EvolveDirector: Approaching Advanced Text-to-Image Generation with Large Vision-Language Models

Rui Zhao[1], **Hangjie Yuan**[2], **Yujie Wei**[2,3], **Shiwei Zhang**[2], **Yuchao Gu**[1], **Lingmin Ran**[1], **Xiang Wang**[2,4], **Zhangjie Wu**[1], **Junhao Zhang**[1], **Yingya Zhang**[2], **Mike Zheng Shou**[1,*]

[1]Show Lab, National University of Singapore [2]Alibaba Group [3] Fudan University

[4] Huazhong University of Science and Technology

Abstract

Recent advancements in generation models have showcased remarkable capabilities in generating fantastic content. However, most of them are trained on proprietary high-quality data, and some models withhold their parameters and only provide accessible application programming interfaces (APIs), limiting their benefits for downstream tasks. To explore the feasibility of training a text-to-image generation model comparable to advanced models using publicly available resources, we introduce EvolveDirector. This framework interacts with advanced models through their public APIs to obtain text-image data pairs to train a base model. Our experiments with extensive data indicate that the model trained on generated data of the advanced model can approximate its generation capability. However, it requires large-scale samples of 10 million or more. This incurs significant expenses in time, computational resources, and especially the costs associated with calling fee-based APIs. To address this problem, we leverage pre-trained large vision-language models (VLMs) to guide the evolution of the base model. VLM continuously evaluates the base model during training and dynamically updates and refines the training dataset by the discrimination, expansion, deletion, and mutation operations. Experimental results show that this paradigm significantly reduces the required data volume. Furthermore, when approaching multiple advanced models, EvolveDirector can select the best samples generated by them to learn powerful and balanced abilities. The final trained model Edgen is demonstrated to outperform these advanced models. The code and model weights are available at https://github.com/showlab/EvolveDirector.

1 Introduction

In the field of AI-generated content, an increasing number of advanced models have showcased their ability to generate realistic and imaginative images, such as Imagen [1], DALL·E 3 [2], Stable Diffusion 3 [3], Midjourney [4]. While these models benefit from publicly available datasets such as ImageNet [5], LAION [6], and SAM [7], they rely more heavily on proprietary, high-quality data collections that surpass the quality of publicly accessible datasets. Models such as Midjourney [4] are particularly noted for deriving substantial benefits from their internal datasets. However, given the significant commercial advantages brought by their impressive capabilities, most advanced models keep their parameters private, hindering reproducibility and democratization. In this paper, we aim to *explore training an open-source text-to-image model with public resources to achieve comparable capabilities with the existing advanced models.*

*Corresponding Author.

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

 https://doi.org/10.52202/079017-3880

Three transparent glass bottles on a wooden table. The one on the left has red liquid and the number 1. The one in the middle has blue liquid and the number 2. The one on the right has green liquid and the number 3.

Four superheroes stand in front of the pyramids, from left to right, Batman, Hulk, Iron Man, and Spider-Man.

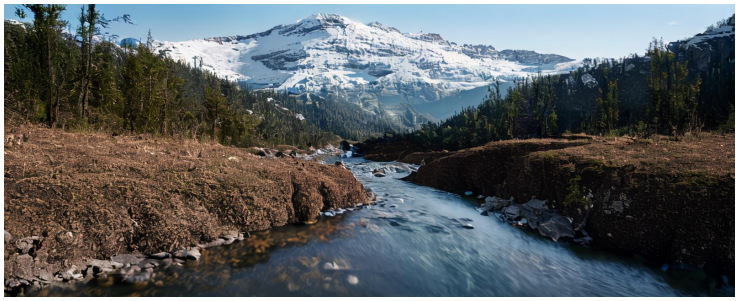
Pallet knife technique, large intentional brushstrokes, masterpiece painting of a leopard silently prowling through the underbrush, dark fantasy.

In a garden, there sits a white cat on the left and a black dog on the right, with a green water cup placed between the two of them.

A captivating book cover features a young woman with an enchanting gaze, dressed in vibrant attire, surrounded by swirling colors, inviting readers into a world of chaos and beauty.

A sleek and modern logo design for NeurIPS 2024. The text "NeurIPS 2024" is written in bold, black capital letters. In the background, there's a minimalist, abstract representation of a brain, with a few key parts highlighted in neon colors. The overall design is clean, visually appealing, and instantly recognizable.

A volcanic eruption in a dense jungle creates a striking scene as lava flows down the mountain, contrasting with the lush green foliage. Dark clouds fill the sky, casting dramatic shadows over the landscape.



At the foot of the snow-capped mountain, a clear stream winds its way, murmuring and flowing gently, resembling a silver ribbon that connects the forests and canyons.

