Exploring State of the Art Diffusion Models

Stanford CS229 Project

Catherine Lee
Department of Computer Science
Stanford University
cation@stanford.edu

Jennifer Zheng
ICME
Stanford University
jenzheng@stanford.edu

Abstract

This paper explores whether new and novel diffusion models can compete in performance and inference time with current state of the art diffusion models. We evaluate specific flavors of latent diffusion models and latent consistency models on text-to-image generation on a variety of benchmarks such as CLIPscore, Human Preference Score, and X-IQE.

1 Introduction

The field of stable diffusion has grown immensely. Every day new methods are coming out to improve the training, inference, and performance of stable diffusion. Currently, latent diffusion models which use denoising autoencoders are state of the art because they improve on traditional diffusion models by applying diffusion in latent space which reduces training time. Latent consistency models further improved the inference time of diffusion models by using a distillation method on a pre-trained diffusion model. There are also methods to improve fine-tuning that have been used in other applications like low rank adaptation. Our problem is to compare across the different state-of-the art diffusion models given the same set of prompts and to evaluate their outputs using metrics such as CLIP Score, Image Reward, and Human Preference Score.

2 Related Work

Diffusion models have shown significant promise in various applications and have been the most popular methods in generating images. Yang et al. (2022) provides a comprehensive survey of these models, categorizing them into three key areas: efficient sampling, improved likelihood estimation, and handling data with special structures. Lin et al. (2023) extends the application of diffusion models to time series forecasting, imputation, and generation, highlighting their potential in these areas.

Diffusion models (DM) (Sohl-Dickstein et al., 2015) are the initial state-of-the-art approaches towards generating images, but typically operate directly in pixel space, resulting in large consumption of computational resources. Latent diffusion models (LDM) (Rombach et al., 2022) applies DM in the latent space of pretrained autoencoders and achieved new state-of-the-art scores for text-to-image synthesis. Consistency Models (CM) (Song et al., 2023) adopted consistency mapping to enable fast one-step generation, becoming a new family of generative models. Latent consistency models (LCM) (Luo et al., 2023a) then adopted the ideas from both CM and LDM, producing high-quality images with minimal inference steps. We then improve the models from the large language modeling (LLM) approach. Parameter-Efficient Fine-Tuning (PEFT) (Houlsby et al., 2019) enables the customization of pre-existing models to reduce computational load and storage demands. Low-Rank Adaptation (LoRA) (Hu et al., 2021) is a PEFT technique that especially enhances computational efficiency by applying a low-rank decomposition. LCM-LORA (Luo et al., 2023b) then incorporates LoRA during the LCM distillation process to reduce the quantity of trainable parameters. Another LLM approach is Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017) methods which align LLMs with users' preferences. Diffusion-DPO (Wallace et al., 2023) adapts Direct Preference

Optimization (DPO) (Rafailov et al., 2024), a simpler alternative to RLHF which directly optimizes a policy that best satisfies human preferences under a classification objective.

3 Dataset and Features

We used the Microsoft COCO Captions dataset (Chen et al., 2015), which contains over one and a half million captions describing over 330,000 images. We used 100 prompts from COCO as our benchmark.

3.1 Metrics

We adapted the evaluation metrics from Chen (2023). The weighted average score is calculated as:

CLIPScore * 15% + AestheticPred * 15% + ImageReward * 20% + HPS * 20% + (X-IQE Fidelity + X-IQE Alignment + X-IQE Aesthetics) * 10%.

The detailed metrics are described below:

1. CLIP Score (Hessel et al., 2021)

   CLIPScore is a metric used to evaluate the quality of an automatic image captioning system. It uses the cosine similarity between the extracted features of the images and the captions to measures the text-image alignment. For an image with visual CLIP embedding v and a candidate caption with textual CLIP embedding c, the CLIP score is defined as

   $$\text{CLIP-S}(\mathbf{c}, \mathbf{v}) = 2.5^* \max(\cos(\mathbf{c}, \mathbf{v}), 0)$$

   The CLIPscore is bounded from 0 to 100 where higher positive scores correlating to superior outcomes.

2. Aesthetic Score Predictor (Schuhmann et al., 2022)

   The aesthetic score measures how good-looking an image is - the larger the more aesthetic. It is determined by a linear estimator on top of CLIP.

3. Image Reward (Xu et al., 2023)

   ImageReward measures the human rating of an image. This metric is the first general-purpose text-to-image human preference reward model. It is trained with ranking data and human annotations.

