all standing in a field of flowers	question : what is in the field? choices: ["**flowers**", "grass", "trees", "rocks"]	flowers	object
	is there a field? choices: ["**yes**", "no"]	field	location
	are there flowers? choices: ["**yes**", "no"]	flowers	object
	what type of place is this? choices: ["**field**", "park", "forest", "mountain"]	field	location
	are the baseball players holding sheep? choices: ["**yes**", "no"]	holding	activity
	are there sheep? choices: ["**yes**", "no"]	sheep	animal
	are there baseball players? choices: ["**yes**", "no"]	baseball players	human
	how many baseball players are there? choices: ["1", "2", "3", "4", "5", "**6**"]	6	human
	how many sheep are there? choices: ["1", "2", "3", "4", "5", "**6**"]	6	animal

Table 4. One example of PaLM2-generated *prompt* and *QA*. *Answer source* and *Answer Type* are also generated by PaLM 2, making it possible for us to get statistics of our training set. We highlight correct answers in **bold** here.

CVPR #***

# of prompts	28,250
# of questions	239,310
- # of binary questions	125,094
- # of multiple-choice questions	114,214
avg. # of questions per prompt	8.5
avg. # of words per prompt	16.7
avg. # of elements per prompt	1.9

Table 5. Statistics of Training Set for DreamSync.

human (945), location (1047), activity (2984), attribute (2925), color (3259), food (1086), spatial (1009), animal (645), material (1610), existence (3072), and other (878).