Figure 1: Images generated by our model Edgen (EvolveDirector-Gen). Edgen can generate high-quality images with multiple ratios and resolutions. Notably, it excels in generating text and avoiding attribute confusion when generating multiple objects, which are significant characteristics of the most advanced text-to-image models available today. The input text prompts are annotated under the corresponding images.

https://doi.org/10.52202/079017-3880

Despite the fact that internal data and model parameters of many advanced models remain inaccessible, they provide publicly accessible application programming interfaces (APIs) that enable users to access their generated distribution. This leads to the construction of synthetic benchmarks, *e.g.*, JourneyDB [8] collects 4.7 million images generated by Midjourney [4]. This benchmark is further utilized for enhancing the training of new generative models [9]. However, this paradigm is not data efficient, posing challenges in terms of substantial computation and expenses. Instead of statically constructing multiple expensive large-scale datasets for each advanced model, we take a step forward in this paper by delving into recovering their generative capabilities in a unified framework with limited samples. We propose EvolveDirector to address this challenging task by shedding light on two research questions: (1) *How many synthetic text-image pairs are sufficient to approximate the generative capability of an advanced model?* (2) *Taking it a step further, is it possible for the base model to obtain generative capabilities beyond the advanced models?*

To explore the first question, we start with a demonstration experiment by training a relatively poor model, a DiT model [10] pre-trained on public dataset ImageNet [5] and SAM [7], to approach the advanced model PixArt-α [9] using increasing data scales. The training data is curated by collecting diverse text prompts and utilizing them to generate images from PixArt-α. The experiments indicate that when we scale the training data (*i.e.*, generated image-text pairs) to 11 million, we can obtain a base model achieving similar capabilities to the target model without access to its internal data. However, the magnitude of 11 million generated data is comparable to the 14 million data used for pre-training the target model. Training a base model in this way incurs significant expenses, not only in terms of time and computational resources but also the costs associated with using fee-based APIs of some advanced models.

For more efficient training, it is crucial to minimize data redundancy and maximize data quality, as the marginal benefit of training on inferior data is limited. The corpus of the 11 million text prompts, generated from the SAM dataset [9], and the images generated by the advanced model exhibit redundancy in several aspects: (1) *lacking imaginative text prompts* due to the photographic nature of SAM images; (2) *high similarities among text prompts*; and (3) *imbalanced data quality*. The generated images using the target advanced model may exhibit low quality due to inferior alignment on some text prompts. To address these problems, we introduce large vision-language models (VLMs) to improve the diversity and quality of training data for efficient training. Our approach involves a continuous evaluation with VLMs to dynamically refine the training dataset by the discrimination, expansion, deletion, and mutation operations. This dynamic curation strategy ensures that only valuable data is retained, significantly reducing the volume of data for training. Experimental results demonstrate that a mere 100k training samples are sufficient for the base model to gain similar performance to that of the target model, which is substantially fewer than the 11 million samples required by the baseline method.

To explore the second question, we applied EvolveDirector to train the base model to approach multiple recent most advanced models in a unified framework, including DeepFloyd IF [11], Playground 2.5 [12], Stable Diffusion 3 [3], and Ideogram [13]. For each text prompt, we invoke advanced models to generate their images, and the VLM selects the best match to train our base model. The final trained model is named Edgen (EvolveDirector-Gen). The experimental results demonstrate that Edgen outperforms the advanced models mentioned above. Although our initial goal was to approximate these advanced models, we ultimately benefited from the VLM in choosing better training samples from their generated data, thereby achieving capabilities superior to any individual model.

The code of EvolveDirector and the model weights of Edgen are released to benefit the downstream tasks. Our contributions are summarized as: (1) Through experiments on massive data, we conclude that the generation abilities of a text-to-image model can be approximated through training on its generated data. (2) We propose the EvolveDirector, a framework that harnesses VLM to direct the training of the base model to learn generation ability from advanced models efficiently. (3) The trained text-to-image model Edgen outperforms the most advanced models.

2 Related Works

Text-to-Image Generation. To advance the overall quality of text-to-image generation, research efforts have been invested in exploring architectural improvement [10, 14, 15, 16] and generation paradigm advancement [17, 18, 3], *etc*. Diffusion models stand out as the de facto text-to-image

 https://doi.org/10.52202/079017-3880

generation paradigm [19, 20, 1, 3, 21, 22, 23, 24, 25], noted for its scalability and stability [26]. They benefit numerous downstream tasks, spanning image editing [27, 28], video generation [29, 30, 31, 32], 3D content generation [33, 34], *etc*. Through the use of highly descriptive and aligned image-text pairs at a substantial scale, text-to-image models that excel in resolutions, safety control, and the capability to render accurate scenes are obtained, *e.g.*, Imagen [1], Midjourney [4], DALL·E 3 [2], Stable Diffusion 3 [3], and Ideogram [13]. However, the exceptional capabilities of most advanced models have led providers to restrict access, typically offering only APIs, which limits their widespread and equitable use. In this paper, we aim to fill this gap by leveraging advanced VLMs to direct base model to replicate the functionality of advanced models. In contrast, some works propose to motivate models to learn from their self-generated images [35, 36].

Evaluating T2I Generation with VLMs. Some automatic evaluation methods [37, 38, 39] are propoesd to combine the LLMs and VQA models to evaluate the generated contents. Thanks to the substantial advancement of large language models [40, 41, 42, 43, 44, 45], the capabilities of VLMs are largely boosted [46, 47, 48, 49, 50, 51]. Utilizing these enhanced capabilities, methods building on VLMs are designed to facilitate the evaluation of text-to-image generation [52, 53, 54, 37, 55]. Notably, the recent research effort, Gecko [54], demonstrates the practicality of leveraging pre-trained LLMs [44] and VLMs [45] for systematic evaluation of text-to-image generative performance, spanning aspects such as text rendering, relational generation, attribute generation, *etc*. However, previous research requires fine-grained evaluation across various aspects, which remains challenging. In EvolveDirector, we simplify this process by requiring only pairwise comparisons, which enables a stable and reliable performance, facilitating the approaching of advanced text-to-image models.

