SELMA: Learning and Merging Skill-Specific Text-to-Image Experts with Auto-Generated Data

Jialu Li* **Jaemin Cho*** **Yi-Lin Sung** **Jaehong Yoon** **Mohit Bansal**
UNC Chapel Hill
{jialuli, jmincho, ylsung, jhyoon, mbansal}@cs.unc.edu

https://selma-t2i.github.io

Abstract

Recent text-to-image (T2I) generation models have demonstrated impressive capabilities in creating images from text descriptions. However, these T2I generation models often fail to generate images that precisely match the details of the text inputs, such as incorrect spatial relationship or missing objects. In this paper, we introduce **SELMA**: **S**kill-Specific **E**xpert **L**earning and **M**erging with **A**uto-Generated Data, a novel paradigm to improve the faithfulness of T2I models by fine-tuning models on automatically generated, multi-skill image-text datasets, with skill-specific expert learning and merging. First, SELMA leverages an LLM's in-context learning capability to generate multiple datasets of text prompts that can teach different skills, and then generates the images with a T2I model based on the prompts. Next, SELMA adapts the T2I model to the new skills by learning multiple single-skill LoRA (low-rank adaptation) experts followed by expert merging. Our independent expert fine-tuning specializes multiple models for different skills, and expert merging helps build a joint multi-skill T2I model that can generate faithful images given diverse text prompts, while mitigating the knowledge conflict from different datasets. We empirically demonstrate that SELMA significantly improves the semantic alignment and text faithfulness of state-of-the-art T2I diffusion models on multiple benchmarks (+2.1% on TIFA and +6.9% on DSG), human preference metrics (PickScore, ImageReward, and HPS), as well as human evaluation. Moreover, fine-tuning with image-text pairs auto-collected via SELMA shows comparable performance to fine-tuning with ground truth data. Lastly, we show that fine-tuning with images from a weaker T2I model can help improve the generation quality of a stronger T2I model, suggesting promising weak-to-strong generalization in T2I models.

1 Introduction

Text-to-Image (T2I) generation models have shown impressive development in recent years [61; 59; 50; 30; 57; 85; 10]. Although these approaches can generate high-quality images based on textual inputs, they still struggle to capture all semantics in the given textual prompts, such as failing to compose multiple subjects [85; 22; 40] and generate correct spatial relationships [48].

Many recent works have been proposed to tackle these challenges in text-to-image generation, aiming to enhance the faithfulness of T2I models to textual inputs. One line of research focuses on supervised fine-tuning on high-quality image-text datasets with human annotations [18] or image-text pairs with re-captioned text prompts [65; 5], as shown in Fig. 1 (a). Another line of research is based on aligning T2I models with human preference annotations [82; 52; 21; 33; 76], as shown in Fig. 1 (b). Other

[*] equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1217

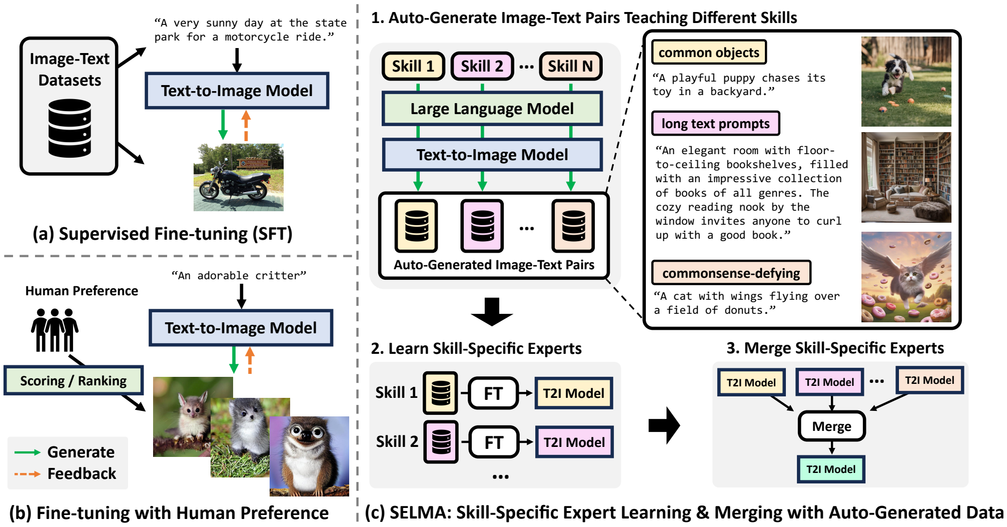


Image-Text Datasets
"A very sunny day at the state park for a motorcycle ride."
Text-to-Image Model
(a) Supervised Fine-tuning (SFT)
Human Preference
"An adorable critter"
Text-to-Image Model
Scoring / Ranking
Generate
Feedback
(b) Fine-tuning with Human Preference
1. Auto-Generate Image-Text Pairs Teaching Different Skills
Skill 1
Skill 2
···
Skill N
Large Language Model
Text-to-Image Model
···
Auto-Generated Image-Text Pairs
common objects
"A playful puppy chases its toy in a backyard."
long text prompts
"An elegant room with floor-to-ceiling bookshelves, filled with an impressive collection of books of all genres. The cozy reading nook by the window invites anyone to curl up with a good book."
commonsense-defying
"A cat with wings flying over a field of donuts."
2. Learn Skill-Specific Experts
Skill 1
FT
T2I Model
Skill 2
FT
T2I Model
···
3. Merge Skill-Specific Experts
T2I Model
T2I Model
···
T2I Model
Merge
T2I Model
(c) SELMA: Skill-Specific Expert Learning & Merging with Auto-Generated Data

Figure 1: Comparison of different fine-tuning paradigms for text-to-image (T2I) generation models. **(a) Supervised Fine-tuning (SFT)**: a T2I model is trained with image-text pairs from existing datasets. **(b) Fine-tuning with Human Preference (*e.g.*, RL/DPO)**: humans annotate their preferences on images by ranking/scoring in terms of text alignments, and a T2I model is trained to maximize the human preference scores. **(c) SELMA**: instead of collecting image-text pairs or human preference annotations, we automatically collect image-text pairs for desired skills with LLM and T2I model, and create a multi-skill T2I model by learning and merging skill-specific expert models.

works focus on introducing additional layouts or object grounding boxes to guide the generation process [37; 81; 84; 22; 16; 89]. Despite achieving significant improvements in aligning generated images with input textual prompts, the success of these approaches relies on the quality of the layouts created from the textual prompts, the collection of high-quality annotations with human efforts, or the existence of large-scale ground truth data, which involves expensive human annotation.

Motivated by LLMs' impressive text generation capability (given open-ended task instructions and in-context examples), and recent T2I models' capability in generating highly realistic photos (based on text prompts), we investigate an interesting question to further improve the faithfulness of state-of-the-art T2I models: "*Can we automatically generate multi-skill image-text datasets with LLMs and T2I models, to effectively and efficiently teach different image generation skills to T2I models?*" In this paper, we propose **SELMA**: **S**kill-Specific **E**xpert **L**earning and **M**erging with **A**uto-Generated Data, a novel paradigm for eliciting the pre-trained knowledge in T2I models for improved faithfulness based on skill-specific learning and merging of experts. SELMA consists of four stages: (1) collecting skill-specific prompts with in-context learning of LLMs, (2) self-generating image-text samples for diverse skills without the need of human annotation nor feedback from reward models, (3) fine-tuning the expert T2I models on these datasets separately, and (4) obtaining the final model by merging experts of each dataset for efficient adaptation to different skills and mitigation of knowledge conflict in joint training. We illustrate the SELMA pipeline in Fig. 1 (c).

In the first and second stages, we use the LLM and the T2I model to generate skill-specific image-text data. The skills include understanding common objects (*e.g.*, puppy in a backyard), handling long prompts (*e.g.*, an elegant room with floor-to-ceiling bookshelves, filled with an impressive collection of books of all genres. The cozy reading nook by the window invites anyone to curl up with a good book."), and displaying commonsense-defying scenes (*e.g.*, cat flying over sky). We aim to teach diverse generation skills to the same T2I model (*i.e.*, self-learning), so that they can handle different types of prompts. To generate image-text pairs for different skills, we first query GPT-3.5 [46] for prompt generation by using only three skill-specific prompts as in-context examples, and filter the generated prompts with ROUGE-L score to maximize prompt diversity to collect 1K prompts in total (Sec. 3.1). Then we use Stable Diffusion models [59; 50] themselves to generate corresponding images from the prompts (Sec. 3.2). We find that our skill-specific training can help mitigate knowledge conflict when jointly learning multiple skills (see Table 2).

In the third and fourth stages, we fine-tune a T2I model with the collected image-text pairs to teach different skills. However, updating the entire model weights can be inefficient; knowledge conflicts within mixed datasets may also lead to suboptimal performance [42]. Thus, in the third stage, we fine-tune T2I models on these self-generated image-text pairs with parameter-efficient LoRA (low-rank adaptation) modules [27] to create skill-specific expert T2I models (Sec. 3.3). In the fourth stage, to build a joint multi-skill T2I model that can have faithful generations across different skills, we merge the skill-specific experts based on LoRA merging [66; 91] (Sec. 3.4).

We validate the usefulness of SELMA with public state-of-the-art T2I models – a family of Stable Diffusion – v1.4 [59], v2 [59], and XL [50] on two text faithfulness evaluation benchmarks (DSG [14] and TIFA [28]), three human preference metrics (Pick-a-Pic [31], ImageReward [82], and HPS [79]), and human evaluation. Empirical results demonstrate that SELMA significantly improves T2I models' faithfulness to input text prompts and achieves higher human preference metrics. Our final LoRA-Merging model achieves 6.9% improvements on DSG, 2.1% improvements on TIFA, and improves the human preference metrics by 0.4 on Pick-a-Pic, 0.39 on ImageReward, and 3.7 on HPS. Furthermore, we empirically show that the T2I models learned from the self-generated images achieve a performance similar to that of learning from ground-truth images (see Fig. 3). Lastly, we further show that fine-tuning with images from a weaker T2I model (*i.e.*, SD v2) can help improve the faithfulness of a stronger T2I model (*i.e.*, SDXL), suggesting promising weak-to-strong generalization in text-to-image models (see Table 3).