A.3. Images Generated by DreamSync for Finetuning Exhibit High Quality

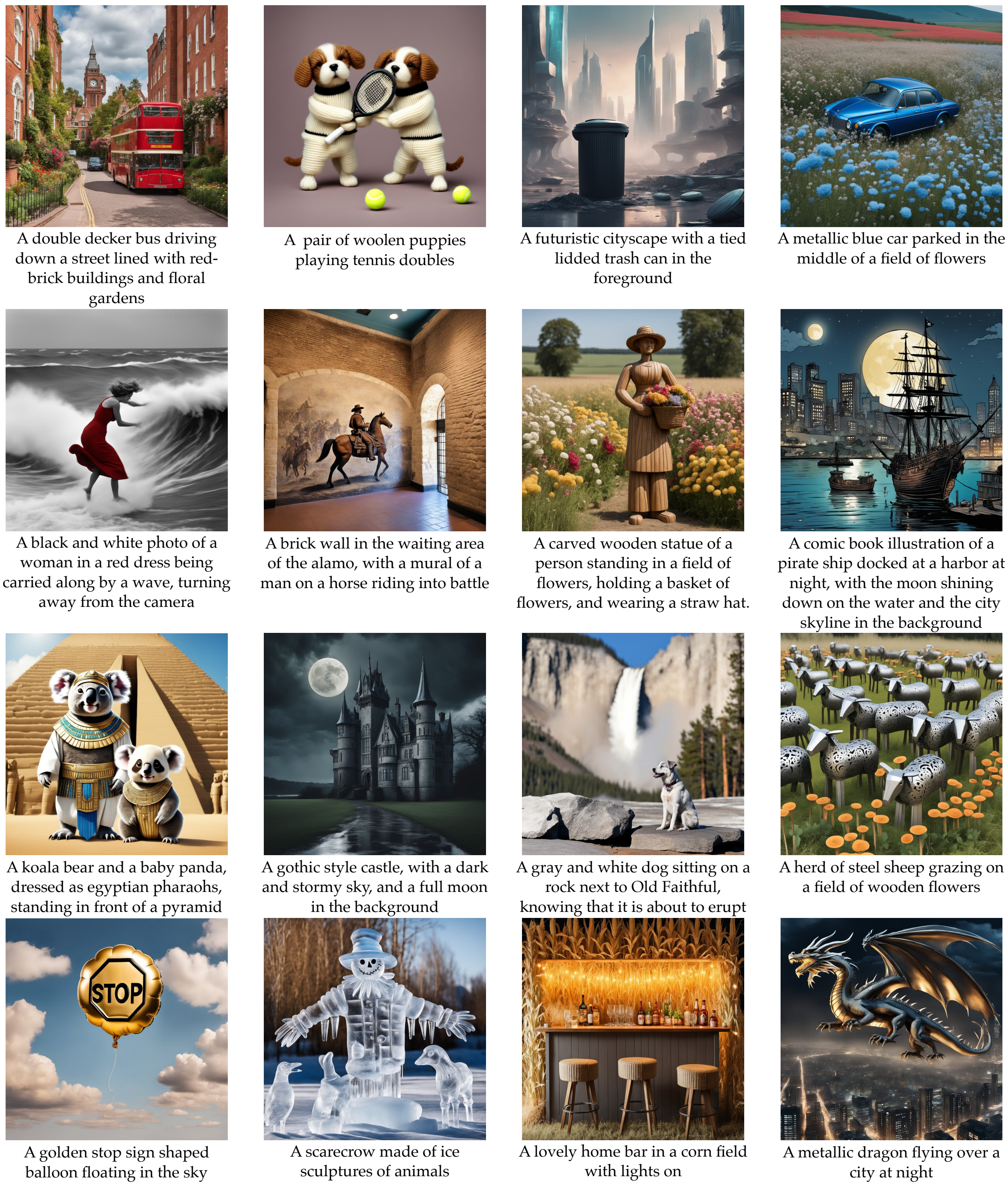


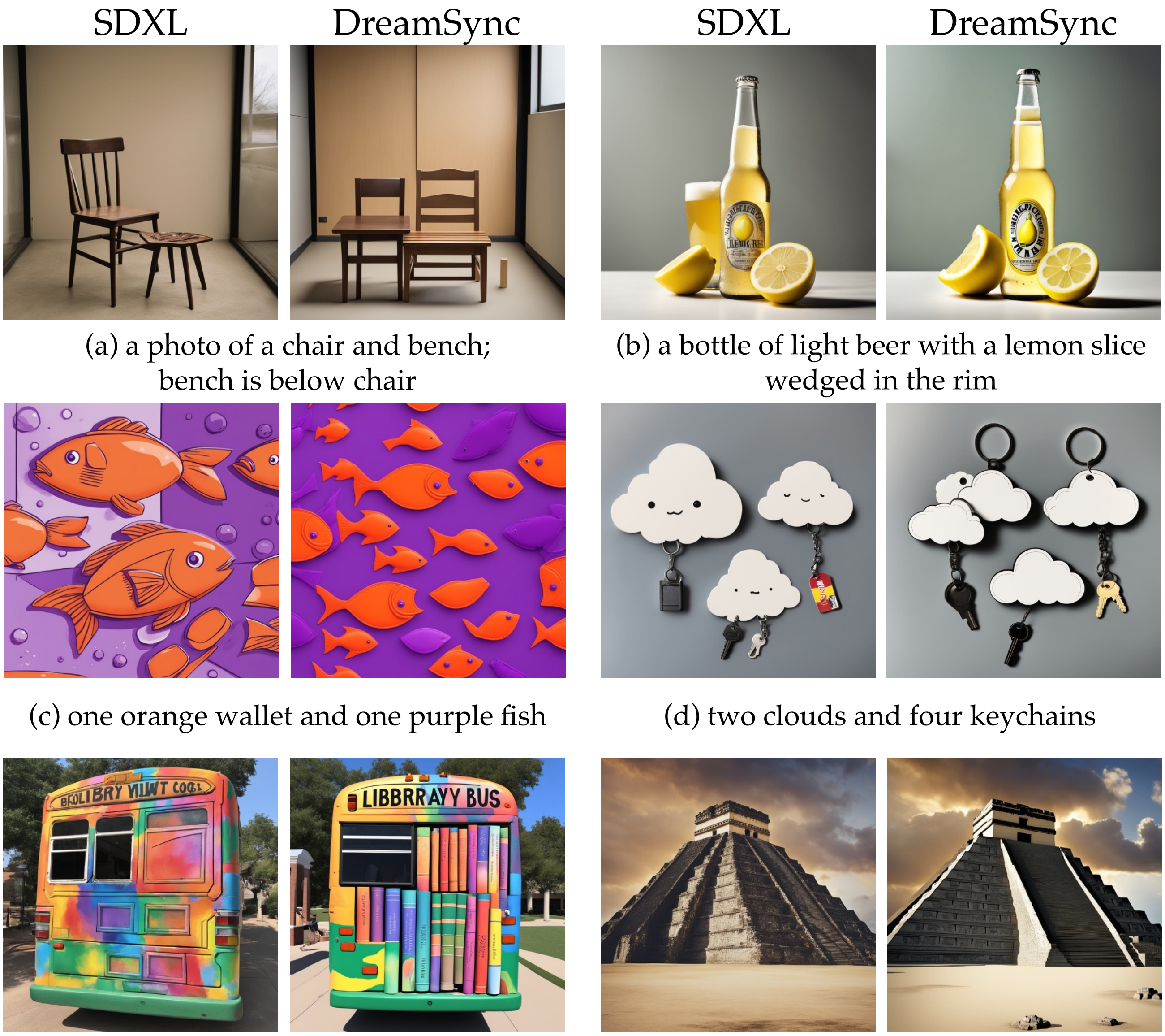
Figure 7. Prompts and Images Generated via DreamSync for Finetuning. Prompts generated by PaLM-2 exhibit high diversity and corresponding synthetic images exhibit high quality and alignment.

B. Failure Modes Analysis

Figure 8 presents several side-by-side examples showcasing common failure modes of DreamSync. For each example, we show the image generated by SDXL on the left, and the image of SDXL + DreamSync on the right. We also indicate some key directions for improvements.

- **Composing multiple objects and attributes is still challenging.** As shown in (a), (b), (c), and (d), SDXL + DreamSync struggles to produce an image that is faithful to the prompt. In (a), DreamSync adds a bench in the image. However, the attributes of chairs and benches are mixed. In (b), DreamSync removes the extra glass in the background, but neither model is able to place the lemon wedge in the rim of the bottom. In (c), DreamSync adds purple fish in the image, but the counting is not correct. In (d), DreamSync produces four objects but they are cloud-keychain combinations.
- **We observe decline of texture details and shadows on some images.** In (e), the alignment between the text and the bus significantly improves. However, the quality of the bus shadow declines. In (f), both images align well with the text. The main difference is in the details of the temple facade. Notice that for most images we observe DreamSync yields images with high quality and visual appeal, as illustrated in Appendix A.3.

