Prompt-Propose-Verify: A Reliable Hand-Object-Interaction Data Generation Framework using Foundational Models

Gurusha Juneja[1*], **Sukrit Kumar**[2*]
[1]IIT Delhi
[2]BITS Pilani, Hyderabad
ee1190480@iitd.ac.in, ksukrit2001@gmail.com

Abstract

Diffusion models when conditioned on text prompts, generate realistic-looking images with intricate details. But most of these pre-trained models fail to generate accurate images when it comes to human features like hands, teeth etc. We hypothesize that this inability of diffusion models can be overcome through well-annotated good-quality data. In this paper, we look specifically into improving the hand-object-interaction image generation using diffusion models. We collect a well annotated hand-object interaction synthetic dataset curated using *Prompt-Propose-Verify* framework and then finetune a stable diffusion model on it. We evaluate the image-text dataset on qualitative and quantitative metrics like CLIPScore, ImageReward, Fedility, and alignment and show considerably better performance over the current state-of-the-art models.

1 Introduction

Large pre-trained vision and language models have ascended to the forefront of the ever-evolving landscape of deep learning. These models display superhuman capabilities across a spectrum of complex tasks involving image synthesis and natural language processing. The growth of these large scale models is driven in part by availability of large amounts of data to train on available from across the web (Lin et al. 2015; Krishna et al. 2017; Schuhmann et al. 2022). Recent evolution of latent diffusion models have shown that they can also generate high-quality images guided by text prompts in a wide-variety of scenarios (Saharia et al. 2022a)

However, despite their ability to produce photo-realistic human images, these models struggle with accurately generating specific features, such as text, teeth, and human hands (see Table-1), with human hands being amongst the most difficult. Generating high quality hand-object interaction images requires understanding of intricate hand poses and interaction. Popular datasets for training diffusion models (Lin et al. 2015; Schuhmann et al. 2022), lack high-quality hand-object interaction images. Consequently, current state-of-the-art diffusion models encounter limitations in their ability to generate such images.

Furthermore, it has come to light that these models exhibit biases toward certain racial and demographic groups, as highlighted in (Basu, Babu, and Pruthi 2023). This is largely attributed to the bias present in the training data. These issues can be resolved by generating high-quality diverse synthetic data sets.

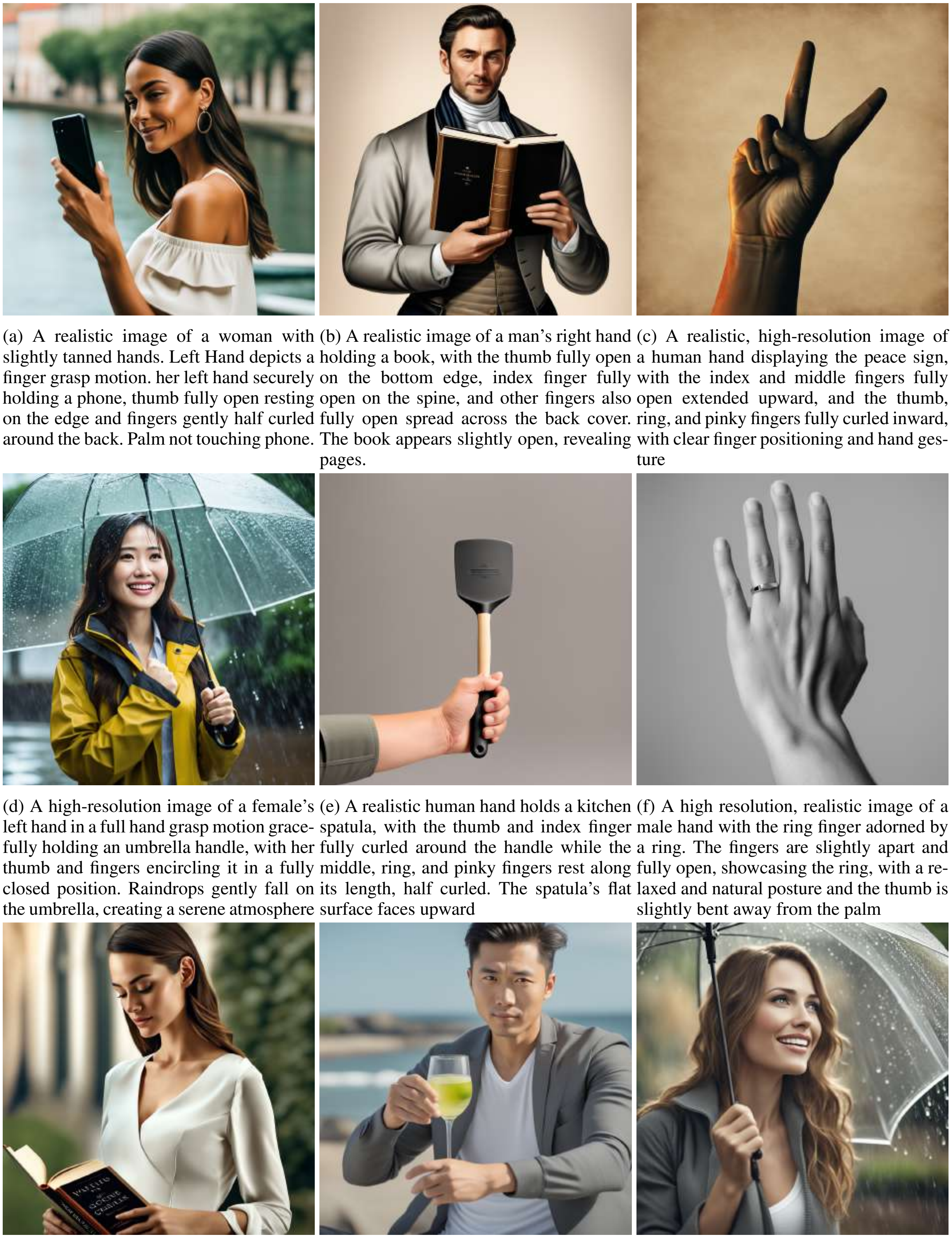
Our contributions: We develop *Prompt-Propose-Verify* - a framework to generate a high-quality well-annotated synthetic Text-Image dataset. Our *Prompt-Propose-Verify* rests on a prompt-propose-verify backbone (see fig-2). Leveraging the world knowledge stored in language models, a *prompter* first generates a detailed prompt. This prompt serves as a conditioning vector to multiple *proposers* to generate a set of proposed images. A *verifier*, trained to accept or reject an image based on the fidelity and alignment of the prompt-image pair, then filters the pairs and augments the dataset with only the most coherent and realistic samples.