2 Related Work

Training Vision-Language Models with Synthetic Images. As recent denoising diffusion models [69; 25] have achieved photorealistic image synthesis capabilities, many works have studied using their synthetic images for training different models. Azizi *et al*. [4], Sariyildiz *et al*. [63], Lei *et al*. [34], inter alia, study training image classification models with synthetic images. For image captioning, Caffagni *et al*. [9] use diffusion models to generate images on the captioning data. For training CLIP [53] models, several works use diffusion models to generate images from existing captions [74] or text generated with language models [23]. There is a recent research direction using synthetic images to train image generation models themselves, and we discuss more details in the following paragraph.

Training Text-to-Image Generation Models with Synthetic Images. A line of recent works train text-to-image (T2I) generation models with synthetic images generated by the same or other models annotated with human preference scores using reinforcement learning [33; 82; 79; 20; 17; 21] or direct preference optimization (DPO) [55; 76]. While these works show promising results in improving model behavior with human preferences, they require expensive human preference annotations. SPIN-Diffusion [87] proposes using self-play [62; 73], which was successfully adopted in Alphago Zero [67] and language models [13; 88], where the model itself becomes a judge and iteratively compares itself with previous iterations. However, self-play algorithm still relies on a set of ground truth image-text pairs as positive examples for supervision. Concurrent/independent to our work, DreamSync [70] trains a T2I model by first creating text prompts with LLMs, sampling multiple images by the T2I model itself, filtering out images with off-the-shelf scorers, and fine-tuning the model on the resulting synthetic image-text pairs [70]. Unlike DreamSync that depends on image filtering (generating 8 images and taking at most one of them for each text prompt, SELMA generates 1 image for each prompt), significantly improving data generation efficiency by using only 2% of image-text pairs compared with DreamSync. Furthermore, we focus on learning multiple skills with T2I models by learning and merging skill-specific LoRA experts to mitigate knowledge interference across different skills, and we show this approach attains much stronger performance without adding any additional inference cost (see Table 2).

3 SELMA: Learning and Merging Text-to-Image Skill-Specific Experts with Auto-Generated Data

We introduce SELMA, a novel framework to teach different skills to a T2I generation model based on auto-generated data and model merging. As illustrated in Fig. 2, SELMA consists of four stages: (1) skill-specific prompt generation with LLM (Sec. 3.1), (2) image generation with T2I Model (Sec. 3.2), (3) skill-specific expert learning (Sec. 3.3), and (4) merging expert models (Sec. 3.4).

https://doi.org/10.52202/079017-1217

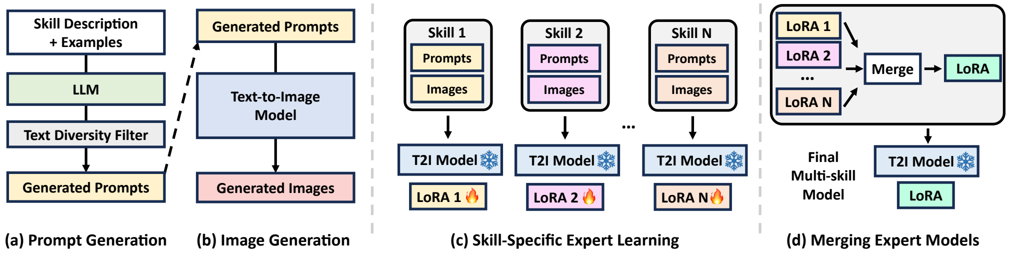


Figure 2: Illustration of the four-stage pipeline of SELMA (Sec. 3).

3.1 Automatic Skill-Specific Prompt Generation with LLM

As shown in Fig. 2 (a), we automatically collect skill-specific prompts (that will be paired with images in Sec. 3.2) to fine-tune T2I models in two steps: (1) using large language models (LLMs) to generate prompts with brief skill descriptions and a few example prompts and (2) filtering the generated prompts to ensure their diversity. In the following, we explain the two steps in detail.

Prompt Generation. We leverage the in-context learning ability of LLMs to generate additional text prompts that follow similar writing styles (*e.g.*, paragraph style) or acquire models' knowledge in the same domain (*e.g.*, count capability). We manually collect three seed prompts with similar writing styles or acquire similar skills (*e.g.*, spatial reasoning) to the target text prompts. Next, we use these seed prompts as in-context learning examples to query GPT-3.5 (`GPT3.5-turbo-instruct`) [46]. We provide additional instructions that encourage diversity in generated prompts, including object occurrences, sentence patterns, and required skills for the T2I model to generate accurate prompts. The detailed prompt template can be found in the Appendix. During prompt generation, we keep expanding the seed prompts with the generated prompts, and always randomly sample three prompts as in-context learning examples from the seed prompts.

Prompt Filtering. To improve the diversity of the collected text prompts, we filter out prompts that are similar to already generated ones. As Taori *et al.* [72] demonstrate that instruction diversity is crucial for improving the instruction following capability of large language models, we follow the same intuition to create diverse text prompts. To ensure the diversity of generated prompts, we first receive a newly generated text prompt from the previous step. Then, we calculate its highest ROUGE-L [38] score with all the previously generated and filtered prompts. Following Taori *et al.* [72], we discard text prompts with ROUGE-L>0.8 to maximize the diversity of generated prompts.

3.2 Automatic Image Generation with Text-to-Image Models

As illustrated in Fig. 2 (b), we generate corresponding images for each generated text prompt using the T2I model. We find that existing diffusion-based T2I models are highly effective in learning from their self-generated images, and even benefit from learning from images generated with weaker T2I models (Table 3). It is important to leverage the knowledge that already exists inside the T2I models (learned from web data during pre-training), and hence we aim to extract this knowledge for creating the skill-specific image-text pairs, and use them to improve T2I models' faithfulness (Sec. 3.3).

3.3 Fine-tuning with Multiple Skill-Specific LoRA Experts

We efficiently adapt the T2I model to different skills by learning skill-specific Low-Rank Adaptation (LoRA) [27] experts. In LoRA fine-tuning, the updates to the original weights $W_0 \in R^{d\times d}$ is decomposed with two low-rank matrices: $W_0 + \Delta W = W_0 + BA$, where $B \in R^{d\times r}$, $A \in R^{r\times d}$ and $r << d$. For each new dataset, we fine-tune the T2I model with LoRA independently, and this introduces $\mathcal{N}$ skill-specific LoRA experts (as shown in Fig. 2 (c)). In Sec. 5.2, we observe that learning and merging skill-specific experts is more effective than learning a single LoRA across all datasets, by helping the T2I model mitigate knowledge conflicts between different skills [42; 11]. However, using multiple skill-specific experts requires the model to know which expert to use for a given input, and this usually requires user annotations on the skill category of inputs. In the next section, we propose to merge skill-specific experts to efficiently construct a single multi-skill model.

https://doi.org/10.52202/079017-1217

3.4 Merging LoRA Expert Models to Obtain a Multi-Skill Model

Recent work of model merging [29; 68; 2; 71; 83] proposes to merge multiple task-specific weights into one, while retaining the original task-specific performances. Moreover, model merging can help mitigate the knowledge conflicts between datasets because we only need to adjust the merging ratios without re-training the task-specific models [86; 56]. Due to these benefits, we extend model merging to learn a final T2I model that can handle multiple skills without knowledge conflicts. Concretely, given $\mathcal{N}$ LoRA experts learned from Sec. 3.3, we merge all LoRA experts into one ($A^{\text{merged}} = \frac{1}{\mathcal{N}} \sum_{n \in \mathcal{N}} A^n$ and $B^{\text{merged}} = \frac{1}{\mathcal{N}} \sum_{n \in \mathcal{N}} B^n$); the resulting single expert can handle all $\mathcal{N}$ skills simultaneously (as shown in Fig. 2 (d)). With this approach, we can reach superior performance over standard multi-task LoRA training and even MoE-LoRA (learning a router with LoRA experts), as shown in Tables 2 and 5, and also eliminate the need to know the skill categories beforehand. Note that while ZipLoRA [66] has demonstrated the use of LoRA merging (merging 2 LoRA modules) in diffusion models, to the best of our knowledge, we are the first to show the effectiveness of LoRA merging on multiple diverse skills (from 5 datasets) in diffusion models.

4 Experimental Setup

4.1 Evaluation Benchmarks

We evaluate models on two evaluation benchmarks that measure the alignment between text prompts and generated images: **DSG** [14] and **TIFA** [28].

DSG consists of 1060 prompts from 10 different sources (160 prompts from TIFA [28], and 100 prompts from each of Localized Narratives [51], DiffusionDB [77], CountBench [47], Whoops [6], DrawText [41], Midjourney [75], Stanford Paragraph [32], VRD [43], PoseScript [19]). Among the ten DSG prompt sources, we mainly experiment with text prompts from five prompt sources that have (1) ground-truth image-text pairs (to compare the usefulness of auto-generated data with ground-truth data) and (2) measuring different skills required in T2I generation (*e.g.*, following long captions, composing infrequent objects). Specifically, we use **COCO** [39] for short prompts with common objects in daily life, **Localized Narratives** [51] for paragraph-style long captions, **DiffusionDB** [77] for human-written prompts that specify many attribute details, **CountBench** [47] for evaluating object counting, and **Whoops** [6] for commonsense-defying text prompts.

TIFA consists of 4,081 prompts from four sources, including COCO [39] for short prompts with common objects, PartiPrompts [85] / DrawBench [61] for challenging image generation skills, and PaintSkills [15] for compositional visual reasoning skills.

4.2 Evaluation Metrics

We quantitatively evaluate the performance of T2I generation models in text faithfulness and human preference metrics. Specifically, to evaluate text faithfulness, we use VQA accuracy from TIFA [28] and DSG [14]. To evaluate human preference score, we use the PickScore [31], ImageReward [82], and HPS [79] See also Sec. 5.6 for human evaluation. Details can be found in Appendix.

4.3 Implementation Details

In the **prompt generation** stage (Sec. 3.1), we use `gpt-3.5-turbo-instruct` [46] to generate text prompts. We collect 1K prompts for each of the five datasets (COCO [39], Localized Narratives [51], DiffusionDB [77], CountBench [47], and Whoops [6]). We refer to the resulting auto-generated datasets as Localized Narrative$^{\text{SELMA}}$, CountBench$^{\text{SELMA}}$, DiffusionDB$^{\text{SELMA}}$, Whoops$^{\text{SELMA}}$, and COCO$^{\text{SELMA}}$, and the resulting combination of 5K auto-generated dataset as DSG$^{\text{SELMA-5K}}$.