Future work may explore if these challenges can be addressed by following extensions to DreamSync: (1) DreamSync could be used in tandem with RL-based method and training-free method to further improve text-to-image faithfulness; (2) prompt engineering methods in DALL-E 3 [2] may help rewriting challenging prompts into simpler ones for models to synthesize; (3) scaling up DreamSync with a more diverse set of prompts and reward models; (4) mixing DreamSync with large-scale pre-training on real images. In summary, as discussed in §6.1, there is still plenty of headroom to improve.



(a) a photo of a chair and bench; bench is below chair

(b) a bottle of light beer with a lemon slice wedged in the rim

(c) one orange wallet and one purple fish

(d) two clouds and four keychains

(e) The library bus has many colorful books painted on it.

(f) El Castillo, A Mayan temple is in the desert.

Figure 8. **Failure modes.** We present qualitative examples of DreamSync failures. First, **it remains challenging to compose multiple objects and bind the attributes correctly**, as shown in (a), (b), (c), and (d). Second, we observe that **the quality of details and shadows decline on some images**, as illustrated in (e) and (f). Overall, SDXL + DreamSync still has room for improvement in terms of text-to-image faithfulness and quality.

C. DSG and Human Rating Evaluation

Details of DSG-1k benchmark. Tab. 6 presents the data sources, quantity, and examples for DSG-1k. Fig. 9 summarizes the 4 broad and 14 detailed semantic categories covered in the benchmark.

Like TIFA, DSG [7] falls into the Question Generation / Answering (QG/A) alignment evaluation framework. Unlike TIFA, DSG introduces a linguistically motivated [8–10] question generation module to ensure the questions generated to hold 4 reliability traits: a) *atomic*: only queries about 1 semantic detail, for unambiguous interpretation; b) *unique*: no duplicated questions; c) *dependency-aware*: prevent invalid queries to VQA/human answerers, e.g. if the answer to a parent question "is there a bike?" is negative, then the child question "is the bike blue?" will not be queried; d) *full semantic coverage*: dovetailing the semantic content of a prompt, no more no less. DSG is powered by a large variant of PaLM 2 [1] for QG and the SoTA VQA module PaLI [5] for QA. For our evaluation task, we adopt DSG-1k (DSG's 1,060 benchmark prompt set) which covers a balanced set of diverse semantic categories and writing styles – including 4 broad categories (e.g. entity/attribute/etc.) and 14 detailed categories (e.g. color/counting/texture/etc.).

Human QA protocol. For human evaluation, we elicit 3 rating responses per prompt/question set (with ∼8 questions per set on average, and a total of 8183 questions). Fig. 10 exemplifies the UI the human raters see. Fig. 11 presents the annotation instructions used to guide the raters. The inner-annotator agreement for this study is 0.684. While the raters respond with `YES/NO/UNSURE`, we find it to be practically useful to numerically convert the answers – 1.0 point for `YES`, 0 for `NO`, and 0.5 for `UNSURE` as partial credit, with the justification that if a semantic detail *can potentially* be grounded in an image yet not necessarily so (e.g. "does this man dress like an engineer?" image: a male in a plain shirt; "is this a cat" image: a blob that *may* be interpreted as a cat), partial credit is fair for not completely failing.

Feature	**Source**	**Sample**	**Example**
Assorted categories	TIFA160 [20]	160	"A Christmas tree with lights and teddy bear"
Paragraph-type captions	Stanford paragraphs [25]	100	"There is a cat in the shelf. Under the shelf are two small silver barbels. On the shelf are also DVD players and radio. Beside the shelf is a big bottle of white in a wooden case."
	Localized Narratives [38]	100	"In this picture I can see food items on the plate, which is on the surface. At the top right corner of the image those are looking like fingers of a person."
Counting	CountBench [34]	100	"The view of the nine leftmost moai at Ahu Tongariki on Easter Island"
Relations	VRD [31]	100	"person at table. person has face. person wear shirt. person wear shirt. chair next to table. shirt on person. person wear glasses. person hold phone"
Written by T2I *real* users	DiffusionDB [51]	100	"a painting of a huangshan, a matte painting by marc simonetti, deviantart, fantasy art, apocalypse landscape, matte painting, apocalypse art"
	Midjourney-prompts [50]	100	"furry caterpillar, pupa, screaming evil face, demon, fangs, red hands, horror, 3 dimensional, delicate, sharp, lifelike, photorealistic, deformed, wet, shiny, slimy"
Human poses	PoseScript [11]	100	"subject is squatting, torso is leaning to the left, left arm is holding up subject, right arm is straight forward, head is leaning left looking forward"
Commonsense-defying	Whoops [3]	100	"A man riding a jet ski through the desert"
Text rendering	DrawText-Creative [30]	100	"a painting of a landscape, with a handwritten note that says 'this painting was not painted by me'"

Table 6. **DSG-1k overview.** To comprehensively evaluate T2I models, DSG-1k provides 1,060 prompts covering diverse skills and writing styles sampled from different datasets.