4. Human Preference Score (HPS) (Wu et al., 2023)

   HPS measures the human preference of an image. It is a classifier trained to better predict human choices. The score is the cosine similarity between the text and image features complemented by incorporating human aesthetic preferences.

5. X-IQE (Chen, 2023)

   X-IQE is a comprehensive and explainable metric based on visual LLMs (MiniGPT-4). We prompt MiniGPT-4 to evaluate our image based on fidelity, image-to-prompt alignment and aesthetics. MiniGPT-4 is trained for general vision-language knowledge and returns explainable responses because it uses Chain-of-Thought prompting. The score is bound from 1 to 10 for fidelity, 1 to 5 on image-to-prompt alignment, and 1 to 10 on aesthetics. The model outputs a json with a score and reasoning for the score. Prompts are provided by Chen (2023)

Among the metrics, CLIPScore (Contrastive Language–Image Pre-training) is the most classical metric. However, it is unclear if CLIPscore can accurately evaluate generated images as it is trained on image-caption pairs, where the images in the training set are not generated images. Futhermore, CLIPscore fails in modeling human preferences on image quality. In contrast, human preference score is trained on image pairs generated from 9 recent diffusion models along with text labels. Each metric has drawbacks so we use a wide variety of metrics.

4 Methods

We applied the diffusion models on a prompt dataset in order of their evolution to compare their computational costs and qualities. The models that we tested are, in order of evolution, latent diffusion models (LDM), latent consistency models (LCM), LCM-LoRA, and LDM and DPO.

1. Latent Diffusion Models (LDM) Rombach et al. (2022)

   Diffusion Probabilistic Models (DM) by Sohl-Dickstein et al. (2015) was the initial state-of-the-art method in density estimation and training stability. The DM modesl gradually denoise the noise-corrupted data by the reverse diffusion process to estimate the score of data distribution.

   The LDM model then builds on top of the DM such that it performs forward and reverse diffusion processes in the data latent space, resulting in a more efficient computation.

2. Latent Consistency Models (LCM) Luo et al. (2023a)

   LCM is inspired by Consistency Models (CM) proposed by Song et al. (2023), which are a family of generative models that enable one-step or few-step generation, allowing the models to be trained as distill pre-trained DM, or standalone generative models.

   LCMs adopt a CM in the image latent space to enhance image generation quality and reduce computational load. Consistency models require much less inference steps in comparison to diffusion models.

3. LCM-LoRA Luo et al. (2023b)

   Low-Rank Adaptation (LoRA) introduced by Hu et al. (2021) freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture to reduce the number of trainable parameters for downstream tasks.

   LCM-LoRA applies LoRA distillation to Stable-Diffusion models to expand LCMs' scope to larger models with significantly less memory consumption.

4. Latent Diffusion Model with Direct Preference Optimization (LDM-DPO) Wallace et al. (2023)

   Direct Preference Optimization (DPO) by Rafailov et al. (2024) is a stable, performant, and computationally lightweight algorithm used to solve the human preference data classification problem. It can control the sentiment of generations and improve response quality, especially in comparison to reinforcement learning from human feedback (RLHF) (Christiano et al., 2017).

   Diffusion-DPO aligns diffusion models to human preferences by directly optimizing on human comparison data. Using this technique, DPO is reformulated to account for likelihood, utilizing the evidence lower bound to derive a differentiable objective.

5 Experiments / Results / Discussion

For a dataset of prompts with size 100, we generate an image with each prompt using different models and evaluate their performance using the metrics introduced in Section 3.1. Our results are shown in Table 1.

	Clip	HPS	Image Reward	Aesthetic
LDM (75 steps)	31.72	0.3135	0.1508	5.3109
LDM-DPO (75 steps)	31.75	0.3157	0.9881	5.6409
LCM (45 steps)	30.79	0.3096	0.4472	5.4586
LCM-LoRA (45 steps)	25.78	0.2742	-0.9510	5.0854

Table 1: Evaluation of the images generated by the different models

We timed the generation of one image for each of the models using 45 inference steps. The computation time in seconds are shown in Table 2. The machine used is Quadro RTX 6000, AMD EPYC processors, CUDA 11.8.