We demonstrate the effectiveness of our *Prompt-Propose-Verify* framework by generating a high-quality text-image dataset of hands interacting with various everyday objects : *TactileTango* . We fine-tune a StableDiffusion XL (SDXL) (Podell et al. 2023) model on this generated dataset. We then compare the abilities of the base SDXL (Podell et al. 2023) model and the one fine-tuned on the *TactileTango* dataset on quantitative and qualitative metrics. We outperform the base model in prompts related to hand-object interactions in both quantitative (**3.32%** and **15.9%** improvement over base model in CLIPScore(Hessel et al. 2022) and ImageReward (Xu et al. 2023b) respectively) and qualitative human evaluation (Avg rating **3.8** vs 2.7). We also further show that fine-tuning the model doesn't reduce overall generalization capability of the model by getting similar results to the base models on the DrawBench (Saharia et al. 2022b) prompts.

2 Related Work

Synthetic dataset generation has emerged as a valuable technique in fields where procuring real data is arduous and unfeasible. Generative Adversarial Networks (GANs) have been proven to be invaluable for synthesizing data, particularly in classification tasks domains such as healthcare (Frid-Adar et al. 2018) and environmental conservation (Chatterjee et al. 2022). In addition to classification, generative models have been used to produce pixel-level annotations (Zhang et al. 2021; Li et al. 2022; Xu et al. 2023a; Kim and Hwang

[*] These authors contributed equally.



(a) A realistic image of a woman with slightly tanned hands. Left Hand depicts a finger grasp motion. her left hand securely holding a phone, thumb fully open resting on the edge and fingers gently half curled around the back. Palm not touching phone.

(b) A realistic image of a man's right hand holding a book, with the thumb fully open on the bottom edge, index finger fully open on the spine, and other fingers also fully open spread across the back cover. The book appears slightly open, revealing pages.

(c) A realistic, high-resolution image of a human hand displaying the peace sign, with the index and middle fingers fully open extended upward, and the thumb, ring, and pinky fingers fully curled inward, with clear finger positioning and hand gesture

(d) A high-resolution image of a female's left hand in a full hand grasp motion gracefully holding an umbrella handle, with her thumb and fingers encircling it in a fully closed position. Raindrops gently fall on the umbrella, creating a serene atmosphere

(e) A realistic human hand holds a kitchen spatula, with the thumb and index finger fully curled around the handle while the middle, ring, and pinky fingers rest along its length, half curled. The spatula's flat surface faces upward

(f) A high resolution, realistic image of a male hand with the ring finger adorned by a ring. The fingers are slightly apart and fully open, showcasing the ring, with a relaxed and natural posture and the thumb is slightly bent away from the palm

(g) A woman's right hand tightly grips a large book, with the thumb, index, and middle fingers half closed around the cover, while the ring and little fingers are fully open. The left hand supports the book with all fingers half closed. (realistic, high resolution)

(h) A chinese man holding a glass. One hand supporting and the other grasping, delicately hold a wine glass. The right hand's fingers are fully closed around the glass. The left hand provides support, with all fingers fully closed. (realistic, high resolution, detailed hand positions)

(i) A high-resolution image of a female's right hand in a full hand grasp motion gracefully holding an umbrella handle, with her thumb and fingers encircling it in a fully closed position. Raindrops gently fall on the umbrella, creating a serene atmosphere

Figure 1: Examples of image-prompt pairs in our *TactileTango* dataset

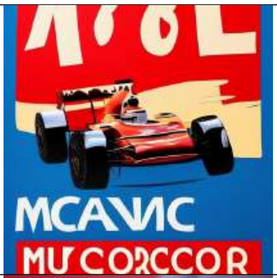
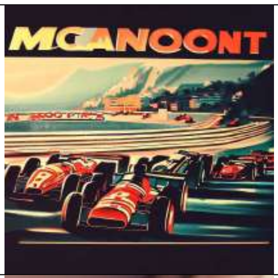
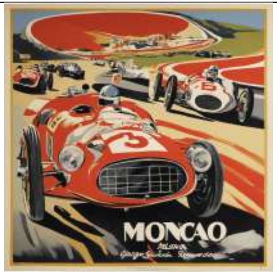
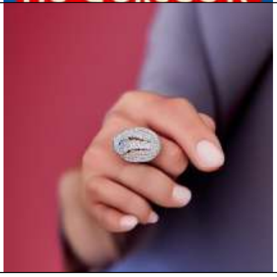
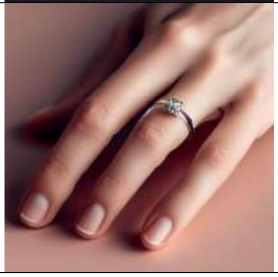
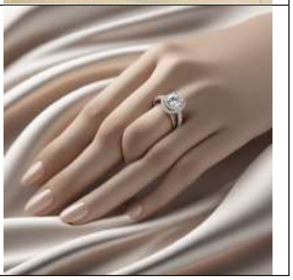
Prompt	Stable Diffusion V1.4	DALL·E 2	StableDiffusionXL
A poster of a car race with the words ”Monaco” on it.			
A human hand with a ring on the ring finger resting on a plain surface			

Table 1: Sample prompts where diffusion models fail to generate good realistic looking-images.