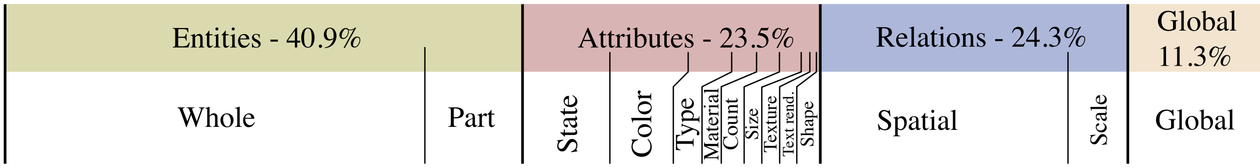


Figure 9. Semantic categories contained in DSG. ***Entity***: `whole` (entire entity, e.g., *chair*), `part` (part of entity, e.g., *back of chair*). ***Attribute***: `color` (e.g., *red book*), `type` (e.g., *aviator goggles*), `material` (e.g., *wooden chair*), `count` (e.g., *5 geese*), `texture` (e.g., *rough surface*), `text rendering` (e.g., letters "Macaroni"), `shape` (e.g., *triangle block*), `size` (e.g., *large fence*). ***Relation***: `spatial` (e.g., *A next to B*); `action` (*A kicks B*). ***Global*** (e.g., *bright lighting*).

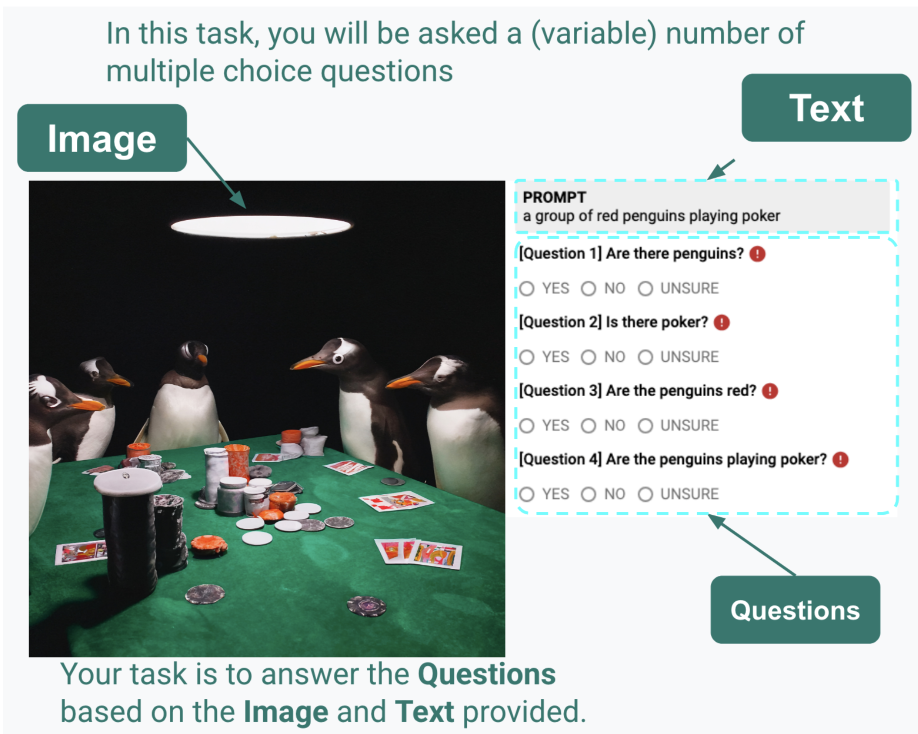


Figure 10. Annotated example of DSG human evaluation query.

INSTRUCTION
Given an image, a question, and a set of choices, choose the correct choice according to the image content.
All the questions are formulated as binary: ''YES'' / ''NO'' with an additional option ''UNSURE''
Select ''UNSURE'' if you think the image does not provide enough information for you to answer the question.

NOTES
- Some images may be of low quality. In such cases, please just select the choice according to your intuition. For ambiguous cases, for example, the question is ''is there a man?'', and the image contains a human but it is unclear whether the human is a man, answer ''no''.
- If a question assumes something incorrect, select ''UNSURE''.

Figure 11. Summary of the human annotation instruction for DSG-1k QA.