	LDM	LDM-DPO	LDM(75)	LCM	LCM-LoRA
Time (seconds)	219.09	15.30	371.77	7.79	13.70
CLIPScore	28.63	26.41	25.33	26.96	25.32
HPS	0.3472	0.3330	0.3140	0.3181	0.2825
ImageReward	1.1003	1.5613	1.1446	1.3684	-2.1720
Aesthetic	4.7034	5.8029	5.2901	5.6402	4.8731
X-IQE					
Fidelity	4	6	4	4	4
Alignment	1	5	1	1	1
Overall Aesthetic	3	1	4	5	3
Weighted Average	6.089	6.410	5.7847	6.2273	4.951065

Table 2: Evaluation of the images for prompt "Man riding a motor bike on a dirt road on the countryside" using 45 inference steps

The images corresponding to the results in Table 2 are shown in Figure 1

(a) LDM (b) LCM (c) LDM-DPO (d) LCM-LoRA

Figure 1: Images generated with prompt "Man riding a motor bike on a dirt road on the countryside" using 45 inference steps using different diffusion models

From the generated images, we observe that Figure 1b and Figure 1c have higher qualities and are closer the prompts (from the authors' perspectives). These two images score higher in all metrics except in HPS and CLIPScore. We suggest that our disagreement with HPS might result from different preferences in the training dataset.

As we calculate the overall weighted average score using the formula in Section 3.1, Figure 1b and Figure 1c do have the highest overall score, suggesting that the weights of the formula is reasonable.

We also observe that under the same inference steps (45) and guidance scales (8.0), Figure 1a has significantly lower quality in image resolution and much longer computation time, suggesting that even though LDM was the most popular method used in Stable Diffusion, the other methods have improved towards much faster computations and higher qualities.

As LDM scores higher than LCM and LDM-DPO in CLIPScore and HPS with a lower quality image, we suggest that the two metrics are trained using the diffusion model, and might be the reason that they are biased towards Figure 1a.

While Figure1d has high resolution and aligns with the prompt, we observe that it has a different style in comparison to all the other generated images in Figure 1, resulting in a significantly lower score in all metrics. This infers that the metrics have a preference over one specific style - how realistic the images are.

As we mentioned in Section 4, LCM requires minimal inference steps in comparison to LDM. We thus conduct the same experiments on LDM with an inference step of 75, shown in Figure 2 and the third column of Table 2.

Figure 2: Image generated with prompt "Man riding a motor bike on a dirt road on the countryside" using 75 inference steps using LDM

We observe that while no significant improvement in the image quality and an even longer computation time, suggesting that it is not worth the computational resources to generate an image with LDM with high inference steps.

We also visualized the prompt in Figure 3 using cross attentions by Tang et al. (2023). Highlighted sections strongly correlate with the specific word in the prompt.

Figure 3: Visualization of "Man riding a motor bike on a dirt road on the countryside" using LDM

6 Conclusion / Future Work

Although Latent Consistency models improve on inference time and training time (not explored here but in Luo et al. (2023a)), their performance does not beat the current state of the art diffusion methods, Latent Diffusion Models. Also, improvements such as Direct Preference Optimization and Low Rank Adaptation for fine tuning do not consistently improve the performance of the respective models. Although LDM-DPO had improved performance over classical LDM, LCM-LoRa did not. This indicates multiple opportunities; latent consistency models, although faster at inference and training, need more improvements to be able to catch up to Latent Diffusion Models. If we had more compute, we would train LCM with LoRA weights and see if we could improve the performance through other means like Direct Preference Optimization during training.

7 Acknowledgement

Team members truly appreciate Kefan Dong, Professor Emily Fox, and Professor Sanmi Koyejo for their wonderful mentorships throughout this project. We thank the CS229 course staffs for their help and GCP credits. Our computing resources come from the Institute of Computational and Mathematical Engineering and the Google Cloud credits provided by the course.

8 Contributions

Jennifer worked on the evaluation of Latent Consistency Model and LCM-LoRA. Catherine worked on the evaluation of Latent Diffusion Model and LDM-DPO model. Both worked on resolving dataset expLoRAtion and splitting dataset into train/test with the Hugging Face interface. Both worked equally on the project proposal, the project milestone, and the final project.

References

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yixiong Chen. 2023. X-iqe: explainable image quality evaluation for text-to-image generation with visual large language models. *arXiv preprint arXiv:2305.10843*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Lequan Lin, Zhengkun Li, Ruikun Li, Xuliang Li, and Junbin Gao. 2023. Diffusion models for time-series applications: a survey. *Frontiers of Information Technology & Electronic Engineering*, 25:19–41.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023a. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. 2023b. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models. *arXiv preprint arXiv:2303.01469*.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659, Toronto, Canada. Association for Computational Linguistics.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2023. Diffusion model alignment using direct preference optimization. *arXiv preprint arXiv:2311.12908*.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Better aligning text-to-image models with human preference. *ArXiv*, abs/2303.14420.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation.

Ling Yang, Zhilong Zhang, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. 2022. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56:1 – 39.