In the **image generation** stage (Sec. 3.2), we use the default denoising steps 50 for all models, and the Classifier-Free Guidance (CFG) [26] of 7.5. In the **LoRA fine-tuning** stage (Sec. 3.3), we use 128 as the LoRA rank. **During inference**, we uniformly merge the specialized LoRA experts into one multi-skill expert (Sec. 3.4). More details can be found in Appendix.

Table 1: Comparison of SELMA and different text-to-image alignment methods on text faithfulness and human preference (see Sec. 5.1 for discussion). SELMA achieves the best performance in all five metrics when adapted on different base models (*i.e.*, SD v1.4, SD v2, and SDXL). Best scores for each model are in **bold**.

Base Model	**Methods**	**Text Faithfulness**		**Human Preference on DSG prompts**		
		DSG$^{\text{mPLUG}}$ ↑	TIFA$^{\text{BLIP2}}$ ↑	PickScore ↑	ImageReward ↑	HPS ↑
SD v1.4 [59]	Base model	67.3	76.6	20.3	-0.22	23.0
	(Training-free)					
	SynGen [58]	66.2	76.8	20.4	-0.24	24.5
	StructureDiffusion [22]	67.1	76.5	20.3	-0.14	23.5
	(RL)					
	DPOK [21]	-	76.4	-	-0.26	-
	DDPO [7]	-	76.7	-	-0.08	-
	(Automatic data generation)					
	DreamSync [70]	-	77.6	-	-0.05	-
	SELMA (Ours)	**71.3**	**79.5**	**20.5**	**0.36**	**25.5**
SD v2 [59]	Base model	70.3	79.2	20.8	0.17	24.0
	SELMA (Ours)	**77.7**	**83.2**	**21.3**	**0.72**	**27.5**
SDXL [50]	Base model	73.3	83.5	21.6	0.70	26.2
	DreamSync [70]	-	85.2	-	0.84	-
	SELMA (Ours)	**80.2**	**85.6**	**22.0**	**1.09**	**29.9**

5 Results and Analysis

5.1 Comparison with Different Alignment Methods for Text-to-Image Generation

We compare SELMA with different alignment methods for T2I generation, including training-free methods (SynGen [58], StructureDiffusion [22]), RL-based methods (DPOK [21], DDPO [7]), and DreamSync [70], a concurrent method based on automatic data generation. We experiment with three diffusion-based T2I models (*i.e.*, SD v1.4, SD v2, and SDXL).

SELMA outperforms other alignment methods for T2I generation. As shown in Table 1, SELMA consistently improves faithfulness and human preference metrics for all three backbones. Specifically, on SD v1.4, SELMA improves the baseline by **2.9%** in TIFA, **4.0%** in DSG, **0.2** in PickScore, **0.58** in ImageReward, and **2.5** in HPS score. Furthermore, SELMA achieves significantly higher performance than other baselines, including the RL-based methods (DPOK/DDPO), which require annotated human preference data, and DreamSync, a concurrent/independent work based on a larger auto-generated dataset (*i.e.*, 28K text prompts; SELMA uses 5K text training prompts in total), and image filtering (*i.e.*, generating 8 images and taking at most one of them for each text prompt; SELMA only generates 1 image for each prompt). Besides, on SD v2 and SDXL, SELMA shows larger improvement in text faithfulness (*i.e.*, **7.4%** improvement on DSG for SD v2, and **6.9%** on DSG for SDXL), demonstrating the effectiveness of SELMA.

5.2 Effectiveness of Learning & Merging Skill-Specific Experts

We compare (1) separately learning multiple LoRA experts on different auto-generated datasets followed by merging and (2) training a single LoRA on a mixture of datasets. For this, we experiment with our five auto-generated image-text pairs: Localized Narrative$^{\text{SELMA}}$, CountBench$^{\text{SELMA}}$, DiffusionDB$^{\text{SELMA}}$, Whoops$^{\text{SELMA}}$, and COCO$^{\text{SELMA}}$ (see Sec. 4.3 for details).

Learning & merging skill-specific LoRA experts is more effective than single LoRA on multiple datasets. Table 2 shows that the LoRA models trained separately on each of the five automatically generated datasets (*No.1.* to *No.5.*) can improve the overall metric over the baseline SD v2 – 70.3%, while the degree of improvements is different for each metric (*e.g.*, 76.4% for fine-tuning with Localized Narrative$^{\text{SELMA}}$, and 73.0% for fine-tuning with DiffusionDB$^{\text{SELMA}}$). However, training multiple skills simultaneously with a single LoRA (*No.6.* to *No.7.*) tends to degrade performance as more datasets are incorporated. This indicates that the T2I model struggles with LoRA to accommodate distinct skills and writing styles from different datasets. A similar phenomenon has been reported in LLaVA-MoLE [11], where the knowledge conflict between multiple datasets can degrade the performance. We're the first to show this knowledge conflict across different skills also

Table 2: Comparison of single LoRA and LoRA Merging (see Sec. 5.2 for discussion). We use SD v2 as our base model and train models with our automatically generated image-text pairs. $\texttt{DATA}^{\texttt{SELMA}}$: auto-generated image-text pairs where prompts are generated with LLMs with three prompt examples from `DATA` that are not included in DSG test prompts (see Sec. 4.3 for details). *LN: Localized Narratives; CB: CountBench; DDB: DiffusionDB.* Best/2nd best scores are **bolded**/underlined.

No.	Model	Auto-Generated Training Dataset					Text Faithfulness		Human Preference on DSG		
		LN^{SELMA} *(Paragraph)*	CB^{SELMA} *(Count)*	$\text{DDB}^{\text{SELMA}}$ *(Real Users)*	$\text{Whoops}^{\text{SELMA}}$ *(Counter-Factual)*	$\text{COCO}^{\text{SELMA}}$ *(Common Objects)*	$\text{DSG}^{\text{mPLUG}}$	$\text{TIFA}^{\text{BLIP2}}$	PickScore	ImageReward	HPS
0.	SDv2						70.3	79.2	20.8	0.17	24.0
1.	+ Single LoRA	✓					76.4	81.4	20.9	0.56	26.2
2.			✓				76.0	81.4	20.8	0.46	25.7
3.				✓			73.0	81.2	20.9	0.46	25.8
4.					✓		73.0	80.7	20.8	0.44	25.3
5.						✓	76.0	81.3	20.9	0.47	25.6
6.		✓	✓	✓			75.1	81.5	20.7	0.37	24.8
7.		✓	✓	✓	✓	✓	74.4	80.2	20.6	0.35	24.9
8.	+ LoRA Merging	✓	✓	✓			76.9	82.9	21.2	0.65	27.3
9.		✓	✓	✓	✓	✓	**77.7**	**83.2**	**21.3**	**0.72**	**27.5**

exists in diffusion models. We find that merging multiple skill-specific LoRA experts (*No.8.* and *No.9.*) achieves the best performance in both text faithfulness and human preference, demonstrating that merging LoRA experts can help mitigate the knowledge conflict between multiple skills.

5.3 Effectiveness of Auto-Generated Data

In this section, we investigate the effectiveness of our automatically generated data by comparing them with ground truth data. We fine-tune SD v2 model using ground truth data from Localized Narratives, CountBench, DiffusionDB, Whoops, and COCO, sampling 1K image-text pairs from each dataset and fine-tuning specialized LoRA experts accordingly.

Fine-tuning with auto-generated data can achieve comparable performance to fine-tuning with ground truth data. As shown in Fig. 3, we observe that fine-tuning with either auto-generated or ground truth data improves from baseline SD v2 performance – 70.3%, when evaluated on the DSG benchmark. Surprisingly, fine-tuning with the generated data via SELMA outperforms the use of ground truth data in most cases, leading to a DSG accuracy improvement of **4.0%** with Localized Narrative style prompts, **1.0%** with CountBench style prompts, **1.9%** with DiffusionDB style prompts, and **0.9%** with COCO style prompts. In short, our approach results in an average improvement of 1.2% brought by fine-tuning only auto-generated data without any need for human-collected ground truth text-image pairs, suggesting that diffusion-based text-to-image models may benefit from the diversity of self-generated images. Furthermore, we investigate whether the improvement is brought by text prompt or image quality. We generate images with SD v2 based on 1K ground truth captions, and fine-tune specialized LoRA experts accordingly. We observe that in most cases, using generated images works better than ground truth images (*e.g.*, Localized Narrative), suggesting T2I models can generate images with comparable alignment as ground truth images. Besides, learning from our LLM-generated captions achieves comparable performance with learning from ground truth captions, suggesting the effectiveness of our text prompt collection process. Lastly, we also show in Appendix that fine-tuning with LLaMA3 [1] generated prompts also improve T2I models' faithfulness in generation, demonstrating that our proposed SELMA is compatible to different LLM-based prompt generator.

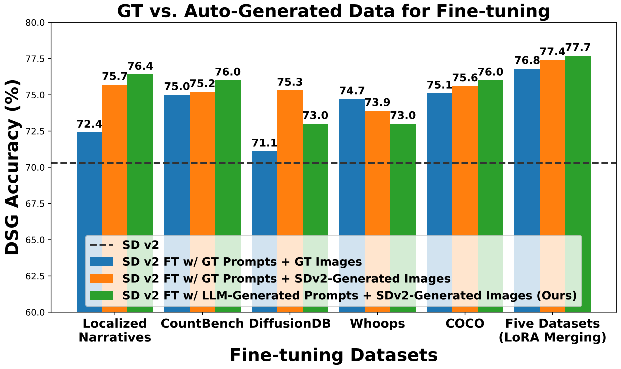


Figure 3: DSG accuracy of SD v2 fine-tuned with different image-text pairs.

Table 3: Comparison of different image generators for creating training images. In addition to using the same model being trained as an image generator, we also experiment with using a smaller model as an image generator (No. 4.). SDXL is bigger/stronger than SD v2. See Sec. 5.4 for discussion.

No.	Base Model	Training Image Generator	Text Faithfulness		Human Preference on DSG		
			$\text{DSG}^{\text{mPLUG}}$ ↑	$\text{TIFA}^{\text{BLIP2}}$ ↑	PickScore ↑	ImageReward ↑	HPS ↑
1.	SD v2	-	70.3	79.2	20.8	0.17	24.0
2.	SDXL	-	73.3	83.5	21.6	0.70	26.2
3.	SD v2	SD v2	77.7	83.2	21.3	0.72	27.5
4.	SDXL	SD v2	**81.3**	83.8	21.5	0.78	28.8
5.	SDXL	SDXL	80.2	**85.6**	**22.0**	**1.09**	**29.9**

5.4 Weak-to-Strong Generalization

In previous experiments, we demonstrate the interesting self-improving capabilities of T2I models, where the training images were generated by the same T2I model. Here, we delve into the following research question: *"Can a T2I model benefit from learning with images generated by a weaker model?"*. The problem of *weak-to-strong* generalization was initially explored in the context of LLMs [8; 60], referred to as superalignment, which involved training GPT-4 [45] using responses generated by a weaker agent, such as GPT-2.