2022).

The recent advancements in diffusion models, particularly in their capacity to generate photo-realistic images (Saharia et al. 2022a), has led to significant progress in enhancing classifiers through the utilization of data generated by these models (Azizi et al. 2023; Trabucco et al. 2023). The ability to condition these model's generations on textual inputs, image editing (Couairon et al. 2022; Yang et al. 2022) and incorporating novel subjects (Ruiz et al. 2022; Gal et al. 2022) has ushered in a new era in the realm of synthetic data, offering precise and fine-grained control over the generation process.

Language models have exhibited remarkable capabilities when processing textual information. Despite achieving super-human performance in various tasks, these models encounter challenges in accomplishing multi-step reasoning tasks. To mitigate this limitation, prior research efforts have introduced the concept of verifiers (Inala et al. 2022; Cobbe et al. 2021), specialized models trained to classify inputs as either valid or invalid. Our approach aims to enhance the reliability of multi-step reasoning processes.

3 Methodology

The proposed framework - *Prompt-Propose-Verify* consists of three stages (see fig - 2)- the prompter Θ, the proposers $\Phi_i(P)$ and the verifier $\zeta(I, P)$. A base prompt p is given to the prompter, in our case GPT-4, which produces a prompt P. The proposers, each of which is a fine-tuned diffusion model produces an image I_i conditioned on input prompt vector. The image prompt pairs $< I, P >$ are then fed into the verifier, parameterized by ζ, that predicts whether the image is accurate and aligns well with the prompt.

3.1 Prompter Θ

Datasets like COCO (Lin et al. 2015) and LAION-5B(Schuhmann et al. 2022) lack intricate pose, structural, and interaction information pertaining to entities like hands. Diffusion models trained on these datasets (Rombach et al. 2022) face challenges in understanding the intricacies of such entities. To address this challenge, we propose to use a prompter to augment the *TactileTango* dataset by introducing detailed textual prompts. The proposer is designed to generate prompts that accentuate the anatomical aspects of hands and the relative spatial positions of hand features concerning objects they interact with.

Pre-trained large language models (LLMs) are trained on extensive internet corpora, and thus encapsulate a wide spectrum of worldly knowledge. This inherent attribute renders them suitable to be used as *prompters* Θ, due to their capability to furnish intricate and comprehensive object descriptions. Specifically, we use GPT-4 to generate detailed prompts P, based on a given base prompt p. To ensure that *TactileTango* covers a wide range of object categories, we prompt GPT-4 to produce texts describing hand-object interactions across a variety of object classes. This acts as the base prompt p. Refer to Table-4 in the appendix for an overview of the distribution of object classes.

Handling possible noise from Language Models Language models quite often hallucinate (Yu et al. 2023; Hanna and Levic 2023; Beutel, Geerits, and Kielstein 2023), this can introduce noise in diffusion models and lower the quality of their generations. To avoid this, we propose to introduce a deterministic component. We define a Domain-Specific-Language (DSL), where each hand position, object position and their interaction can be expressed as a code in DSL. Given the DSL program, we run it through a rule-based program that tells if the position of each finger, palm, object interaction is physically possible given the object shape and