Knowledge Distillation. KD [56] aims to transfer knowledge from well-trained teacher models to a simpler student model. The primary focus of most research in KD lies in investigating distillation losses with the output predictions of teacher model [57, 58], intermediate feature maps [59, 60], or feature correlations [61, 62] to distill knowledge. Recently, distillation methods based on diffusion models have garnered attention, primarily by distilling outputs from intermediate steps of the diffusion process to expedite the sampling process [63, 64, 65, 66]. Despite sharing the same goal of approaching the performance of advanced models, we would like to highlight our training paradigm is an orthogonal procedure to KD. To approach the advanced models with only APIs available, we avoid the need for distillation losses or acquiring intermediate results and instead choose a data-driven paradigm. Furthermore, our designed paradigm is also orthogonal to dataset distillation [67] as we evaluate and refine data during training rather than relying on a pre-existing large dataset.

Online Learning. Online learning has garnered significant interest for its capacity to enable models to adapt to real-time and dynamic data scenarios [68, 69]. This attention extends to a variety of real-world tasks, including semi-supervised learning [70, 71], unsupervised learning [72, 73], and continual learning [74, 75, 76], *etc*. In EvolveDirector, we harness the potential of online learning and powerful VLMs to evolve models towards advanced generation models, offering an efficient and scalable framework capable of dynamically adapting to evolving data.

3 Method

In this section, we outline the EvolveDirector framework in Sec. 3.1, describe the detailed operations of the VLM within this framework in Sec. 3.2, and discuss the training strategies in Sec. 3.3.

3.1 Overview of EvolveDirector

The proposed framework EvolveDirector, as shown in Fig. 2, comprises three parallel processes: (1) interacting with advanced T2I models to get training images, (2) maintaining the dynamic training set empowered by the Vision-Language Model (VLM), (3) training the base model on the dynamic training set. The dynamic training set is updated by the VLM and advanced T2I models, to ensure the data are high value for training, thereby achieving efficient training and reducing the overall required data volume.

Interaction with Advanced Models. While the model configurations and weights of many advanced T2I models are not publicly accessible, they often provide APIs for interactions. In EvolveDirector, we interact with these APIs by submitting text prompts and receiving the corresponding generated images. The one that aligns with the given text prompt better will be selected by the VLM and

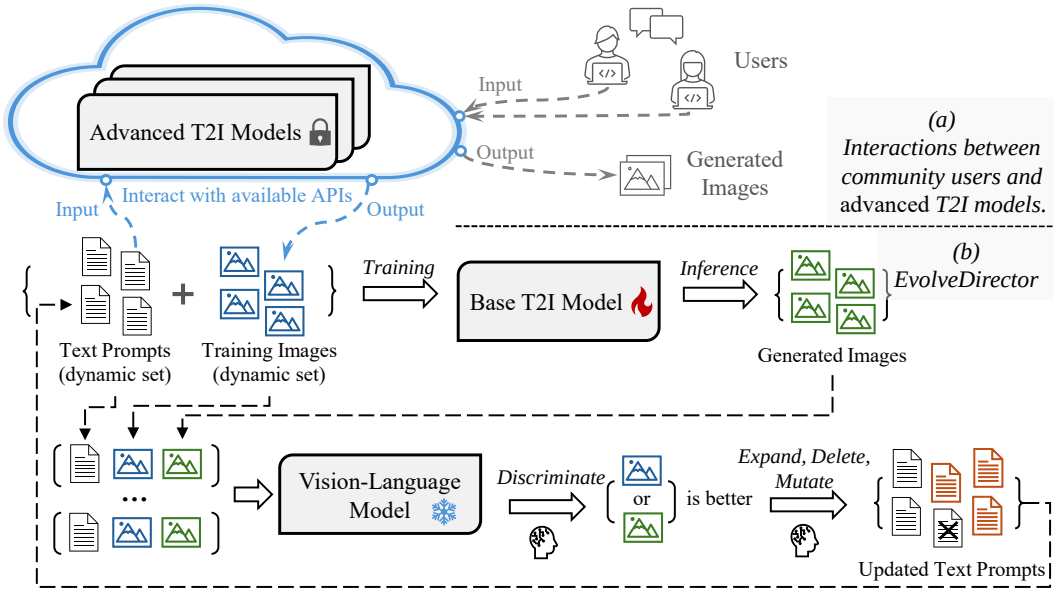


Figure 2: The overview of the proposed framework EvolveDirector. (a) Advanced T2I models provide accessible APIs, allowing users to input text prompts and get the generated images. (b) The base model is trained on the dynamic dataset, consisting of text prompts and corresponding images generated by advanced models via API calls. The VLM continuously evaluates the base model and, according to its performance, dynamically updates and refines the dataset through discrimination, expansion, deletion, and mutation operations based on its evaluations.

included in the training set. The selection criteria and details of these advanced models are elaborated in Sec. 4.4.

Dynamic Training Set. The training set is dynamically updated during the training of the base model. It continuously incorporates high-value samples, *i.e.* text prompts on which the base model underperforms compared to advanced models. Simultaneously, it dynamically excludes low-value samples, *i.e.* those on which the base model performs comparably to the advanced models. The VLM evaluates the value of samples, with a detailed procedure outlined in Sec 3.2. The advanced T2I models are continuously called to generate new images for the new text prompts and update them into the dynamic training set.