CVPR
#***

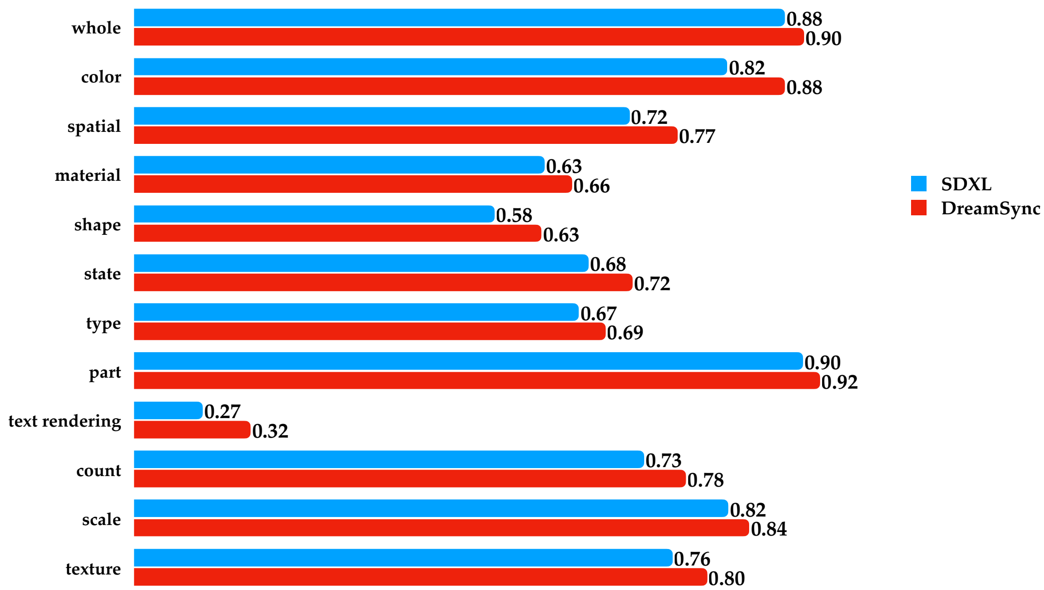


Figure 12. Detailed Human Evaluation Results on DSG-1K. By applying DreamSync upon SDXL, the human evaluation of alignments improved on all categories.

Detailed Human Evaluation Results on DSG-1K. We present detailed evaluations on DSG-1K by semantic categories listed in Figure 9. The results are shown in Figure 12. By applying DreamSync upon SDXL, the human evaluation on alignments improved on all categories.

Single-Question Human Evaluation. Besides the large-scale human annotation, we also did a light-weight single-question

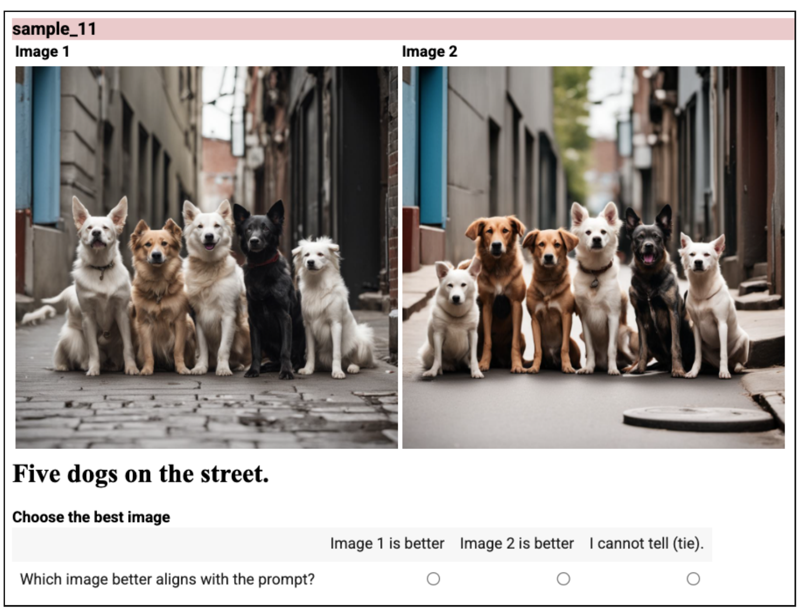


Figure 13. Example human rater screen for the author raters. We display two image side-by-side, where we randomize which is Image 1 vs. Image 2. We ask a single question to the raters, referencing the prompt that is displayed below the images. The raters were given that instructions `‘Rate images based on how many ‘components’ of the prompt are captured in the image. If both images depict every part of the prompt, then you should choose "I can’t tell (tie)." Otherwise, the image with more correct components is better.’’` The raters were also shown four examples, with desired ratings and an explanation for the choices.

human evaluation for text prompt alignment. This study was completed by three of the paper's authors. Although this study yields a quite low inter-annotator agreement, we hope it would provide valuable insights on how to set up human evaluation for measuring textual faithfulness of generated images. For this study, we generated one image with SDXL and DreamSync. See Figure 13 for an example rating screen. We randomized the order of the images and prompts. Three authors were asked `‘‘Which image better aligns with the prompt?’’` They could choose Image 1 is better, Image 2 is better, or that they cannot tell (indicating a tie). We use 200 prompts in total with 100 prompts from TIFA and another 100 from DSG.

As mentioned, the inner-annotator agreement was quite low for this study. Only for 42.5% of the 200 prompts did the human raters all agree in their answers. This is likely due to the fact that it is hard to judge overall prompt alignment directly when given two side-by-side images. Indeed, the majority of prompts led to the raters choosing that they cannot tell which image is better. Using the scoring rules from the DSG study described above (with 1 point going to the model with a direct vote, and with 0.5 going to each model for a tie vote), then we have that DreamSync scores 50.08 while SDXL scores 49.92.