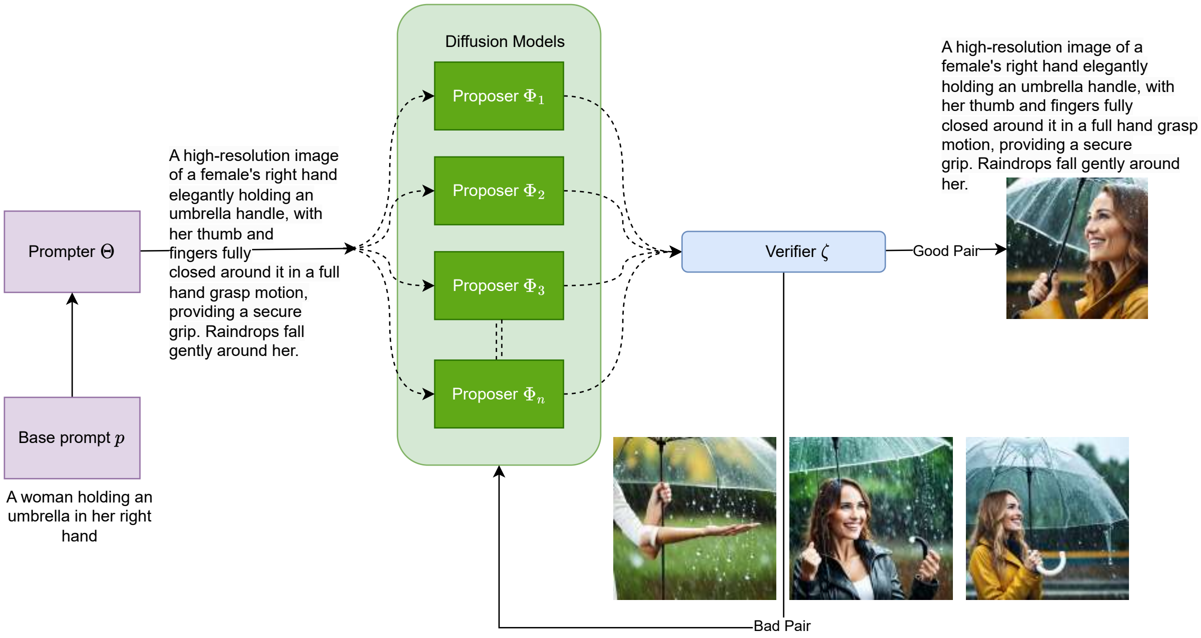

Diffusion Models
Proposer Φ_1
Proposer Φ_2
Proposer Φ_3
Proposer Φ_n
Prompter Θ
Base prompt p
A woman holding an umbrella in her right hand
A high-resolution image of a female's right hand elegantly holding an umbrella handle, with her thumb and fingers fully closed around it in a full hand grasp motion, providing a secure grip. Raindrops fall gently around her.
Verifier ζ
Good Pair
A high-resolution image of a female's right hand elegantly holding an umbrella handle, with her thumb and fingers fully closed around it in a full hand grasp motion, providing a secure grip. Raindrops fall gently around her.
Bad Pair

Figure 2: **Data generation pipeline of *Prompt-Propose-Verify*** , given a base prompt p, the prompter Θ enriches the base prompt using the world knowledge stored in its weights. Each proposer, specialized in a unique specific object, then proposes an image I_i. These image-prompt pairs are then filtered using the verifier ζ based on aesthetics, accuracy, and alignment. The prompt-image pairs accepted by the verifier are added to *TactileTango*

size. We then give this program to GPT-4 to refine it into a language prompt. This step ensures that none of the prompts contain hand positions that are not physically possible. Example of one such DSL program is given in the appendix G.

Ensuring fairness of the generated dataset Previous studies have demonstrated the existence of bias and disproportionate representation in large-scale diffusion models (Luccioni et al. 2023; Basu, Babu, and Pruthi 2023). Deep-learning models tend to amplify bias present in the training data (Hall et al. 2022), to mitigate this issue we ensure that *TactileTango* represents populations from different races, ethnicities, and genders appropriately. We explicitly model the Prompter Θ to generate a versatile array of prompts. The distribution of ethnicities within the dataset is given in Table 4. The meta prompt given to GPT-4 to generate input for diffusion models is given in the appendix.

3.2 Proposer Φ

Diffusion models learn the distribution of data by denoising in a stepwise manner. Given an initial noise map $\epsilon \in \mathcal{N}(0, \mathbf{I})$ conditioned on an input context vector $c = \tau(P)$ where τ is a text encoder and P is the text prompt. Attributing to a lack of data, these models face challenges in generalizing hand images to different and complex scenarios. To curate a dataset with good hand images, it becomes imperative for the generator model to be able to produce realistic hand images with human-like anatomy. One natural choice is to finetune a diffusion model to generate good hands. However, given the scarcity of well-annotated hand datasets of fine quality, we decided to leverage DreamBooth(Ruiz et al. 2022) to finetune a diffusion model in a few-shot manner. DreamBooth(Ruiz et al. 2022) implants a novel subject into the output space of diffusion models using a unique identifier for the subject.

Hands have multiple degrees of freedom and limitless granularity of movement. To ensure the fidelity of images being generated we fine-tune multiple *proposer* Φ models using DreamBooth (Ruiz et al. 2022). Inspired by the classifications made by (Jian et al. 2023), we first classify hand-object interaction into categories based on hand poses. We then fine-tune diffusion models using DreamBooth(Ruiz et al. 2022) for each category. At inference time, given a text prompt P from the prompter, we use these *proposers* to generate images I_i. Each of these paired text-image entities undergoes evaluation by the verifier. Any pair accepted by the verifier is then incorporated into the dataset. If all pairs are rejected by the verifier, the proposers are again prompted to generate fresh batch of images.

3.3 Verifier ζ

Although specialized proposers generate high-fidelity images given the prompt, they may still fail to generalize to novel scenarios (see Fig-3) attributing to the complex nature of hand-object interaction. This invokes the need for a robust verifier that ensures that only high-quality and aligned image-prompt pairs are added to the data set.

The choice of the verifier architecture depends upon the task at hand, owing to the complex nature of hand images, we finetune a ViLT (Kim, Son, and Kim 2021) based verifier, that is trained on a subset of 5000 manually annotated text-images pairs generated by the proposers finetuned using DreamBooth (Ruiz et al. 2022).

4 Experimental Results

In this section we present qualitative and quantitative results on the downstream task performance using the *TactileTango* dataset which was created using our *Prompt-Propose-Verify* framework. A glimpse of the images generated by the finetuned model can be found in the supplementary material. Detailed implementation details can be found in Appendix A.

4.1 Finetuning model using *TactileTango* data

We generate $\sim$ 10k high-quality annotated text-image pairs using *Prompt-Propose-Verify* as described in Sec-3 to curate the dataset - *TactileTango* . We then finetune the Stable Diffusion XL (Podell et al. 2023) model on *TactileTango* using LoRA (Hu et al. 2021).

We compare the performance of finetuned versus base Stable Diffusion XL (Podell et al. 2023) and DreamBooth finetuned models using the CLIPScore (Hessel et al. 2022) and normalized ImageReward (Xu et al. 2023b), we observe that (see Table-2), the CLIPScore for finetuned model is the highest out of the three models and outperforms the Base model by ***3.32%*** and DreamBooth ensembles by ***2.03%***. Similarly, for the ImageReward metric, we outperform the Base model by ***15.90%*** and DreamBooth ensembles by ***34.21%***.