Training of the Base Model. Based on the dynamic training set, we train a Diffusion Transformer (DiT) [10] as our base text-to-image model. Specifically, the model is built upon the improved architecture proposed by PixArt-α [9]. Besides, we incorporate layer normalization after the Q (query) and K (key) projections in the multi-head cross attention blocks [77] to further stabilize the training, $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{f_Q(Q) \cdot f_K(K)^T}{\sqrt{d_k}}\right) V$, where $f_Q(\cdot)$ and $f_K(\cdot)$ are the layer normalizations after the projections.

3.2 Vision-Language Model as Director

The VLM acts as a director to guide the construction of more valuable dynamic datasets. To simplify the task difficulty and better leverage the capabilities of VLM, we present it with a choice of two images as a multiple-choice question, as shown in the top left corner of Fig. 3. One image $I_{advanced}$ is generated by the advanced model, while another one I_{base} is generated by the base model, respectively. In practice, the order of the two images is randomized. VLM is called to compare them and decide which one aligns better with the given text prompt T. Regarding the choice of VLM, there are two potential scenarios as follows.

(1) $I_{advanced}$ outperforms I_{base}. The inferior performance on this text prompt indicates that the base model is still under-trained. Therefore, EvolveDirector utilizes the VLM to generate more N_S variations of this text prompt T, as shown in the lower-left corner of Fig. 3. Then the advanced T2I

https://doi.org/10.52202/079017-3880

model generates corresponding images to expand the dynamic training set, as shown in the right side of Fig. 3. The original samples will continuously be involved in the training.

(2) $I_{advanced}$ does not outperform I_{base}. If the base model is comparable with the advanced model, indicating sufficient learning, the VLM will remove that prompt T as a low-value sample from the set to economize on training resources.

Besides the expansion and deletion operations, EvolveDirector also performs mutation operations with a certain probability. This operation permits the VLM to generate more diverse text prompts independent of any existing ones, thereby encouraging the model to explore and learn from a broader domain of text prompts.

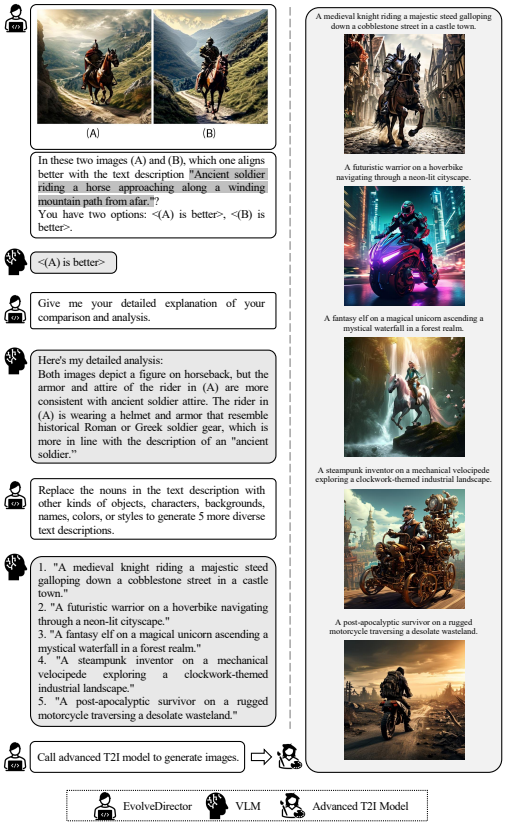


Figure 3: An example of the interaction between the EvolveDirector, VLM, and advanced T2I model. For brevity, auxiliary instructions to the VLM are omitted in this figure.

3.3 Training Strategies

Online Training. To boost training efficiency, we develop EvolveDirector as an online training framework. Specifically, the base model undergoes uninterrupted training, without pausing for the advanced model or the VLM to execute. Instead, it sends a command to start VLM evaluation every 100 epochs, termed a checking epoch. At the checking epoch, a subset of text prompts is sampled from the dataset with a specific ratio R_S. They are fed into a replica of the base model to generate images. Then the VLM evaluation and the subsequent execution of EvolveDirector begin, as illustrated in Fig. 3. Finally, a command is sent to the trainer of the base model to update the dynamic dataset. A detailed analysis of hyperparameters, including select ratio R_S and extension number N_S, is provided in the supplementary.

Stable Training. In our task setup, the scale of training data is significantly less than the million scale. Under this circumstance, the original architecture [9] demonstrates considerable instability during training, and the generation collapse is observed. Thus we follow the work [77] and apply the layer normalization after the query and key projections to improve the training stability. The experimental results detailed in the supplementary demonstrate the effectiveness of this adaptation.

Multi-Scale Training. The ability to generate images across various scales and aspect ratios is a significant capability of advanced T2I models. To facilitate more efficient training, we initially train the base model on images with a fixed resolution of 512px. Subsequently, we extend the training to images of higher resolution with multiple aspect ratios, thereby enabling the model to generate multi-scale and multi-ratio images. Following [9], we construct buckets with different aspect ratios and image sizes. In EvolveDirector, for each text prompt, a size bucket is randomly sampled from the buckets and advanced T2I models are called to generate images with this size. To avoid generation collapse, if the selected bucket falls outside the optimal size range of the advanced model, it will be resized to the closest appropriate size.

4 Experiments

4.1 Training Details

We train the base model on 16 A100 GPUs for 240 GPU days, with a batch size of 128 and 32 for images at 512px and 1024px resolution, respectively. The VLM evaluation process is distributed across 8 A100 GPUs to facilitate its speed. The open-source advanced models are deployed on our

devices with simulated APIs for interaction. For closed-source models, EvolveDirector interacts with them through their public APIs.