Weaker T2I models can help stronger T2I models. As shown in Table 3, fine-tuning SDXL with generated images from SD v2 (*No.4.*) remarkably enhances performance over the SDXL baseline (*No.2.*) in both text faithfulness and human preference. In addition, this approach achieves competitive performance compared with fine-tuning SDXL with SDXL-generated images (*No.5.*), indicating a promising potential for weak-to-strong generalization in diffusion-based T2I generation models. To the best of our knowledge, this is the first work to find promising improvements in the weak-to-strong generalization for text-to-image diffusion models.

5.5 Comparison with Prompt Generation with LLaMA3

Table 4: DSG and TIFA accuracy of SDXL fine-tuned with prompt data generated with LLaMA3 and GPT-3.5.

Model	Prompt Generator	Image Generator	$\text{DSG}^{\text{mPLUG}}$ ↑
SDXL	-	-	73.3
SDXL	LLaMA3	SDv2	78.0
SDXL	LLaMA3	SDXL	78.6
SDXL	GPT3.5	SDv2	81.3
SDXL	GPT3.5	SDXL	80.2

In this section, we demonstrate that our proposed paradigm is compatible with different prompt generator LLMs. Specifically, we experiment with LLaMA3 (8B) [1], a publicly available open-source LLM and compare the results with GPT-3.5 (`gpt-3.5-turbo-instruct`) based setups described in the main paper. With both LLMs, we generate five sets of skill-specific prompts, and each set contains one thousand skill-specific prompts. As shown in Table 4, we find that fine-tuning SDXL with data generated with LLaMA3 achieves 78.6% on average on DSG, improving the baseline by 5.3%, closing the gap to GPT-3.5 based results. This demonstrates that SELMA is flexible and compatible with different prompt generator LLMs. Besides, we further experiment with fine-tuning SDXL with data generated with both a weaker image generator SDv2 and a weaker prompt generator LLaMA3. Our results show that this model achieves similar performance as the model fine-tuned with images generated with SDXL, demonstrating that weak-to-strong generalization holds with weaker data generators.

5.6 Human Evaluation

In addition to automatic evaluation using text faithfulness benchmarks (DSG and TIFA) and human preference metrics (PickScore, ImageReward, and HPS), we further perform a human evaluation to compare the performance of SDXL and SDXL fine-tuned with SELMA on $\text{DSG}^{\text{SELMA-5K}}$ (details in Sec. 4.3). We randomly select 200 prompts from DSG and ask three annotators to determine "Which image aligns with the caption better?" given the text prompt and generated images from both SDXL and SDXL+SELMA. We provide win/tie/lose options to the annotators, and we report the win *vs.* lose percentage in the following. The user interface, instructions, and the detailed statistics are provided in appendix.

https://doi.org/10.52202/079017-1217

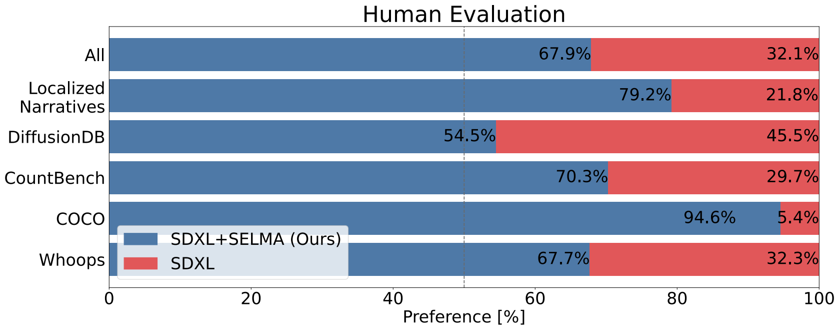


Figure 4: Human Evaluation on 200 sampled text prompts from DSG, where we show the win *vs.* lose percentages of SDXL and SDXL+SELMA (Ours).

SDXL+SELMA is preferred than SDXL in terms of text alignment. Fig. 4 shows that on all five DSG splits, images generated with SDXL+SELMA are preferred **67.9%** of the time, compared to 32.1% for the baseline SDXL. Furthermore, on the five datasets fine-tuned with similar text prompts, SDXL+SELMA achieve a preference rate of **94.6%** on COCO split and **79.2%** on Localized Narratives. This substantial preference over the baseline model demonstrates the effectiveness of SELMA in enhancing T2I models.

Table 5: Comparison with different fine-tuning methods on SD v2 with our auto-generated data, in text faithfulness and human preference. See Sec. 5.7 for discussion.

No.	Methods	Text Faithfulness		Human Preference on DSG		
		DSG$^{\text{mPLUG}}$ ↑	TIFA$^{\text{BLIP2}}$ ↑	PickScore ↑	ImageReward ↑	HPS ↑
0.	SDv2	70.3	79.2	20.8	0.17	24.0
1.	+ LoRA Merging (SELMA)	**77.7**	**83.2**	**21.3**	**0.72**	**27.5**
2.	+ LoRA Merging + DPO	75.1	81.4	20.8	0.44	26.0
3.	+ MoE-LoRA	77.2	83.0	**21.3**	0.68	27.2

5.7 Training Method Ablations

We experiment with various training configurations for SELMA to validate our design choices for fine-tuning. As our current experiments are based on supervised fine-tuning with LoRA Merging, we additionally explore Direct Preference Optimization (DPO) [55; 76] as an alternative to supervised fine-tuning and employing Mixture of Lora Experts (MoE-LoRA) [80] instead of LoRA Merging. See the Appendix for the implementation details. Table 5 demonstrates that while fine-tuning with DPO and MoE-LoRA significantly improves the T2I models' text faithfulness and human preference (*No.2 & 3. vs. No.0.*), simple inference-time LoRA merging achieves the best overall performance. In the end, we adopt LoRA merging and supervised fine-tuning as the default configuration in SELMA for its simplicity and efficiency.

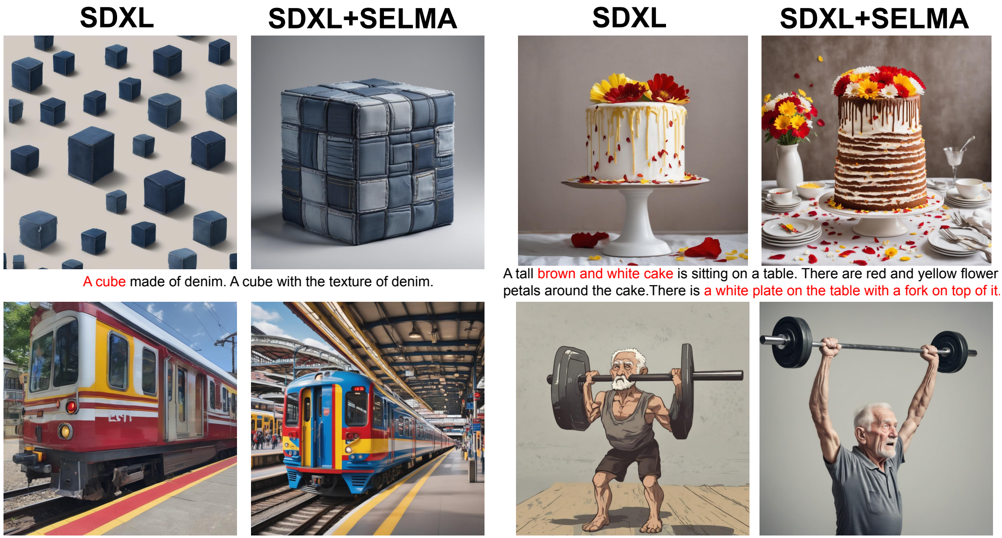
5.8 LoRA Expert Size Ablations

Table 6: Performance of scaling the LoRA ranks on DSG.

Model	LoRA Rank	DSG$^{\text{mPLUG}}$ ↑
Base model (SDv2)	-	70.3
Single LoRA	128	74.4
Single LoRA	256	74.9
Single LoRA	640	71.5
LoRA Merging	128	77.7

In this section, we demonstrate that simply scaling the LoRA size doesn't help mitigate the knowledge conflict between different skills. We experiment with using a single LoRA with different ranks (128 / 256 / 640) and compare them to our default LoRA merging (rank=128). Table 6 shows that increasing the rank of LoRA from 128 to 256 slightly improves the performance (i.e., 74.9 vs. 74.4), but further scaling the rank of the LoRA to 640 significantly drops the performance (i.e., 71.5 vs. 74.4). The performance drop when using LoRA with higher ranks (i.e., rank=640) is similar to the observation in Figure 3 in [24]. This result indicates the effectiveness of our skill-specific learning and merging of LoRA experts.

https://doi.org/10.52202/079017-1217



A cube made of denim. A cube with the texture of denim.

A tall brown and white cake is sitting on a table. There are red and yellow flower petals around the cake.There is a white plate on the table with a fork on top of it.

A train is opening its doors. The train is currently parked at a train station. The train is blue, the doors are red, and it has white stripes on it.There is a long yellow line near the train area.

An old man lifts a barbell above his head.

Figure 5: Example images generated with SDXL and SDXL+SELMA. SELMA shows better performance in object composition, attribute binding, and long text prompt following. We highlight the parts of the prompts in red where SDXL makes errors while SDXL+SELMA generates correctly.

5.9 Qualitative Examples

We show some qualitative examples of images generated with SDXL fine-tuned with SELMA paradigm in Fig. 5. We find that fine-tuning with SELMA improves SDXL's capability in composing infrequently co-occurred attributes (*i.e.*, "cube" and "denim" in the top-left image), composing multiple objects mentioned in the text prompts (*i.e.*, "brown and white cake", "table", "red and yellow flower", and "fork" in the top-right image), following details in the long paragraph-style text prompts (*i.e.*, "blue train with white stripes" and "long yellow line near train area" in the bottom-left image), and generating images that challenge commonsense (*i.e.*, "Old man lifts a barbell" in bottom-right image). These qualitative examples demonstrate the effectiveness of SELMA in improving T2I models' text faithfulness and human preference.