Deep Learning models are susceptible to catastrophic forgetting when trained on fine-grained concepts (Kirkpatrick et al. 2017).To assess the impact of fine-tuning the base model using *TactileTango* on its performance when processing prompts involving objects other than hands, we conduct a comparative analysis of its CLIPScore and ImageReward against the base model on the *DrawBench*(Saharia et al. 2022b) prompt list . Table-2 shows training on *TactileTango* does not yield a significant reduction in CLIPScore or ImageReward thereby indicating that the model does not suffer from catastrophic forgetting or concept drift.

4.2 Human study on images generated by finetuned model

The CLIPScore and ImageReward metrics we used are trained on text-image pairs with limited amount and generic data points, exhibit limitations in effectively distinguishing between good-quality and bad hand image generations. To provide a more robust assessment of the model's performance after fine-tuning on the *TactileTango* dataset, we conducted a qualitative human study involving five participants excluding the authors. Each participant evaluated a total of 30 images across three key dimensions: Fidelity, Prompt Alignment, and an overall rating on a scale ranging from 1 (indicating poor) to 5 (indicating excellent). Of the 12 prompts used, more half included hand-object interactions, while the remainder were randomly selected from the DrawBench prompt list. Detailed information regarding the user study is available in Appendix C. The results of the human evaluation, as depicted in Table 2, consistently favored the fine-tuned model across all three dimensions in comparison to both the base model and the DreamBooth ensembles. In the terms of Fidelity the finetuned model (**3.73**) outperforms the Base model (2.60) and the DreamBooth ensembles (2.86). In the prompt-alignment dimension the finetuned model (**3.73**) outperforms the Base models (2.66) and the DreamBooth ensembles (2.80). Similarly in the overall score the finetuned model (3.8) outperforms the Base model (2.7) and the DreamBooth ensembles (2.7). The model finetuned on the *TactileTango* dataset exhibited comparable performance to the base model for general prompts sourced from DrawBench.

4.3 Ablation Study

To prove the effectiveness of our proposed *Prompt-Propose-Verify* framework, we perform an abalation study. To show the effectiveness of each individual component in the pipeline we generate 3 datasets and fine-tune 3 SDXL models as discussed in Sec 4.1.

To perform the ablation study we generate these 3 datasets, where we remove one out of the 3 components and generate text-image pairs :

1. **Without Prompter**: We fine-tune a model using just the base prompt p not enriched by the Prompter Θ.
2. **Without Proposers**: We fine-tune a model using a dataset where the images are not generated by specialized DreamBooth fine-tuned diffusion Proposers Φ.
3. **Without Verifier**: We fine-tune a model using a dataset where the images generated are not 'verified' by the Verifier ζ

We compute the CLIPScore and ImageReward metrics for all these 3 finet-tuned models and the results are shown in Table 2. We show a significant increase in both the metrics when all 3 components of our framework are used instead of just using 1 particular component.

To show the overall effectiveness of our framework we also fine-tune a diffusion model on a dataset of real hand images however with the non-detailed prompt 'hands'. The model fine-tuned using the *TactileTango* dataset outperforms the generic dataset one. More details on the generic dataset can be found in the Appendix D.

5 Discussion and Future Work

In this work, we introduce *Prompt-Propose-Verify* , a novel approach that streamlines the creation of high-quality, fair, and detailed Text-Image synthetic datasets, circumventing the labor-intensive manual annotation processes associated with traditional datasets like (Lin et al. 2015; Krishna et al. 2017) or the extraction of unstructured data from the internet (Schuhmann et al. 2022). Our framework's efficacy is demonstrated through the generation of a detailed Text-Image dataset focused on hand-object interaction.We finetune a SDXL model on this dataset resulting in notable improvements over the base model across both qualitative and quantitative evaluation metrics. Whilst we focused on text-to-image Diffusion models the generated Text-Image data

	Quantitative		Qualitative		
Model Name	CLIPScore ↑	ImageReward* ↑	Fidelity	Alignment	Overall
Hand Images Prompts					
Base	31.64	0.44	2.60	2.66	2.70
DreamBooth	32.04	0.38	2.86	2.80	2.70
Finetuned ***(Ours)***	**32.69**	**0.51**	**3.73**	**3.73**	**3.8**
w/o Prompter (Ablation)	28.51	0.18	-	-	-
w/o Proposer (Ablation)	31.54	0.50	-	-	-
w/o Verifier (Ablation)	30.90	0.46	-	-	-
Generic hand dataset (Ablation)	31.60	0.35	-	-	-
DrawBench Prompts					
Base	**28.00**	0.61	4.43	4.13	**4.44**
Finetuned *(Ours)*	27.97	**0.64**	**4.46**	**4.26**	4.43

Table 2: Comparison of model finetuned on *TactileTango* with base stable diffusion model and DreamBooth finetuned model. * ImageReward scores are normalized to 0.0 to 1.0 using minimum and maximum values of ImageReward for a given model. Qualitative results for ablations are not shown due to resource unavailability.