Key Takeaway from Human Evaluation. Comparing the fine-grained large-scale human evaluation and the single-question human evaluation, we encourage researchers who are interested in evaluating the text-image alignment to ask annotators detailed and fine-grained questions. It yields significantly better inter-annotator agreement than asking a general single question about alignment. Our large-scale human evaluation with a better agreement suggests that DreamSync improves the textual faithfulness of SDXL on `DSG-1k`, resonating with our automatic evaluation.

D. Randomly-Sampled SDXL+DreamSync Images

Aside from the failure cases discussed in Figure 8, we would like to showcase more randomly-sampled examples of SDXL and DreamSync. We sample 100 prompts. Among these prompts, Figure 14 shows the examples where the VQA scores of applying DreamSync are significantly different from the base model, SDXL, i.e. the absolute difference of mean score are significantly different: $\left|\mathcal{S}_M(T, G^{\text{DreamSync}}(T)) - \mathcal{S}_M(T, G^{\text{SDXL}}(T))\right| > 0.5$. Meanwhile, Figure 15 presents examples where the DreamSync does not improve the VQA scores upon SDXL.

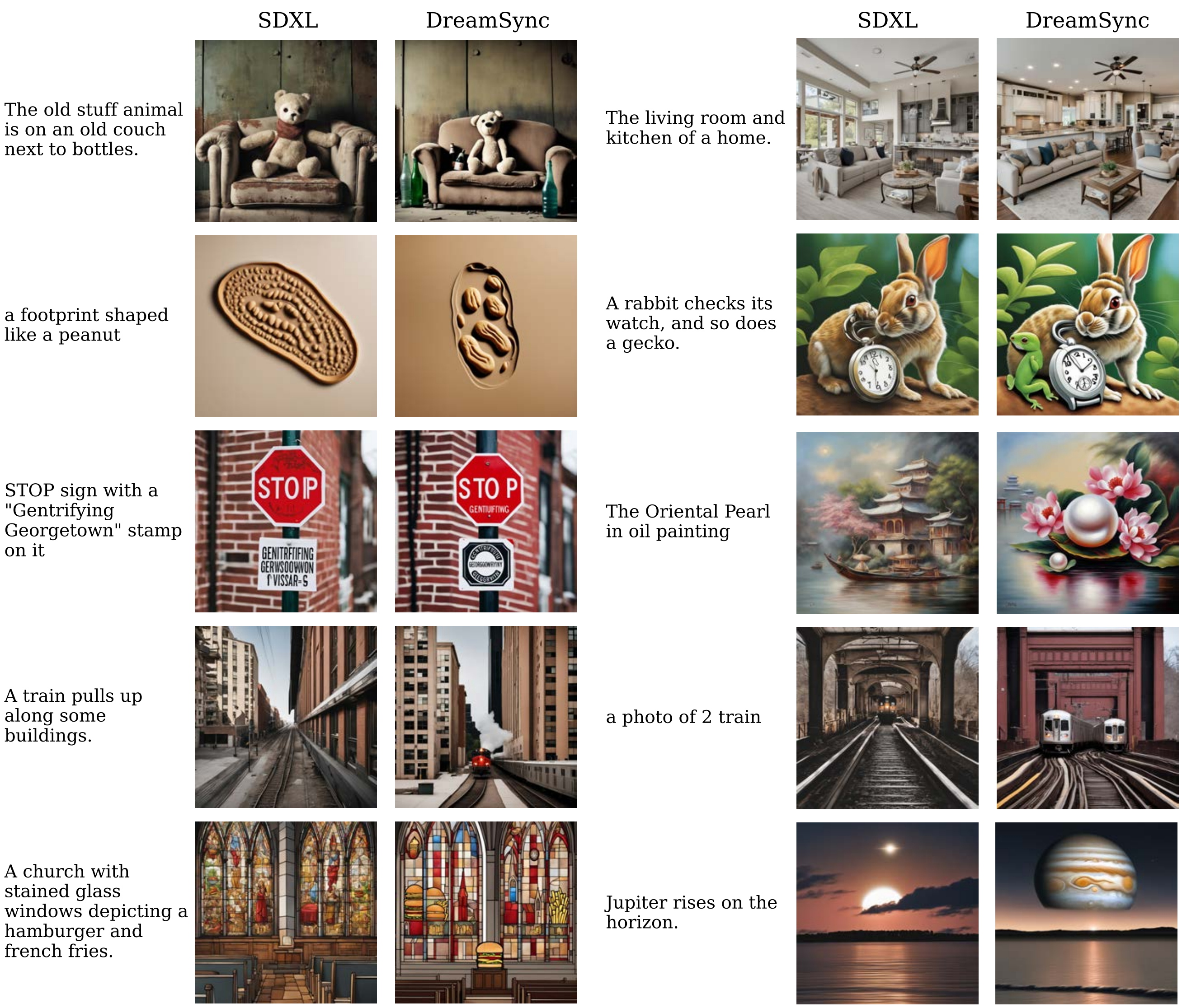


Figure 14. Random samples where DreamSync are significantly different from SDXL. Both models are sampled with the same seed.