6 Conclusion

We propose SELMA, an novel paradigm to improve state-of-the-art T2I models' faithfulness in generation and human preference by eliciting the pre-trained knowledge of T2I models. SELMA first collects self-generated images given diverse generated text prompts without the need for additional human annotation. Then, SELMA fine-tunes separate LoRA models on different datasets and merges them during inference to mitigate knowledge conflict between datasets. SELMA demonstrates strong empirical results in improving T2I models' faithfulness and alignments to human preference and suggests potential weak-to-strong generalization for diffusion-based T2I models.

Acknowledgement

We thank the reviewers and Elias Stengel-Eskin, Prateek Yadav, and Yushi Hu for the thoughtful discussion. This work was supported by DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL-2112635, DARPA Machine Commonsense (MCS) Grant N66001-19-2-4031, ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, and a Bloomberg Data Science Ph.D. Fellowship. The views contained in this article are those of the authors and not of the funding agency.

https://doi.org/10.52202/079017-1217

References

[1] AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[2] Ainsworth, S.K., Hayase, J., Srinivasa, S.S.: Git re-basin: Merging models modulo permutation symmetries. The International Conference on Learning Representations (ICLR) (2023), https://api.semanticscholar.org/CorpusID:252199400

[3] Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C.K., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M.Y., Suk, H., Zhang, S., Suo, M., Tillet, P., Zhao, X., Wang, E., Zhou, K., Zou, R., Wang, X., Mathews, A., Wen, W., Chanan, G., Wu, P., Chintala, S.: Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. p. 929–947. ASPLOS '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3620665.3640366, https://doi.org/10.1145/3620665.3640366

[4] Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic Data from Diffusion Models Improves ImageNet Classification. TMLR (2023), http://arxiv.org/abs/2304.08466

[5] Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)

[6] Bitton-Guetta, N., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., Schwartz, R.: Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2616–2627 (2023)

[7] Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. The International Conference on Learning Representations (ICLR) (2024)

[8] Burns, C., Izmailov, P., Kirchner, J.H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., et al.: Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. arXiv preprint arXiv:2312.09390 (2023)

[9] Caffagni, D., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R.: Synthcap: Augmenting transformers with synthetic data for image captioning. In: Foresti, G.L., Fusiello, A., Hancock, E. (eds.) Image Analysis and Processing – ICIAP 2023. pp. 112–123. Springer Nature Switzerland, Cham (2023)

[10] Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. ICML (2023)

[11] Chen, S., Jie, Z., Ma, L.: Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. arXiv preprint arXiv:2401.15947 (2024)

[12] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. ICLR (2023)

[13] Chen, Z., Deng, Y., Yuan, H., Ji, K., Gu, Q.: Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models (2024), http://arxiv.org/abs/2401.01335

[14] Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. ICLR (2024)

https://doi.org/10.52202/079017-1217

[15] Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In: ICCV (2023)

[16] Cho, J., Zala, A., Bansal, M.: Visual programming for text-to-image generation and evaluation. In: NeurIPS (2023)

[17] Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly Fine-Tuning Diffusion Models on Differentiable Rewards. In: ICLR (2024), http://arxiv.org/abs/2309.17400

[18] Dai, X., Hou, J., Ma, C.Y., Tsai, S., Wang, J., Wang, R., Zhang, P., Vandenhende, S., Wang, X., Dubey, A., et al.: Emu: Enhancing image generation models using photogenic needles in a haystack. arXiv preprint arXiv:2309.15807 (2023)

[19] Delmas, G., Weinzaepfel, P., Lucas, T., Moreno-Noguer, F., Rogez, G.: Posescript: 3d human poses from natural language. In: European Conference on Computer Vision. pp. 346–362. Springer (2022)

[20] Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. TMLR (2023), http://arxiv.org/abs/2304.06767

[21] Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. In: NeurIPS (2023)

[22] Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. ICLR (2023)

[23] Hammoud, H.A.A.K., Itani, H., Pizzati, F., Torr, P., Bibi, A., Ghanem, B.: SynthCLIP: Are We Ready for a Fully Synthetic CLIP Training? (2024), http://arxiv.org/abs/2402.01832

[24] He, H., Li, J.B., Jiang, X., Miller, H.: Sparse matrix in large language model fine-tuning. arXiv preprint arXiv:2405.15525 (2024)

[25] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: NeurIPS. pp. 1–25 (2020), http://arxiv.org/abs/2006.11239

[26] Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI

[27] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. ICLR (2022)

[28] Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. ICCV (2023)

[29] Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. The International Conference on Learning Representations (ICLR) (2023), https://api.semanticscholar.org/CorpusID:254408495

[30] Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10124–10134 (2023), https://api.semanticscholar.org/CorpusID:257427461

[31] Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. Advances in Neural Information Processing Systems **36** (2023)

[32] Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 317–325 (2017)

https://doi.org/10.52202/079017-1217

[33] Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)

[34] Lei, S., Chen, H., Zhang, S., Zhao, B., Tao, D.: Image Captions are Natural Prompts for Text-to-Image Models (2023), http://arxiv.org/abs/2307.08526

[35] Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. ACL (2022)

[36] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)

[37] Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22511–22521 (2023)

[38] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)

[39] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)

[40] Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)

[41] Liu, R., Garrette, D., Saharia, C., Chan, W., Roberts, A., Narang, S., Blok, I., Mical, R., Norouzi, M., Constant, N.: Character-aware models improve visual text rendering. ACL (2023)

[42] Liu, S., Liang, Y., Gitter, A.: Loss-balanced task weighting to reduce negative transfer in multi-task learning. In: AAAI Conference on Artificial Intelligence (AAAI) (2019), https://api.semanticscholar.org/CorpusID:84836014

[43] Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 852–869. Springer (2016)

[44] maintainers, T., contributors: Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision (2016)

[45] OpenAI: Gpt-4 technical report (2023), https://api.semanticscholar.org/CorpusID:257532815

[46] OpenAI: Openai models (2023), https://platform.openai.com/docs/models

[47] Paiss, R., Ephrat, A., Tov, O., Zada, S., Mosseri, I., Irani, M., Dekel, T.: Teaching clip to count to ten. ICCV (2023)

[48] Phung, Q., Ge, S., Huang, J.B.: Grounded text-to-image synthesis with attention refocusing. arXiv preprint arXiv:2306.05427 (2023)

[49] von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers (2022)

[50] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. ICLR (2024)

https://doi.org/10.52202/079017-1217

[51] Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 647–664. Springer (2020)

[52] Prabhudesai, M., Goyal, A., Pathak, D., Fragkiadaki, K.: Aligning text-to-image diffusion models with reward backpropagation. arXiv preprint arXiv:2310.03739 (2023)

[53] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., Wook, J., Chris, K., Aditya, H., Gabriel, R., Sandhini, G., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021), http://arxiv.org/abs/2103.00020

[54] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[55] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In: NeurIPS (2023), http://arxiv.org/abs/2305.18290

[56] Ramé, A., Couairon, G., Shukor, M., Dancette, C., Gaya, J.B., Soulier, L., Cord, M.: Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. Conference on Neural Information Processing Systems (NeurIPS) (2023), https://api.semanticscholar.org/CorpusID:259096117

[57] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021)

[58] Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. Advances in Neural Information Processing Systems **36** (2024)

[59] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[60] Saha, S., Hase, P., Bansal, M.: Can language models teach weaker agents? teacher explanations improve students via personalization. In: Advances in neural information processing systems (NeurIPS) (2023)

[61] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022)

[62] Samuel, A.L.: Some studies in machine learning using the game of checkers. IBM Journal of Research and Development **3**(3), 210–229 (1959). https://doi.org/10.1147/rd.33.0210

[63] Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it Till You Make it: Learning Transferable Representations from Synthetic ImageNet Clones. In: CVPR (2023). https://doi.org/10.1109/cvpr52729.2023.00774

[64] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. In: Proceedings of Neurips Data-Centric AI Workshop (2021)

[65] Segalis, E., Valevski, D., Lumen, D., Matias, Y., Leviathan, Y.: A picture is worth a thousand words: Principled recaptioning improves image generation. arXiv preprint arXiv:2310.16656 (2023)

[66] Shah, V., Ruiz, N., Cole, F., Lu, E., Lazebnik, S., Li, Y., Jampani, V.: Ziplora: Any subject in any style by effectively merging loras. ArXiv **abs/2311.13600** (2023), https://api.semanticscholar.org/CorpusID:265351656

[67] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature **529**(7587), 484–489 (jan 2016). https://doi.org/10.1038/nature16961

[68] Singh, S.P., Jaggi, M.: Model fusion via optimal transport. Conference on Neural Information Processing Systems (NeurIPS) (2020), https://api.semanticscholar.org/CorpusID:204512191

[69] Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)

[70] Sun, J., Fu, D., Hu, Y., Wang, S., Rassin, R., Juan, D.C., Alon, D., Herrmann, C., van Steenkiste, S., Krishna, R., et al.: Dreamsync: Aligning text-to-image generation with image understanding feedback. arXiv preprint arXiv:2311.17946 (2023)

[71] Sung, Y.L., Li, L., Lin, K., Gan, Z., Bansal, M., Wang, L.: An empirical study of multimodal model merging. Empirical Methods in Natural Language Processing (Findings) (2023)

[72] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)

[73] Tesauro, G.: Temporal difference learning and td-gammon. Commun. ACM **38**(3), 58–68 (mar 1995). https://doi.org/10.1145/203330.203343, https://doi.org/10.1145/203330.203343

[74] Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners. In: NeurIPS (2023), http://arxiv.org/abs/2306.00984

[75] Turc, I., Nemade, G.: Midjourney user prompts & generated images (250k) (2023)

[76] Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. arXiv preprint arXiv:2311.12908 (2023)

[77] Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. ACL (2023)

[78] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6

[79] Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human Preference Score: Better Aligning Text-to-image Models with Human Preference. In: ICCV (2023). https://doi.org/10.1109/iccv51070.2023.00200, http://arxiv.org/abs/2303.14420

[80] Wu, X., Huang, S., Wei, F.: Mole: Mixture of lora experts. In: The Twelfth International Conference on Learning Representations (2023)

[81] Xie, J., Li, Y., Huang, Y., Liu, H., Zhang, W., Zheng, Y., Shou, M.Z.: Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7452–7461 (2023)

https://doi.org/10.52202/079017-1217

[82] Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. NeurIPS (2023)

[83] Yadav, P., Tam, D., Choshen, L., Raffel, C., Bansal, M.: Ties-merging: Resolving interference when merging models. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)