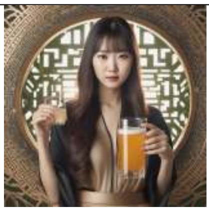
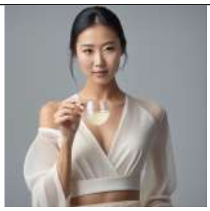
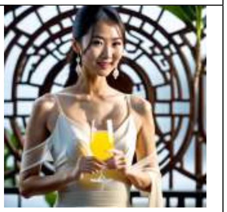
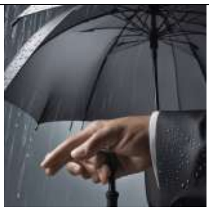
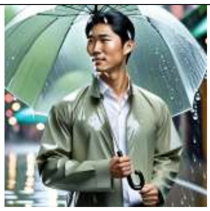
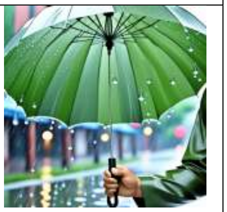
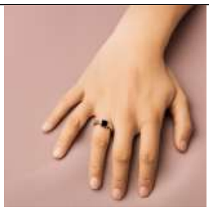
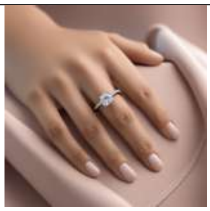
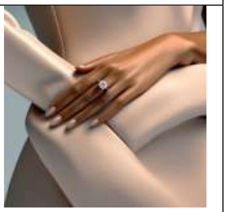
Detailed input Prompts	**Base**	**DreamBooth**	**Finetuned**
An Asian female holding a drink glass. Right hand wraps around the base of a small glass, with all fingers half closed around the object. The left hand supports the glass with all fingers half closed. The glass is in full contact with the right hand's fingers. (realistic, high resolution, clear hand anatomy)			
A realistic 4k image of an asian man's right hand fully wrapping an umbrella, with the thumb half closed and the fingers fully closed around the object. The left hand is not visible. High resolution details of hand and object.			
A realistic, high-resolution image of a tanned female right hand in a supportive gesture, with all fingers fully-open. The tiny ring is positioned at the base of the middle finger, without touching the palm or any other fingers. Rested on a satin cloth. The fabric appears smooth and luxurious.			

Table 3: Comparision of model fine-tuned using *TactileTango* with DreamBooth finetuned and base stable diffusion models.

can be used to improve performance of image captioning and other multi-modal models.

Looking ahead, interesting questions arise: Can we leverage verifier feedback to refine input prompts or further finetune proposer models? Moreover, when extended to accommodate 3D data, could *Prompt-Propose-Verify* be employed to generate supervised data for training for robotic hand-object manipulation tasks? These avenues hold promise for future exploration.

Appendix

A Implementation Details

Fine-tuning SDXL using DreamBooth We fine-tune multiple SDXL models using DreamBooth specialized on generating hands in different poses like holding an object, grasping, open palm, having accessories, making hand gestures etc. We source 5 images from unsplash.com for each pose type and train the model for 500 training steps with an instance prompt and the unique identifier for the class being [sks].Training the DreamBooth model used ∼17 GB of VRAM. We decided to use the SDXL Base 1.0 from StabilityAI as the base model. Instead of fine-tuning a SDXL model using DreamBooth we can use other Diffusion models like DALL.E2, MidJourney and other fine-tuned variants of SDXL which have a high class prior and generate good quality images for that class. We had to fine-tune the proposer due to very poor prior hand generation capabilities of the model. See B for more.

Prompter ϕ We use GPT-4 to generate high quality positive and negative prompts for the DreamBooth tuned model using few-shot learning. We provide a few examples to GPT of good positive and negative prompts using In-Context learning and ask GPT to come up with a base prompt of a hand interacting with everyday objects and refine it. The prompts are then routed to the approriate DreamBooth model based on the content of the prompt.

Training of the Verifier ζ We generated ∼2000 images using the Prompter-Proposer pipeline. Then these images were labelled into Binary classes [Good/Bad] based on the Prompt-Alignment and the quality of the generated hands. We then used a pre-trained ViLT model and add a fully connected layer to network acting as the final output logit classifying each $< I_i, P >$ as either good or bad. The training data was processed using the ViLT processor for the respective model and the model was trained for 10 Epochs with the Adam optimizer with a LR scheduler. Other verifiers can be explored depending upon the dataset being generated, CLIP-Score might be a good filter if the CLIP model embeddings are sensitive to the data to filter out the bad images. Simple CNN based classifiers might also suffice for single pose use-cases. Multiple different verifiers can also be used in cases where we have multiple classes and poses, each proposer could feed to it's verifier.

Fine-tuning SDXL on *TactileTango* using LoRA The SDXL Base model's U-Net and text encoders (SDXL has 2 text encoders) were finetuned on the *TactileTango* dataset for 10 epochs using LoRA (Hu et al. 2021). We used an initial lr=10^{-4} and batch size = 4 and a cosine lr scheduler. The training process took around $\sim$ 9 hours to complete on an A100 GPU and consumed $\sim$ 64GB of VRAM.

Image generation using SDXL We use HuggingFace Diffusers' diffusion pipeline to do inference on the diffusion models. We first feed the prompt to the base SDXL model (or the LoRA (Hu et al. 2021) adapted DreamBooth/Fine-tuned model) and generate the latents. These latents are then fed to the SDXL 1.0 Refiner along with the prompt to generate the final image. We do 100 steps of denoising : 80 for the base model and 20 for the refiner. We follow a guidance scale of 7 and use the DPMSolverMultistepScheduler.

B Distribution of fingers in base SDXL model

We generate 100 different images of hands in various different poses and interacting with different objects using the default 1.0 SDXL model and then the images were manually assigned into bins depending on the perceived number of fingers visible on the screen, another category for bad generation was also added for when the fingers were clearly not visible. Out of the 100 images 97 of them had visible and distinct fingers. We then plotted an histogram to get an idea about the observed frequency of the number of fingers and saw that 6 fingers was the most frequent generation class of the SDXL model.

C User study details

Definition of the dimensions We used 3 dimensions to rate the generated images : All these dimensions were rated from 1-5 where 1 was Bad and 5 was excellent.
Fidelity : How good does the generated image look ? Does it look aesthetically pleasing ? Is the hand-object interaction correct and images look grounded in reality ?
Prompt-Alignment: Does the generated image follow what was asked in the prompt ?
Overall Score : Rating of the image based on overall quality and look.