4.2 Selection of Vision-Language Models

The ability of the VLM to analyze images is crucial to the EvolveDirector framework. There are multiple powerful VLMs available, including CogVLM [82], CogAgent [78], Qwen-VL (Qwen-VL-Chat, Qwen-VL-Plus, and Qwen-VL-Max) [42], InternVL [79], LLaVA and LLaVA-Next [51, 80], and GPT-4V [81]. We evaluate their newest version on 600 pairs of questions to calculate the alignments with human raters. Each question consisted of a text prompt and two images generated based on that prompt, presented to 5 different human raters. The VLMs are tested in two aspects: (1) Discrimination, and (2) Expansion. To be more specific, (1) initially, the VLMs were required to select which image aligned more closely with the given text prompt. The output is scored by 0 for wrong or 1 for correct by human raters. (2) Subsequently, they are instructed to generate more variations of the given text prompt. The score of "Accuracy" evaluates whether the generated text prompts contain any linguistic errors and whether they are in the same syntactic structure as the given text prompt. The score of "Diversity" evaluates the diversity among the generated text prompts. These two scores range from 1 (worst) to 5 (best). The results are shown in Tab. 1. The LLaVA-Next and GPT-4V achieve the top performance. Considering that LLaVA-Next is totally free to use, we select it as the VLM for EvolveDirector.

Table 1: Alignment of VLMs with Human Preferences. The best value is highlighted in blue, and the second-best value is highlighted in green.

	Discrimination	Expansion	
		Accuracy	Diversity
CogAgent [78]	0.675	3.81	3.76
Qwen-VL-Max [42]	0.825	4.72	4.56
InternVL [79]	0.793	4.70	3.67
LLaVA-Next [80]	0.840	4.85	4.69
GPT-4V [81]	0.847	4.82	4.72

4.3 Ablation Studies

Models. For ablation studies, we select the Pixart-α [9] as the unified advanced model to approach. We utilize a DiT model pre-trained on publicly accessible data, *i.e.* the ImageNet and SAM dataset, as the base model.

Metrics. We sample $10,000$ text prompts to feed into models trained under different ablation settings to generate images and calculate their FIDs with the images generated by the advanced model. These text prompts are not seen by these models in the training stages. To conduct human evaluation, we randomly selected 300 sets of text prompts and their corresponding generated images, which are $2,400$ in total. These images are paired in twos, each consisting of one image from the advanced model and one from an ablation model corresponding to the same text prompt, arranged in random order. Then they are presented to 5 different human raters to choose which image matches the text prompt better. In this manner, each ablation model is paired with the advanced model for comparison, and their selected ratios are recorded to reflect their relative performance compared to the advanced model.

Table 2: Ablation Studies. The best value is highlighted in blue, and the second-best value is highlighted in green.

Discrimination	Expansion	Mutation	Data Scale	FID ↓	Human Evaluation ↑
✗	✗	✗	11M	7.36	48.89
✗	✗	✗	1M	11.49	39.44
✗	✗	✗	100K	15.19	32.22
✓	✗	✗	100K	15.41	32.61
✗	✓	✗	100K	13.05	36.44
✓	✓	✗	100K	7.61	47.00
✓	✓	✓	100K	7.45	48.53

Results. The results of FIDs and human evaluation are shown in Tab. 2. (1) *Directly Training on Generated Data.* The first three models were directly trained on images generated by the advanced model, using image quantities of 10 million, 1 million, and 100 thousand respectively. The experimental results show that the model trained on 10 million data reaches a comparable level to the advanced model in terms of human preference (48.89% V.S. 51.11%), and achieves a low FID score (7.36). This indicates that the generative capabilities of the advanced model can be learned through training on its large-scale generated data. However, if the training data is reduced to 10% or even 1% of the original amount, the performance of the trained model will significantly decrease.

(2) *Training with EvolveDirector.* In the last four rows of Tab. 2, models trained with EvolveDirector under different ablation settings are evaluated. The upper bound of data volume is set to 100k for all models. The first model is trained on an initial number of 100K data and dynamically deletes data. The last three data models start training on an initial number of 2K data and dynamically

https://doi.org/10.52202/079017-3880

add new data to learn. The first model applies the *"Discrimination"* function of the VLM model to discriminate the generated samples of the base model. Samples comparable to the output of the advanced model are removed from the training dataset. The results show that dynamically deleting samples does not cause much performance degradation. The second model does not use the VLM to evaluate the base model but instead randomly selected training samples for *"Expansion"*, i.e. generating more variants of given prompts and training samples. The results show that this operation brought a slight performance improvement due to the expansion of text prompt diversity. The third model utilizes VLM to evaluate the base model and performs reasoned expansion and deletion based on the evaluation results. As the results indicate, there is a significant performance increase, which highlights the importance of the evaluation of VLM. Lastly, the complete version of EvolveDirector is tested, which further applies the *Mutation*, i.e. randomly generating entirely new text prompts with a 10% probability. This operation further improves the performance of the model, because it encourages the model to explore more diverse images. With the full version of EvolveDirector, the model trained on a dynamic dataset with a data cap of 100K, achieved performance comparable to the model trained on 10M generated data, indicating that the proposed framework could significantly reduce the amount of training data required to approach the performance of the advanced model.