[84] Yang, Z., Wang, J., Gan, Z., Li, L., Lin, K., Wu, C., Duan, N., Liu, Z., Liu, C., Zeng, M., et al.: Reco: Region-controlled text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14246–14255 (2023)

[85] Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. TMLR **2**(3), 5 (2023)

[86] Yu, L., Bowen, Y., Yu, H., Huang, F., Li, Y.: Language models are super mario: Absorbing abilities from homologous models as a free lunch. ArXiv **abs/2311.03099** (2023), https://api.semanticscholar.org/CorpusID:265034087

[87] Yuan, H., Chen, Z., Ji, K., Gu, Q.: Self-Play Fine-Tuning of Diffusion Models for Text-to-Image Generation (2024), http://arxiv.org/abs/2402.10210

[88] Yuan, W., Pang, R.Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., Weston, J.: Self-Rewarding Language Models (2024), http://arxiv.org/abs/2401.10020

[89] Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)

[90] Zhang, Y., Jiang, L., Turk, G., Yang, D.: Auditing gender presentation differences in text-to-image models. arXiv preprint arXiv:2302.03675 (2023)

[91] Zhong, M., Shen, Y., Wang, S., Lu, Y., Jiao, Y., Ouyang, S., Yu, D., Han, J., Chen, W.: Multi-lora composition for image generation. arXiv preprint arXiv:2402.16843 (2024)

https://doi.org/10.52202/079017-1217

Appendix

In this appendix, we present the following:

- Details of evaluation metrics we use (Appendix A).
- Evaluation on HPS v2.1 benchmark (Appendix B).
- Additional qualitative examples with SELMA on SDXL backbone (Appendix C).
- Skill-specific VQA accuracy on both TIFA and DSG benchmarks (Appendix D).
- Human evaluation details (Appendix E).
- Additional comparison with fine-tuning with filtered generated data (Appendix F).
- Bias evaluation in fine-tuned T2I models (Appendix G).
- Implementation details of SELMA (Appendix H).
- Implementation details of two training configuration variants: DPO and MoE-LoRA (Appendix I).
- Prompts we used to query LLM to generate new text data (Appendix J).
- Limitations and broader impact of SELMA approach (Appendix K).
- License information for data and model used in this paper (Appendix L).

A Evaluation Metrics

We quantitatively evaluate the performance of T2I generation models in text faithfulness and human preference metrics.

Text faithfulness. To evaluate T2I model's faithfulness in generation, we use VQA accuracy from TIFA and DSG. Specifically, TIFA and DSG utilize LLMs to generate questions given a text prompt and utilize the VQA model to check whether it can answer the questions correctly given the generated image. The image is considered to have better faithfulness to text prompts if the VQA model can answer the question more correctly. For TIFA, we use BLIP-2 as the VQA model following Sun *et al*. [70], For DSG, we use mPLUG-large [35] as the VQA model, as PaLI [12] is not publicly accecssible, and Hu *et al*. [28] shows that mPLUG achieves higher human correlation than BLIP-2.

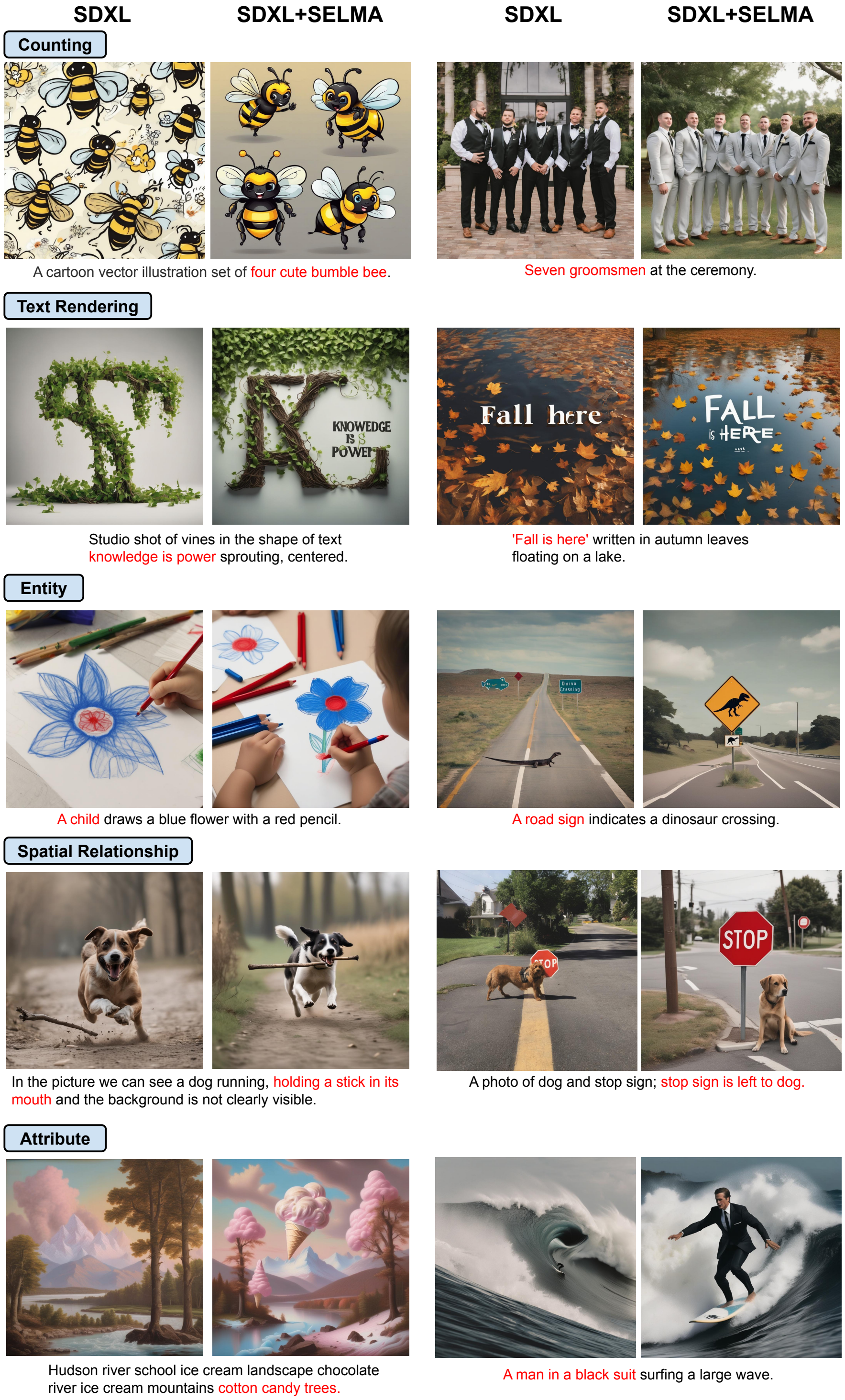
Human preference metrics. To evaluate how the generated images align with human preference, we use the PickScore [31], ImageReward [82], and HPS [79]. PickScore and HPS are based on CLIP [54] trained on the Pick-a-Pic dataset [31] and Human Preference Score dataset [79] respectively, which both have annotations of human preference over images. ImageReward is a BLIP [36] based reward model fine-tuned on human preference data collected on DiffusionDB. We calculate PickScore, ImageReward, and HPS on the 1060 DSG prompts. We also provide the evaluation results on HPS prompts in the appendix.

B Evaluation on HPS v2.1 Benchmark

In the main paper, we calculate HPS score [79] on text prompts on DSG benchmark, following DreamSync [70]. In this section, we additionally show the HPS score on the prompts from HPS v2.1 benchmark. HPS benchmark contains 3200 unique prompts from four different categories: anime, concept-art, paintings, and photo. We calculate the HPS score based on its HPS v2.1 model trained on higher quality datasets. As shown in Table 7, when adapting SELMA to different stable diffusion base model, our approach significantly improves the baseline performance (*i.e*., 2.5 for SD v1.4, 4.3 for SD v2, and 1.4 for SDXL), achieving better performance than all the released model on the HPS benchmark.[2]

[2] Benchmark performance can be found: https://github.com/tgxs002/HPSv2

https://doi.org/10.52202/079017-1217



A cartoon vector illustration set of four cute bumble bee.

Seven groomsmen at the ceremony.

Studio shot of vines in the shape of text knowledge is power sprouting, centered.

'Fall is here' written in autumn leaves floating on a lake.

A child draws a blue flower with a red pencil.

A road sign indicates a dinosaur crossing.

In the picture we can see a dog running, holding a stick in its mouth and the background is not clearly visible.

A photo of dog and stop sign; stop sign is left to dog.

Hudson river school ice cream landscape chocolate river ice cream mountains cotton candy trees.

A man in a black suit surfing a large wave.

Figure 6: Qualitative example images generated with SDXL and SDXL+SELMA (Ours) from DSG [14] test prompts requiring different skills. SELMA helps improve SDXL in various skills, including counting, text rendering, spatial relationships, and attribute binding. We highlight the parts of the prompts in red where SDXL makes errors while SDXL+SELMA generates correctly.

Table 7: Evaluation on HPS v2.1 evaluation benchmark. SELMA achieves signifiantly better scores on HPS evaluation benchmark compared with baselines, and outperforming all other baselines reported in HPS v2.1 benchmark.

Method	**HPS v2.1 Evaluation Benchmark**				
	Anime	Concept-Art	Paintings	Photo	Average
SD v1.4 [59]	26.0	24.9	24.8	25.7	25.4
+ SELMA	**28.2**	**27.6**	**27.8**	**28.0**	**27.9**
SD v2 [59]	27.1	26.0	25.7	26.7	26.4
+ SELMA	**32.0**	**30.3**	**30.0**	**30.4**	**30.7**
SDXL [50]	33.3	32.1	31.6	28.4	31.3
+ SELMA	**34.7**	**32.7**	**32.6**	**30.8**	**32.7**

Table 8: Detailed skill-specific comparison of SD models *vs*. SD models+SELMA on TIFA benchmark.