We randomly selected 6 different prompts and generated 1 image each from 3 Models (Base, DreamBooth, FT-SDXL).We also selected 6 Random prompts from the DrawBench prompts list (Saharia et al. 2022b)DrawBench prompts list and generated 1 image each from the 2 Models (Base, FT-SDXL). Then a form was created using Microsoft Forms which had the Image, Prompt and the rating areas across the three dimensions. The order of the questions was randomized for each participant and the participant did not source model of the generated images. We then computed the average rating across each dimension for each of the groups by model and prompt type.

D Training Dataset

A few labeled Text-Image pairs are shown in fig 1. A few cases of poor-hand generation by DreamBooth models are shown in fig 3. The breakup of the dataset is given in Table 4. Additional examples of the images generated in the *TactileTango* dataset are attached with the supplementary material.

	Race					
Object Type	Light skinned	Dark Skinned	Asian	Indian	Latin	Total
Kitchen objects	100	200	200	200	200	900
Sports Objects	100	100	200	200	200	800
Electronics	100	100	200	200	200	800
Musical Instruments	200	200	200	200	300	1100
Hardware tools	200	200	200	200	200	1000
Art supplies	100	100	100	100	100	500
Medical Instruments	100	100	100	100	100	500
Gardening tools	100	100	100	100	100	500
Vehicle Interior	100	100	100	100	100	500
Straight hand	100	200	300	200	200	1000
Household supplies	100	100	100	100	100	500
Office supplies	100	100	100	100	100	500
Miscellaneous	300	300	300	300	300	1500
Total	1700	1900	2300	2100	2100	10100

Table 4: Breakup of the generated *TactileTango* dataset by Race and Object type

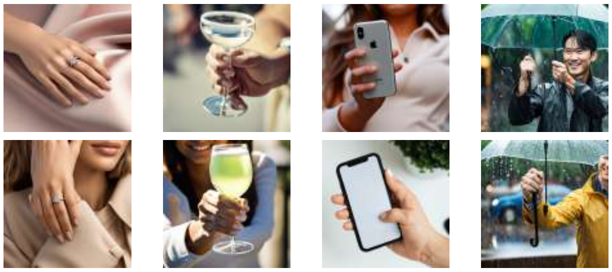


Figure 3: Examples of poor hands being generated by the DreamBooth tuned models.

The generic dataset used to fine-tune the SDXL model was the camenduru/hands dataset on HuggingFace.

E Failure Modes of framework

In this section we highlight a few failure modes of the *Prompt-Propose-Verify* . The proposer might not exactly follow the prompt exactly or might give disproportionate weight to a few tokens in the initial Prompt P, so the generation might not follow the prompt exactly. This issue can be fixed by giving a higher guidance scale during generation or by using something like ReFL (Xu et al. 2023b) or by creating a better Verifier ζ. The Verifier also might not be able to work well if we have multiple poses and occlusion in the generated images. This can be resolved by creating specific verifiers for proposers so that each verifier is specialized on a generation type and can distinguish between good or bad generations easily.

The output and generation quality of are framework is also heavily reliant on the effectiveness of the backbone Proposer and Prompter networks used. The output quality of the generated images will vary according to the base diffusion model used. The quality and details of the text pairs corresponding to the images will also differ based on the LM being used.

F Meta-Prompt for GPT-4

The meta-prompt used for generating the prompts and the negative prompts from GPT-4 for the *TactileTango* dataset are given below.

First, for generating the program.

> You are given a base sentence for hand-object-interaction. You have to write a detailed program building upon the base sentence. The program contains information about the hands, finger positions and object. Write the program based on the posssible physical positions of the hand, do not write something that cannot be physically executed. Following is the syntax of the program.
>
> Fully closed meaning finger tips touching palm. Half closed meaning wrapped around the object. Fully open meaning straight.
>
> [Begin]
>
> Right Hand:
>
> Motion Type: Full Finger Grasp / Full Finger Wrap / Finger Tip Grasp / Support / Lever Grasp / Press / Two Finger Grasp / Three Finger Grasp
>
> Thumb: Fully Open / Half Closed / Fully Closed
>
> Index Finger: Fully Open / Half Closed / Fully Closed
>
> Middle Finger: Fully Open / Half Closed / Fully Closed
>
> Ring Finger: Fully Open / Half Closed / Fully Closed
>
> Little Finger: Fully Open / Half Closed / Fully Closed
>
> Left Hand:
>
> Motion Type: Full Finger Grasp / Full Finger Wrap / Finger Tip Grasp / Support / Lever Grasp / Press / Two Finger Grasp / Three Finger Grasp
>
> Thumb: Fully Open / Half Closed / Fully Closed
>
> Index Finger: Fully Open / Half Closed / Fully Closed

Middle Finger: Fully Open / Half Closed / Fully Closed

Ring Finger: Fully Open / Half Closed / Fully Closed

Little Finger: Fully Open / Half Closed / Fully Closed

Object:

Object Name: Name Of Object

Object Size wrt Hand : Tiny / Small / Size Of Palm / Larger Than Palm

Position wrt Palm: Fully Touching Palm / Not Touching Palm / Partially Touching Palm

Contact with Thumb: Full Thumb/No Thumb/Tip of Thumb/Base of Thumb

Contact with Index Finger: Full Finger/No Finger/Tip of Finger/Base of Finger

Contact with Middle Finger: Full Finger/No Finger/Tip of Finger/Base of Finger

Contact with Ring Finger: Full Finger/No Finger/Tip of Finger/Base of Finger

Contact with Little Finger: Full Finger/No Finger/Tip of Finger/Base of Finger

[End]

Now write a code for the following base sentence: base Think step by step. First describe the position of hands in text and reason about the positoins and their physical satisfiability then write the code at the end.

Second, for generating natural language prompt from program.