It is worth noting that there is still a slight gap between the final model and the advanced model, which can be attributed to the law of diminishing returns. This occurs because it becomes increasingly difficult to identify truly high-value samples as the performance approaches that of the advanced model. However, this phenomenon vanishes when EvolveDirector learns from multiple advanced models simultaneously, as experiments in Sec. 4.4 demonstrated. This is because the larger performance gaps between multiple models facilitate the easier identification of high-value samples.

4.4 Approaching Advanced Models

We select several latest advanced models to approach their powerful generation ability, including Playground 2.5 [12], Stable Diffusion 3 [3], and Ideogram [13]. Playground 2.5 is famous for its aesthetic generative effects, while Stable Diffusion 3 and Ideogram are known for their strong performance in various aspects including text generation and multi-object generation. Besides, we select a relatively old model, DeepFloyd IF (but just released in April 2023) [11], for its amazing text generation ability. The base model is the same as the one in Sec. 4.3, and we continue to train the base model that has already approached the Pixart-α to approach the selected multiple advanced models. During training, EvloveDiretcor will feed each text prompt to all of them to generate corresponding images, and then use VLM to select the best one of them as the image from the advanced model. The selected image then undergoes evaluation as detailed in Sec.3.2.

Table 3: Select Ratios of Advanced Models. The highest value is highlighted in blue, and the lowest value is highlighted in gray.

	Overall	Specific Domain		
		Human	Text	Multi-Object
DeepFloyd IF [11]	18.57%	9.12%	21.21%	12.12%
Playground 2.5 [12]	25.71%	31.33%	9.09%	25.24%
Ideogram [13]	29.52%	29.18%	34.24%	33.36%
Stable Diffusion 3 [3]	26.19%	30.36%	35.45%	29.27%

Select Ratios of Advanced Models. We have calculated the proportions of images generated by these models that were selected by the VLM in the training stage, as shown in Table 3. Among the generated images, those generated by Ideogram are selected with the highest ratio. Besides, we evaluate the select ratios of images in specific domains, such as human generation, text generation, and multiple object generation. Examples of these three types are shown in Fig. 5. The results in Table 3 show that different advanced models achieve the highest select ratios in different specifics. For example, for generating text in images, the Stable Diffusion 3 outperforms others, while the Playground 2.5 is much worse than others. This may be caused by that the internal training data of Playground 2.5 does not include sufficient high-quality images with text in them. This also demonstrates the significance of diverse high-quality data.

Quantitative Comparison. To evaluate the performance of the trained Edgen, the base model, and the advanced models, we sample text prompts and feed them into each model to generate images. One image generated by Edgen and one image generated by the comparison model are combined together. Finally, same with the scale of comparison in previous works [9], 300 text prompts and 1800 image combinations are shown to human raters to select which one aligns with the given text prompt better. Each combination is evaluated by 5 different human raters. The results are shown in Fig. 4. We first analyze the performance of each model across all tested text prompts, as shown on the left side of Fig. 4. It is noteworthy that during the training, the VLM selects the best ones among

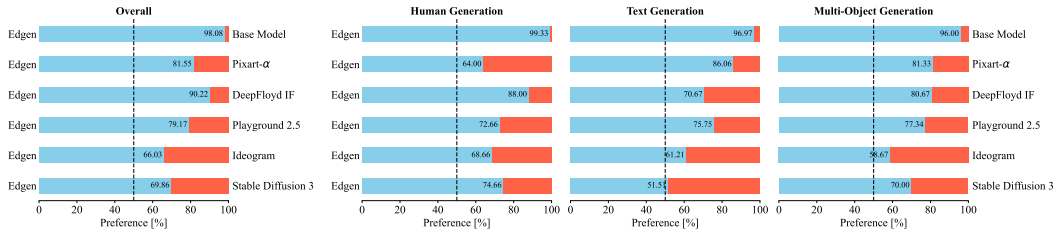


Figure 4: Human evaluation of the images generated by the base model, Edgen trained by the proposed EvolveDirector, and multiple advanced models.