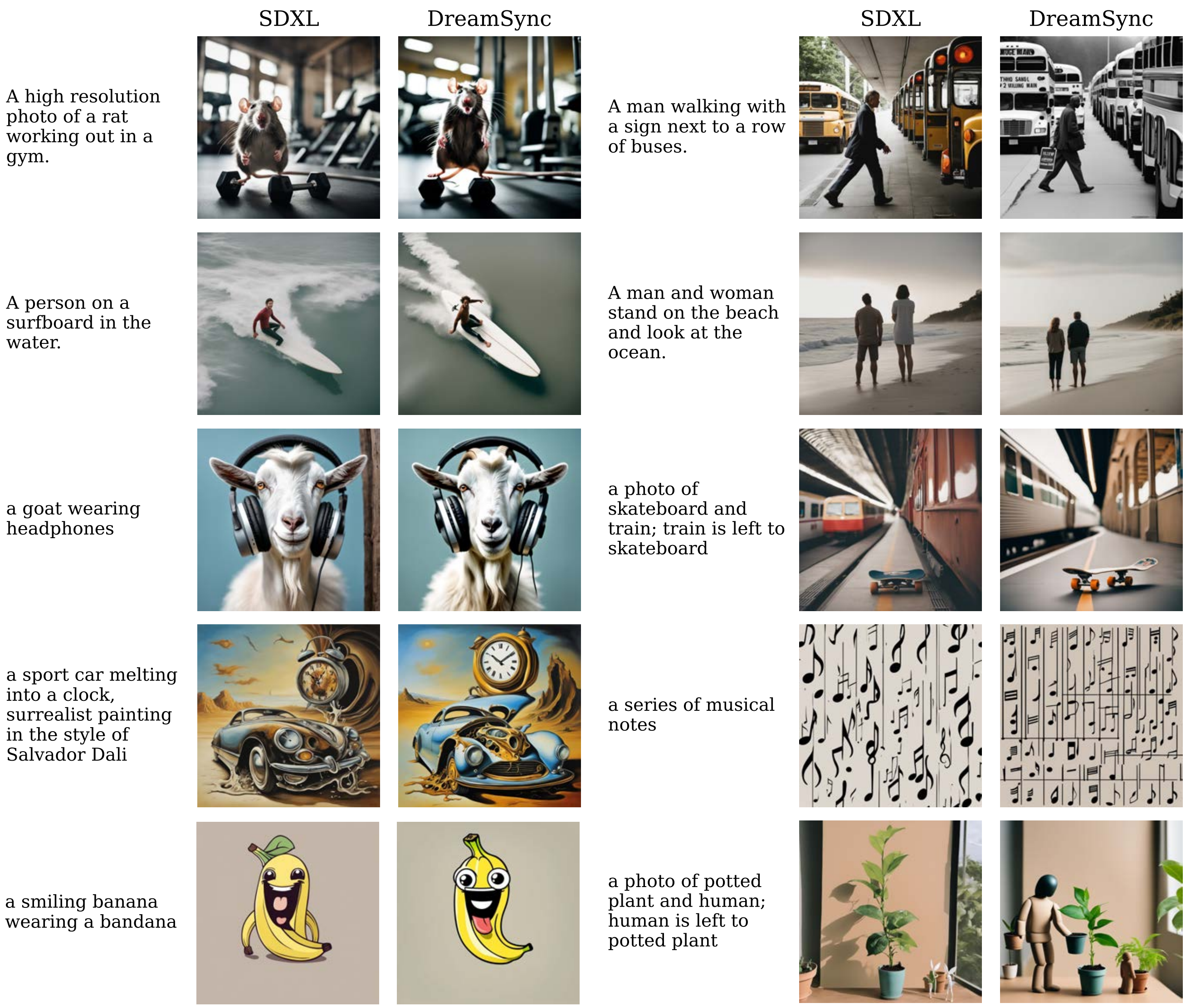


Figure 15. Random samples where DreamSync barely change SDXL's VQA scores. Both models are sampled with the same seed. We hypothesize that because for simple prompts, SDXL is already good enough to compose them.

E. Qualitative Comparison with SD v1.4-based Methods in Table 1

Among the 6 examples shown in Figure 16, DreamSync has 3 absolute successes, wheres SynGen, DDPO and StructureDiffusion each has 2, DPOK has 1 and the base model SD v1.4 has 0 absolute success. These results match well with Table 1.