You are a prompt designer for a text to image system. Text to image systems do not produce ood hands, you have to write a prompt with your main focus on the hands. 1. You are given a base code which has the core contents of the prompt. You ar ealso given a base prompt that has additional information about the hand and objects. You need to follow that code and base prompt to produce hand object interaction prompt. 2. Write elaborate prompts focusing on the hands. Give the detailed description of finger positions, hand positions, hand gestures, etc. 3. Append terms like realistic, 4k, high resolution to the prompt. You also have to write a negative prompt. The negative prompt would include all the things that we do not want in the image. For example, bad hands, extra fingers, broken fingers, etc.

A few examples are given below: [Positive Prompt: A human hand tightly grips a knife, with the index and middle fingers around the handle half closed, thumb opposite fully open. The ring and pinky fingers curl inward fully closed. The downward-pointing blade gleams in the light.] [Negative Prompt: blurry, disfigured, extra fingers, bad anatomy,too many fingers, cartoon, painting, illus- tration,six fingers,(worst quality, low quality, normal quality:2)] Now, write a positive and negative prompt for the following input: Input code: code Input Prompt: prompt Make sure the prompts that you write are no more than 50 words and the object is clearly mentioned in the prompt. Give the final prompt in brackets as [pos- itive prompt] and [bad prompt].

G Example of DSL Program

Sample output of the DSL program generated :

```
Right_Hand:
- Motion_Type: Support
- Thumb: Fully_Open
- Index_Finger: Fully_Open
- Middle_Finger: Fully_Open
- Ring_Finger: Fully_Open
- Little_Finger: Fully_Open
Left_Hand:
- Motion_Type: Full_Hand_Grasp
- Thumb: Fully_Closed
- Index_Finger: Fully_Closed
- Middle_Finger: Fully_Closed
- Ring_Finger: Fully_Closed
- Little_Finger: Fully_Closed
Object:
- Object_Name: Tea Filled Cup
- Object_Size_wrt_Hand: Size_Of_Palm
- Position_wrt_Palm: Not_Touching_Palm
- Contact_with_Thumb: Full_Thumb
- Contact_with_Index_Finger: Full_Finger
- Contact_with_Middle Finger: Full_Finger
- Contact_with_Ring_Finger: Full_Finger
- Contact_with_Little_Finger: Full_Finger
```

Based on this prompt, the output generation prompt generated by GPT-4 is as follows :

> Right hand is in a supportive position with all fingers fully open. Left hand is in a full grasp motion with all fingers fully curled. The tea filled cup is placed on the right hand. The left hand holds the handle of cup, not touching the palm. The tea gleams in the light with realistic details

References

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic Data from Diffusion Models Improves ImageNet Classification. arXiv:2304.08466.

Basu, A.; Babu, R. V.; and Pruthi, D. 2023. Inspecting the Geographical Representativeness of Images from Text-to-Image Models. arXiv:2305.11080.

Beutel, G.; Geerits, E.; and Kielstein, J. T. 2023. Artificial hallucination: GPT on LSD? *Critical Care*, 27(1): 148.

Chatterjee, S.; Hazra, D.; Byun, Y.-C.; and Kim, Y.-W. 2022. Enhancement of Image Classification Using Transfer Learning and GAN-Based Synthetic Data Augmentation. *Mathematics*, 10(9).

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Couairon, G.; Verbeek, J.; Schwenk, H.; and Cord, M. 2022. DiffEdit: Diffusion-based semantic image editing with mask guidance. arXiv:2210.11427.

Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; and Greenspan, H. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321: 321–331.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. arXiv:2208.01618.

Hall, M.; van der Maaten, L.; Gustafson, L.; Jones, M.; and Adcock, A. 2022. A Systematic Study of Bias Amplification. arXiv:2201.11706.

Hanna, E.; and Levic, A. 2023. Comparative Analysis of Language Models: hallucinations in ChatGPT: Prompt Study.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Inala, J. P.; Wang, C.; Yang, M.; Codas, A.; Encarnación, M.; Lahiri, S. K.; Musuvathi, M.; and Gao, J. 2022. Fault-Aware Neural Code Rankers. arXiv:2206.03865.

Jian, J.; Liu, X.; Li, M.; Hu, R.; and Liu, J. 2023. AffordPose: A Large-scale Dataset of Hand-Object Interactions with Affordance-driven Hand Pose. arXiv:2309.08942.

Kim, J.-H.; and Hwang, Y. 2022. GAN-Based Synthetic Data Augmentation for Infrared Small Target Detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–12.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. arXiv:2102.03334.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123: 32–73.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Barriuso, A.; Fidler, S.; and Torralba, A. 2022. BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations. arXiv:2201.04684.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. arXiv:2303.11408.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022a. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022b. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

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.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402.

Trabucco, B.; Doherty, K.; Gurinas, M.; and Salakhutdinov, R. 2023. Effective Data Augmentation With Diffusion Models. arXiv:2302.07944.

Xu, A.; Vasileva, M. I.; Dave, A.; and Seshadri, A. 2023a. HandsOff: Labeled Dataset Generation With No Additional Human Annotations. arXiv:2212.12645.

Xu, J.; Liu, X.; Wu, Y.; Tong, Y.; Li, Q.; Ding, M.; Tang, J.; and Dong, Y. 2023b. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. arXiv:2304.05977.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2022. Paint by Example: Exemplar-based Image Editing with Diffusion Models. arXiv:2211.13227.

Yu, X.; Cheng, H.; Liu, X.; Roth, D.; and Gao, J. 2023. Automatic Hallucination Assessment for Aligned Large Language Models via Transferable Adversarial Attacks. arXiv:2310.12516.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort. In *CVPR*.