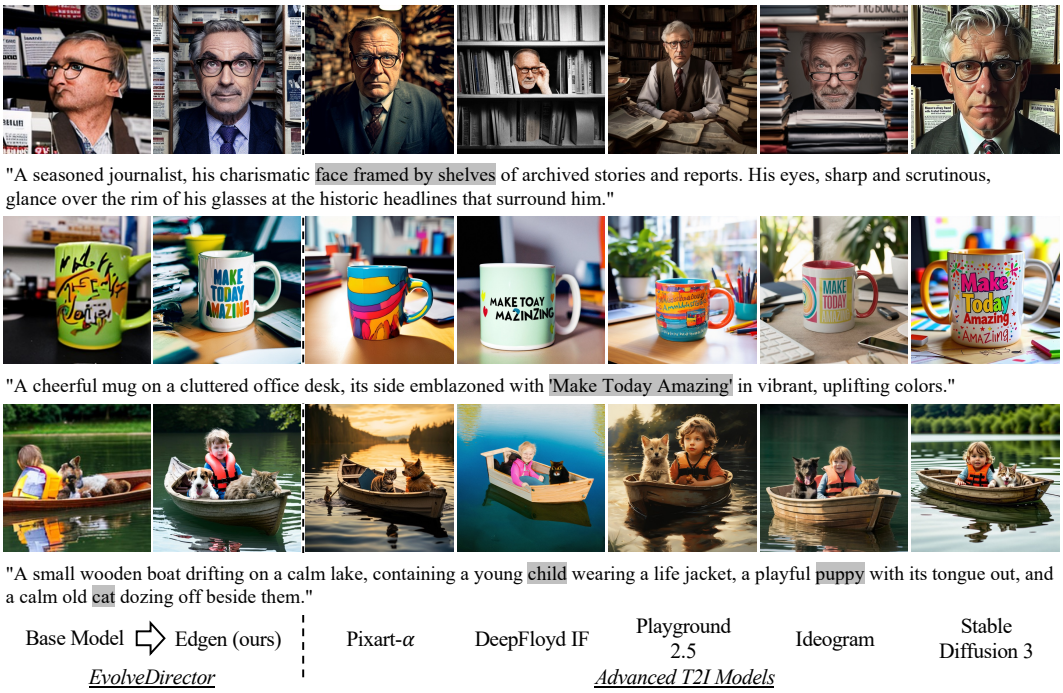


Figure 5: Images generated by the base model, Edgen trained by our EvolveDirector, and multiple advanced models. The results in three rows showcase the generation of human, text, and multi-object.

the images generated by multiple advanced models to be used as training data. Therefore, although EvolveDirector initially aims to train the base model to approach them, the final trained model Edgen outperforms all of the advanced models. Besides, we evaluate the performance of these models on specific types of text prompts, with the results displayed on the right side of Fig. 4.

Qualitative Comparison. In Fig. 5, we showcased three groups of generated images. The three rows of results respectively demonstrate the generation capabilities for humans, text, and multiple objects. For the first group, only the images generated by Edgen, DeepFloyd IF, and Ideogram successfully reflect the “face framed by shelves of ...”. As shown in the second row, only Edgen, Ideogram, and Stable Diffusion 3 have the ability to generate correct text in images. For the results shown in the third row, three objects need to be generated, i.e. the child, puppy, and cat. Only Edgen and Ideogram success and other models lost some objects. These results show that Edgen has already learned powerful generation abilities and outperforms some advanced models.

5 Conclusion

In this paper, we propose EvolveDirector, a framework that targets approaching the generation capabilities of advanced text-to-image models by only utilizing their publicly accessible APIs. By harnessing the capabilities of large vision-language models for evaluating image-text alignment,

https://doi.org/10.52202/079017-3880

EvolveDirector significantly reduces the volume of training data required, thus saving considerable training costs, especially those associated with API usage. Experimental results demonstrate that the resultant model Edgen, inheriting the generation capabilities from multiple advanced models, achieves superior performance in various aspects. The limitations and future work are discussed in the supplementary.

Acknowledgements

This project is supported by the Mike Zheng Shou's Start-Up Grant from NUS. It was partially done by Zhao Rui during his internship at Alibaba Group, and Zhao Rui is partially supported by the Alibaba Research Intern Program.

References

[1] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[4] Midjourney. Midjourney. https://www.midjourney.com, 2023.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[8] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, et al. Journeydb: A benchmark for generative image understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[11] Stability AI. Deepfloyd if. https://github.com/deep-floyd/IF, 2023.

[12] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

[13] Ideogram. Ideogram. https://ideogram.ai/, 2024.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[15] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[16] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[19] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, et al. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023.

[21] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[22] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-\sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv preprint arXiv:2403.04692*, 2024.

[23] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*, 2024.

[24] Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model. *arXiv preprint arXiv:2402.12376*, 2024.

[25] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[26] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[27] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[28] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[29] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.

[30] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023.

 https://doi.org/10.52202/079017-3880

[31] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7621–7630, 2024.

[32] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024.

[33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[34] Rui Zhao, Wei Li, Zhipeng Hu, Lincheng Li, Zhengxia Zou, Zhenwei Shi, and Changjie Fan. Zero-shot text-to-parameter translation for game character auto-creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21023, 2023.

[35] Jiao Sun, Deqing Fu, Yushi Hu, Su Wang, Royi Rassin, Da-Cheng Juan, Dana Alon, Charles Herrmann, Sjoerd van Steenkiste, Ranjay Krishna, et al. Dreamsync: Aligning text-to-image generation with image understanding feedback. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023.

[36] Jialu Li, Jaemin Cho, Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. Selma: Learning and merging skill-specific text-to-image experts with auto-generated data. *arXiv preprint arXiv:2403.06952*, 2024.

[37] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023.

[38] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

[39] Jay Zhangjie Wu, Guian Fang, Haoning Wu, Xintao Wang, Yixiao Ge, Xiaodong Cun, David Junhao Zhang, Jia-Wei Liu, Yuchao Gu, Rui Zhao, et al. Towards a better metric for text-to-video generation. *arXiv preprint arXiv:2401.07781*, 2024.

[40] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[41] OpenAI. GPT-4 technical report, 2023.

[42] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[44] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[45] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[46] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[47] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[48] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[49] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[52] Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

[53] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36, 2024.

[54] Olivia Wiles, Chuhan Zhang, Isabela Albuquerque, Ivana Kajić, Su Wang, Emanuele Bugliarello, Yasumasa Onoe, Chris Knutsen, Cyrus Rashtchian, Jordi Pont-Tuset, et al. Revisiting text-to-image evaluation with gecko: On metrics, prompts, and human ratings. *arXiv preprint arXiv:2404.16820*, 2024.

[55] Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-image generation. *arXiv preprint arXiv:2310.18235*, 2023.

[56] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[57] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018.

[58] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198, 2020.

[59] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171, 2019.

[60] Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3779–3787, 2019.