Method	**TIFA skills**												
	Animal/Human	Object	Location	Activity	Color	Spatial	Attribute	Food	Counting	Material	Other	Shape	Average
SD v1.4 [59]	83.7	78.3	80.3	71.7	73.0	58.9	74.5	81.8	63.3	76.6	47.3	**65.2**	75.8
+ SELMA	**87.1**	**83.0**	**84.8**	**75.9**	**74.4**	**62.3**	**76.0**	**88.1**	**66.2**	**78.5**	**52.2**	59.4	**79.5**
SD v2 [59]	86.5	82.6	83.8	75.6	76.8	62.4	75.4	85.2	**66.5**	**82.2**	55.4	**75.0**	79.2
+ SELMA	**89.7**	**88.0**	**87.6**	**80.3**	**80.3**	**66.0**	**77.2**	**91.0**	65.8	81.3	**63.2**	68.1	**83.2**
SDXL [50]	90.3	86.4	86.6	80.0	78.6	67.7	78.3	90.6	67.4	**84.2**	67.7	**62.3**	82.9
+ SELMA	**93.4**	**90.4**	**89.5**	**83.6**	**81.1**	**69.6**	**78.5**	**92.1**	**68.8**	83.7	**68.7**	60.9	**85.6**

C Additional Qualitative Exmaples

In Fig. 6, we show additional qualitative examples of SDXL and SDXL+SELMA from DSG [14] test prompts requiring different skills. SELMA helps improve SDXL in various skills, including counting, text rendering, spatial relationships, and attribute binding. For counting skill prompts, SDXL+SELMA generates “four bees” and “seven groomsmen” correctly following the text prompts. For text rendering skill prompts, SDXL+SELMA can render the text (“knowledge is power” and “Fall is here”) more accurately, while it still lacks the capability to render the text in the texture of vines or autumn leaves. For entity skill (placing correct objects) prompts, the SDXL sometimes misses some entities mentioned in the text prompt (*i.e.*, “A child”, and “A road sign”), while SDXL+SELMA can successfully generate them. For spatial relationship skill prompts, SDXL+SELMA generated images (*i.e.*, “holding a stick in its mouth”, and “stop sign left to dog”). Lastly, for attribute skill prompts, SDXL+SELMA binds objects with their corresponding attributes (*i.e.*, “cotton candy trees” and “A man in black suit”) more accurately than SDXL. These qualitative results demonstrate the effectiveness of SELMA.

D Skill-specific VQA Accuracy on TIFA and DSG

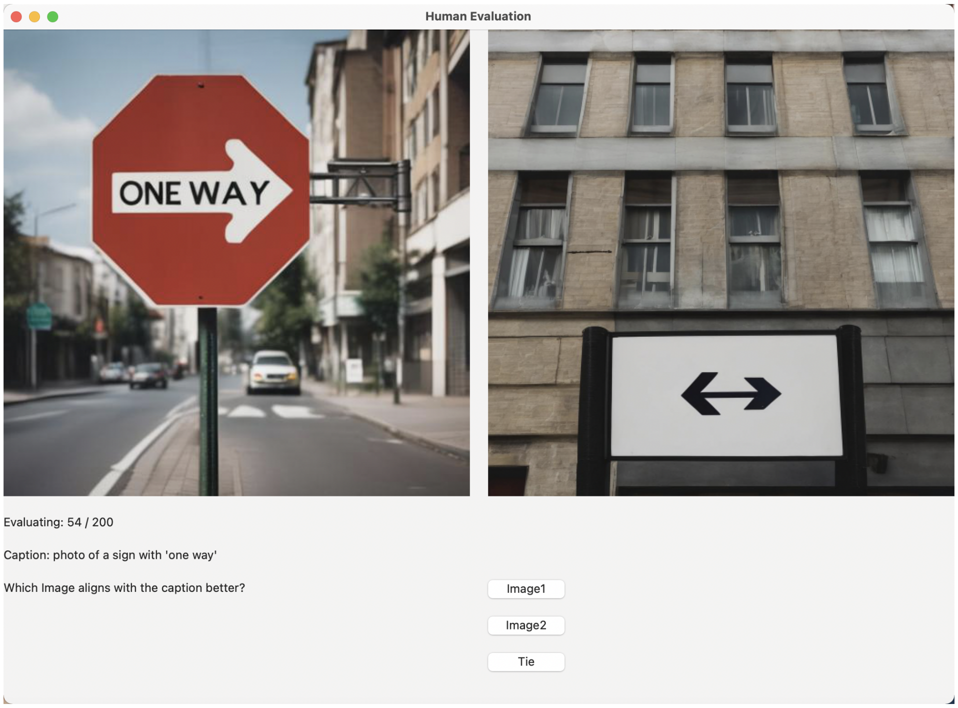
In this section, we show the detailed VQA accuracy for each skill category on TIFA and DSG benchmarks. Since DreamSync [70] does not provide the skill-specific scores, we report the skill-specific scores of SD models on TIFA and DSG and based on our experiments in Table 8 and Table 9; we observe there are less than 1% score differences of SD/SDXL models in TIFA average accuracy between the results in Sun *et al*. and ours.

As shown in Table 8 and Table 9, on both TIFA and DSG benchmarks, SELMA improves the generation faithfulness in most of the categories. Comparing SDXL and SDXL+SELMA, the SDXL finetuned with SELMA approach shows large improvement especially in entity (*i.e.*, 3.1% on animal/human on TIFA compared with SDXL, 5.6% in whole on DSG, 12.2% in part on DSG), as well as spatial relationship (*i.e.*, 1.9% on TIFA, and 8.0% on DSG), and counting skills (*i.e.*, 1.4% on TIFA, and 13.2% on DSG). Besides, we also observe that SELMA significantly improves the text rendering for SD v2 and SDXL (*i.e.*, 16.4% compared with SDXL, and 6.5% compared with SD v2 on DSG), but not SD v1.4 (*i.e.*, 1.8% decrease compared with SD v1.4 on DSG).

https://doi.org/10.52202/079017-1217

Table 9: Detailed skill-specific comparison of SD models *vs*. SD models+SELMA on DSG benchmark. We show the skill categories that have more than 50 questions.

Method	DSG skills												
	Whole	Color	Shape	Spatial	Part	State	Count	Action	Global	Material	Type	Text Rendering	Average
SD v1.4 [59]	78.6	**62.5**	46.0	61.1	68.1	58.2	62.4	59.9	**59.4**	42.3	**73.0**	**52.7**	67.2
+ SELMA	**83.7**	**62.5**	**52.0**	**66.2**	**72.1**	**63.3**	**66.1**	**72.1**	57.1	**59.7**	67.6	50.9	**71.3**
SD v2 [59]	80.8	68.6	50.0	63.6	72.3	63.6	**69.3**	62.9	**61.9**	55.7	66.8	60.9	70.3
+ SELMA	**88.0**	**80.8**	**65.4**	**71.0**	**78.7**	**71.5**	66.3	**78.4**	61.0	**69.2**	**81.4**	**67.4**	**77.7**
SDXL [50]	84.8	74.7	58.0	69.4	71.1	60.7	59.8	71.7	**61.5**	63.9	71.7	60.0	73.3
+ SELMA	**90.4**	**81.3**	**64.0**	**77.4**	**83.3**	**68.2**	**73.0**	**79.4**	60.2	**77.3**	**75.4**	**76.4**	**80.2**



Evaluating: 54 / 200

Caption: photo of a sign with 'one way'

Which Image aligns with the caption better?

Figure 7: Example user interface for human evaluation on DSG prompts.

E Human Evaluation Details

We conduct the human evaluation (described in the main paper Sec. 5.5) on 200 randomly sampled prompts from DSG, with three external annotators. We show the annotation interface in Fig. 7. The image order between the two models is randomized to avoid the leakage of information about which image is generated with which model.

In Table 10, we show the detailed annotator votes for win, lose, and tie for SDXL and SDXL+SELMA. SDXL+SELMA has significantly higher win votes compared with SDXL on all the 200 sampled text prompts (*i.e.*, 241 win *vs*. 114 lose), demonstrating the effectiveness of SELMA.

Table 10: Human Evaluation on 200 sampled text prompts from DSG. We show the detailed win/lose/tie counts on all samples and samples from each dataset.

Eval Dataset	Win	Lose	Tie
All	241	114	245
Localized Narratives	19	5	12
DiffusionDB	6	5	34
CountBench	26	11	17
COCO	35	2	47
Whoops	21	10	26

F Comparison with Fine-tuning with Filtered Generated Data

We demonstrate that our automatically collected data is of high quality to improve the faithfulness of T2I models. Specifically, we filter out the generated images based on TIFA score (>0.9) and VILA score (>0.6), and maintain 1K data for each set of skill-specific prompts. Fine-tuning with filtered images achieves 80.3% DSG score with SDXL backbone, which is similar performance as fine-tuning with un-filtered version (i.e., 80.2% on DSG), indicating the overall high quality of the data, which does not lead to performance degradation.

G Bias Evaluation in T2I Models

In this section, we investigate whether fine-tuning with T2I-generated images will increase the bias in existing T2I models. Specifically, we evaluate the T2I model with GEP score [90], which measures the gender bias in T2I models. Specifically, a learned cross-modal classifier is used to distinguish attributes (*e.g.*, dress) in the image, and GEP score calculates attribute-wise difference based on output score from the cross-modal classifier. A lower GEP score indicates less gender bias. Fine-tuning SD v2 with our generated data achieves 0.04413 GEP score, while fine-tuning with ground truth data achieves 0.04976. This indicates that fine-tuning with our generated data does not amplify the bias in existing T2I models compared with fine-tuning with ground truth data.

H Implementation Details of SELMA

In the **prompt generation** stage (Sec. 3.1), we use `gpt-3.5-turbo-instruct` [46] to generate text prompts by providing three prompts for each skill as in-context examples. For each of the five datasets (COCO [39], Localized Narratives [51], DiffusionDB [77], CountBench [47], and Whoops [6]), we collect 1K prompts starting with three prompts randomly sampled from them, ensuring the prompts are not included in the DSG test prompts (*i.e.*, 5K prompts in total). We refer to the resulting auto-generated datasets as Localized Narrative$^{\text{SELMA}}$, CountBench$^{\text{SELMA}}$, DiffusionDB$^{\text{SELMA}}$, Whoops$^{\text{SELMA}}$, and COCO$^{\text{SELMA}}$. We refer to the resulting combination of 5K auto-generated dataset as DSG$^{\text{SELMA-5K}}$.

In the **image generation** stage (Sec. 3.2), we use the default denoising steps 50 for all models, and the Classifier-Free Guidance (CFG) [26] of 7.5. In the **LoRA fine-tuning** stage (Sec. 3.3), we use 128 as the LoRA rank. We fine-tune LoRA in mixed precision (*i.e.*, FP16) with a constant learning rate of 3e-4 and a batch size of 64. We fine-tune LoRA modules for 5000 steps, which is approximately 313 epochs. **During inference**, we uniformly merge the specialized LoRA experts into one multi-skill expert (Sec. 3.4). We evaluate model checkpoints every 1000 steps and pick the model with the best text faithfulness on DSG benchmark. Fine-tuning LoRA for SD v1.4, SD v2, and SDXL takes 6 hours, 6 hours, and 12 hours on a single NVIDIA L40 GPU, respectively. We use Diffusers [49] for our experiments.