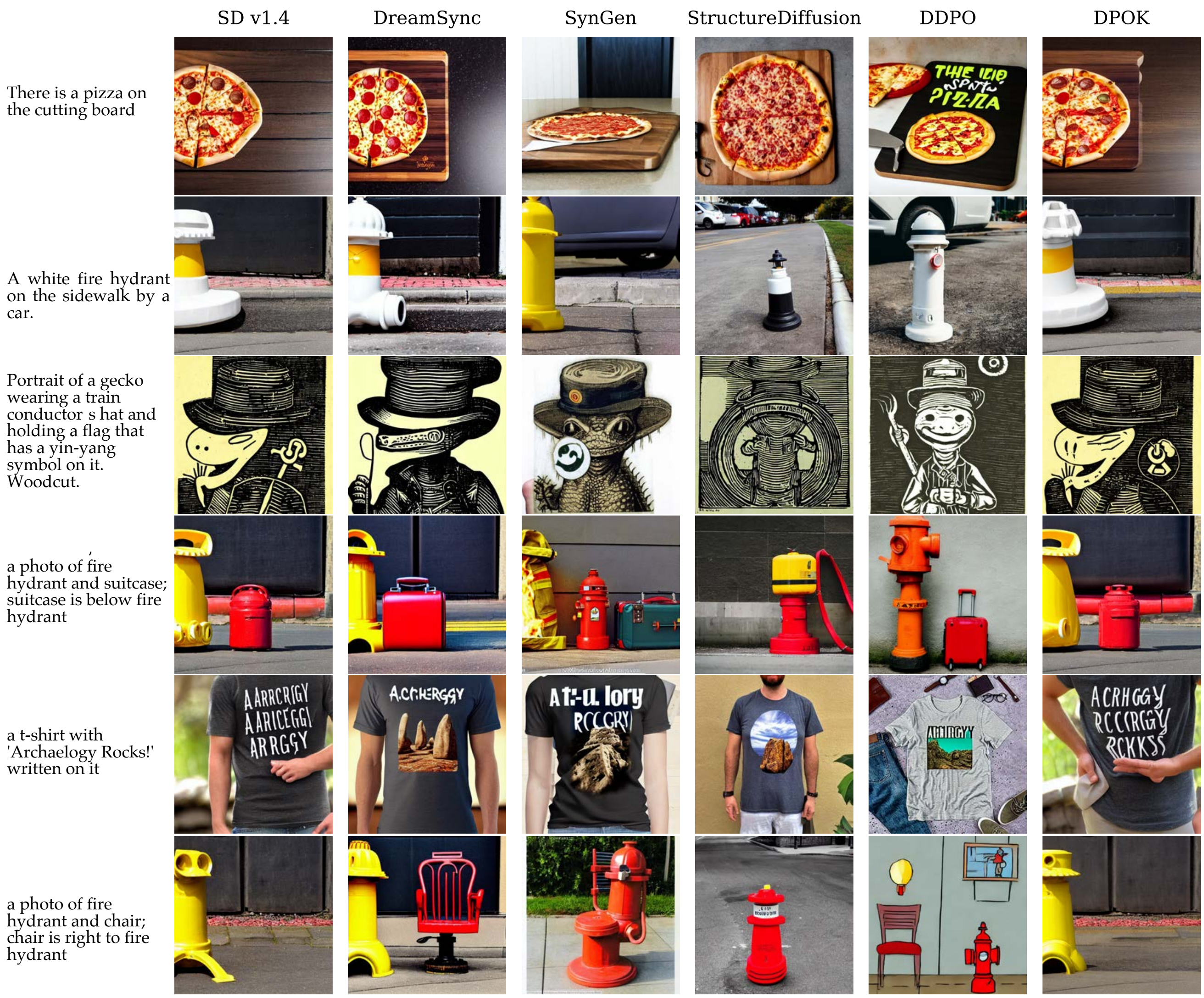


Figure 16. Qualitative Comparison with all models mentioned in Table 1 with base model SD v1.4. Images are generated with the same seed. DreamSync improves the base model's alignment to prompts. Unlike RL-based method (e.g. DDPO), DreamSync does not introduce biases to cartoon. Unlike training-free methods (e.g. SynGen and StructureDiffusion), DreamSync does not degrade image aesthetics.

805
F. Technical Details for Reproducing DreamSync

Sampling	
Number of Inference Steps	50
LoRA α	0.5
Prompts per Iteration	10,000
Images per Prompt	8
Sampling Precision	FP16
Filtering	
θ_{VQA}	0.9
$\theta_{\text{Aesthetics}}$	0.6
Percentage of Prompt-Image Pairs Passing the Filters	20% $\sim$ 30% (see Figure 5)
Selected Prompt-Image Pairs for Fine-tuning	2,000 $\sim$ 3,000
LoRA Fine-tuning	
LoRA Rank	128
Initial Learning Rate	0.0001
Learning Rate Scheduler	Cosine
LR Warmup Steps	0
Batch Size	8
Gradient Accumulation Steps	2
Total Steps	2,500
Resolution	1024 $\times$ 1024
Random Flip	Yes
Mixed Precision	No (i.e. FP32)
GPUs for Training	4 $\times$ NVIDIA A6000
Finetuning Time	$\sim$ 4 Hours

Table 7. Technical Details for Reproducing DreamSync with Base Model SDXL.

Filtering	
θ_{VQA}	0.85
$\theta_{\text{Aesthetics}}$	0.5
LoRA Fine-tuning	
Finetuning Time	$\sim$ 1 Hours

Table 8. Technical Details for Reproducing DreamSync with Base Model SD v1.4. Same Hyper-parameters as Table 7 are omitted.