I Implementation Details of DPO and MoE-LoRA

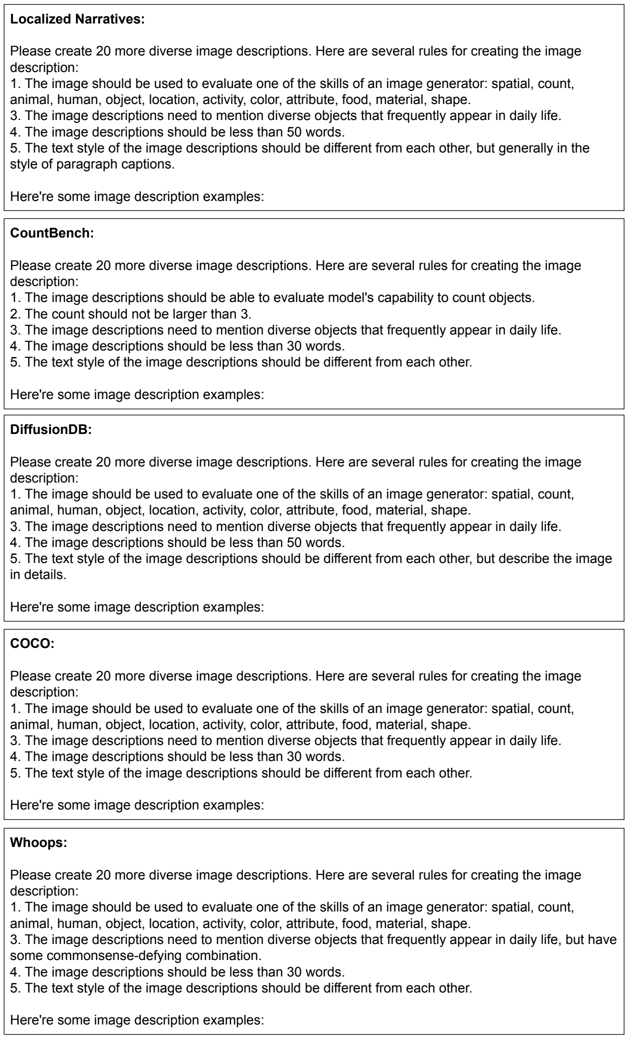
In this section, we provide the implementation details of two training approaches we experiment with (described in Sec. 5.6 in the main paper).

Direct Preference Optimization (DPO). We fine-tune LoRA models with DPO proposed in [76]. Specifically, we sample two images with T2I models and calculate the image-text alignment with CLIP score [54]. We use the image with a higher CLIP score as the positive example and the image with a lower CLIP score as the negative example. We fine-tune the LoRA models to learn to generate images closer to the positive image distribution and push away from generating images similar to the negative image distribution. Similarly, we fine-tune with DPO on five datasets with different text styles and skills and merge LoRA expert models during inference time by averaging the LoRA weights. In DPO training, we use a constant learning rate 3e-4 and fine-tune LoRA for 5K steps. We evaluate the model on DSG every 1K steps and pick the best checkpoint.

Mixture of Lora Experts (MoE-LoRA). MoE-LoRA [80] utilizes a gating function (router) to decide which experts to use during training and inference. The gating function predicts weights for each expert based on layer inputs and picks the top K experts to use at each layer. Specifically, the gating function we use is a simple linear mapping function, where $\{w_i\}_{i=1}^{K} = W_g x$. x is the input to each layer, W_g is the learnable gating weights, and $\{w_i\}_{i=1}^{K}$ are the predicted weights. The outputs of each expert are added together with the normalized weights from the gating function. In MoE-LoRA, we initialize five LoRA experts fine-tuned on different datasets containing different text styles and skills. We freeze the learned LoRA weights and only fine-tune the gating function

https://doi.org/10.52202/079017-1217

on the collected five datasets. We also compare with learning LoRA weights along with the router, which achieves worse performance (75.9 on DSG). We activate all five experts during training and inference (*i.e.*, $K = 5$). In MoE-LoRA training, we use a constant learning rate 1e-5 and fine-tune LoRA for 5K steps. We evaluate the model on DSG every 1K steps and pick the model with highest text faithfulness score.



Localized Narratives:

Please create 20 more diverse image descriptions. Here are several rules for creating the image description:
1. The image should be used to evaluate one of the skills of an image generator: spatial, count, animal, human, object, location, activity, color, attribute, food, material, shape.
3. The image descriptions need to mention diverse objects that frequently appear in daily life.
4. The image descriptions should be less than 50 words.
5. The text style of the image descriptions should be different from each other, but generally in the style of paragraph captions.

Here're some image description examples:

CountBench:

Please create 20 more diverse image descriptions. Here are several rules for creating the image description:
1. The image descriptions should be able to evaluate model's capability to count objects.
2. The count should not be larger than 3.
3. The image descriptions need to mention diverse objects that frequently appear in daily life.
4. The image descriptions should be less than 30 words.
5. The text style of the image descriptions should be different from each other.

Here're some image description examples:

DiffusionDB:

Please create 20 more diverse image descriptions. Here are several rules for creating the image description:
1. The image should be used to evaluate one of the skills of an image generator: spatial, count, animal, human, object, location, activity, color, attribute, food, material, shape.
3. The image descriptions need to mention diverse objects that frequently appear in daily life.
4. The image descriptions should be less than 50 words.
5. The text style of the image descriptions should be different from each other, but describe the image in details.

Here're some image description examples:

COCO:

Please create 20 more diverse image descriptions. Here are several rules for creating the image description:
1. The image should be used to evaluate one of the skills of an image generator: spatial, count, animal, human, object, location, activity, color, attribute, food, material, shape.
3. The image descriptions need to mention diverse objects that frequently appear in daily life.
4. The image descriptions should be less than 30 words.
5. The text style of the image descriptions should be different from each other.

Here're some image description examples:

Whoops:

Please create 20 more diverse image descriptions. Here are several rules for creating the image description:
1. The image should be used to evaluate one of the skills of an image generator: spatial, count, animal, human, object, location, activity, color, attribute, food, material, shape.
3. The image descriptions need to mention diverse objects that frequently appear in daily life, but have some commonsense-defying combination.
4. The image descriptions should be less than 30 words.
5. The text style of the image descriptions should be different from each other.

Here're some image description examples:

Figure 8: Prompts used to query GPT-3.5 auto-generate new prompts targeting different skills.

J Skill-Specific Prompt Generation Details

We show the prompts we use to query GPT3.5 to generate 1K prompts for each skill. As shown in Fig. 8, we use different prompts to generate SELMA data. For example, we specify "paragraph captions" to generate text prompts that can be used to teach model to follow long text prompts, and specifying "evaluate model's capability to count objects" to collect a set of prompts for improving model's counting capability. Besides, in all the prompt generation, we emphasize that the image should "mention diverse objects" to maximize the semantic diversity in generated prompts.

K Limitations and Broader Impact

Text-to-image generation models can be used in many real-world applications, such as creating content for media and entertainment. Our proposed SELMA improves T2I models' faithfulness with auto-generated data, reducing human annotation efforts for high-quality image-text pairs.

However, we also note that our SELMA has several limitations. SELMA relies on a strong image generator and an instruction-following LLM. Note that SELMA is model-agnostic and can be implemented with publicly accessible models (GPT-3.5/LLaMA3 and Stable Diffusion models). Also, since our fine-tuning works well with a small number of image-text pairs (*i.e.*, for each skill, we only generate 1K text prompts and generating one image per each text prompt), the cost of LLM inference (*i.e.*, $27.78 for querying GPT-3.5 for generating prompts in all the experiments including ablation studies) and image generation (8s per image for image generation with SDXL on a single NVIDIA L40 GPU with 48GB memory) is minimal. Besides, although SELMA helps boost T2I models' performance significantly in both text faithfulness and alignment to human preference, fine-tuning with SELMA does not guarantee the resulting model to follow the text prompts in every detail. To use T2I models trained with SELMA, researchers should first carefully study their capabilities in relation to the specific context in which they are being applied.

L Licenses

We provide the licenses of the existing assets we use in this paper in Table 11.

Table 11: A list of the licenses of the existing assets used in this paper.

Asset	License
PyTorch [3]	BSD-style
Huggingface Transformers [78]	Apache License 2.0
Torchvision [44]	BSD 3-Clause “New” or “Revised” License
Diffusers [49]	Apache License 2.0
Stable Diffusion [59]	CreativeML Open RAIL-M
COCO dataset [39]	CC BY 4.0
Localized Narrative dataset [51]	CC BY 4.0
DiffusionDB [77]	MIT License
Whoops [6]	CC BY 4.0
CountBench [47] (LAION-400M [64] subset)	CC BY 4.0
GPT3.5 [46]	OpenAI Terms of Use
LLaMA3 [1]	Meta LLaMA3 License

https://doi.org/10.52202/079017-1217

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims and contributions are clearly stated in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations in Appendix K.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (*e.g.*, independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, *e.g.*, if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-1217

Justification: We don't make theoretical contribution.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include implementation details in main paper Sec. 4.3, and further add more details in Appendix H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (*e.g.*, in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (*e.g.*, a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (*e.g.*, with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (*e.g.*, to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-1217

Answer: [Yes]

Justification: We include code in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (*e.g.*, for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (*e.g.*, data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include implementation details in main paper Sec. 4.3, and further add more details in Appendix H.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report single-run results for all the experiments following previous work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (*e.g.*, Normally distributed errors).

https://doi.org/10.52202/079017-1217

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (*e.g*. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide this information in the implementation details in Appendix H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (*e.g*., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conforms every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (*e.g*., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts in Appendix K.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (*e.g*., disinformation, generating fake profiles, surveillance), fairness considerations (*e.g*., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-1217

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (*e.g.*, gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (*e.g.*, pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We don't have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (*e.g.*, code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited previous work properly in Appendix L and strictly followed the license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (*e.g.*, CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (*e.g.*, website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

https://doi.org/10.52202/079017-1217

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We'll release our code, data and model checkpoints upon acceptance under CC license.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper doesn't involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper doesn't involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1217