[61] Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Heterogeneous knowledge distillation using information flow modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2339–2348, 2020.

 https://doi.org/10.52202/079017-3880

[62] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

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

[64] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

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

[66] Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023.

[67] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

[68] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

[69] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Random averages, combinatorial parameters, and learnability. *Advances in Neural Information Processing Systems*, 23, 2010.

[70] Islam Nassar, Munawar Hayat, Ehsan Abbasnejad, Hamid Rezatofighi, and Gholamreza Haffari. Protocon: Pseudo-label refinement via online clustering and prototypical consistency for efficient semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11641–11650, 2023.

[71] Salah Ud Din, Junming Shao, Jay Kumar, Waqar Ali, Jiaming Liu, and Yu Ye. Online reliable semi-supervised learning on evolving data streams. *Information Sciences*, 525:153–171, 2020.

[72] Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. Online deep clustering for unsupervised representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6688–6697, 2020.

[73] Qi Qian, Yuanhong Xu, Juhua Hu, Hao Li, and Rong Jin. Unsupervised visual representation learning by online constrained k-means. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16640–16649, 2022.

[74] Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, and Hongming Shan. Online prototype learning for online continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18764–18774, 2023.

[75] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

[76] Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8250–8259, 2021.

[77] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

[78] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*, 2023.

[79] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[80] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[81] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[82] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

 https://doi.org/10.52202/079017-3880

A Limitation and Future Work

While EvolveDirector achieves significant strides in approximating the generation capabilities of advanced text-to-image models, the resultant model can face challenges related to bias. Our model can inadvertently inherit biases present in the images generated by the advanced models. Additionally, our reliance on VLMs to evaluate generated images could introduce their own biases into the selection process. To mitigate these issues, it is better to integrate debiasing methods and incorporate human feedback in future developments, aiming to enhance the robustness and fairness of EvolveDirector.

B Broader Impacts

Our research leverages the foundational capabilities of large vision-language models to reproduce advanced text-to-image models. This initiative represents a pioneering approach to make state-of-the-art generative technologies more accessible and cost-effective. Our methodology substantially reduces the volume of data required for model training, thereby reducing computational demands and minimizing environmental impacts associated with image generation. The resultant model Edgen has the potential to revolutionize the creation of digital media owing to its inheritance of capabilities from multiple advanced models. For instance, its capabilities enable the creation of diverse media content and marketing materials, such as logos, by effectively rendering text into realistic and imaginative visual representations.

However, we also acknowledge potential negative societal impacts. As with many generative technologies, there is an inherent risk of bias in terms of gender and race in the generated content. Furthermore, the improper use of the resultant model, such as inputting harmful or obscene text prompts or creating unauthorized or deceptive images of public figures, could facilitate misinformation or impersonation. To mitigate these risks, it is imperative to implement thorough oversight to secure its ethical use.

C Layer Normalization

As discussed in the main paper, due to the scale of training data being significantly less than the previous million scale, the original architecture [9] is not stable during training. As shown in Fig. A (a) and (b), when the training steps increase from 0.5k to 2k, the main object in the generated image begins to be destroyed. To address this problem, we incorporate layer normalization in the multi-head cross-attention. To be more specific, the layer normalization is added after the Q (query) and K (key) projections. This can significantly stabilize the training of the base model, Training model with the layer normalization can avoid generating destroyed images, as shown in Fig. A (c).

(a) Training steps: 0.5k; w/o layer normalization after QK projections | (b) Training steps: 2k; w/o layer normalization after QK projections | (b) Training steps: 2k; w/ layer normalization after QK projections

Figure A: Illisturation of the impact of incorporating layer normalization after QK projections on generation.

D Detailed Instructions to VLM

Instruction for Discrimination. We input the combination of two images and instruct the VLM to choose which one aligns with the given text prompt better. The instruction is as follows, “In these two images (A) and (B), which one aligns better with the text description "*text prompt*"? You have two options: <(A) is better>, <(B) is better>. Simply state your choice, no need for a detailed explanation."

Instruction for Expansion. The instructions for expansion are as follows, “Replace the nouns in the text description: "*text prompt*" with other kinds of objects, characters, backgrounds, names, colors, or styles to generate *extend num* more diverse text descriptions. Arrange them in the format of a list ["Text description 1", "Text description 2", ...].”

Instruction for Mutation. The instructions for expansion are as follows,

"Now, exercise your imagination to generate one new text description for visual content that is completely unrelated to the previous images. It should have a completely different structure from the previous descriptions. *enhanced prompt*. Arrange it in the format of a list just like ["xxxxx"]."

The *enhanced prompt* is randomly sampled from the following ones:

- "It should be rough and short."
- "It should contain less than 30 words and be highly detailed."
- "It should contain over 30 words with different granular levels of detail."
- "It should contain over 50 words with a lot of details."

Structure the Outputs of VLM. The diversity of output formats from VLM can pose challenges for automated parsing. We found that by providing specific instructions to the VLM, its output format can be standardized. Specifically, when prompting the VLM to generate more text prompts, we offer instructions such as "*Arrange them in the format of a list ["Text description 1", "Text description 2", ...].*" This approach directs the VLM to generate outputs in a consistent format.

E Hyper-parameters Setting

The hyper-parameters of EvolveDirector include the select ratio of images selected from the dynamic training set for evaluation (select ratio R_S) and the number of training samples expanded based on a single sample (number of extensions N_E). These two parameters mainly affect the growth rate of training samples in the dynamic training set. Too fast a growth rate can lead to the need to generate a large number of images in the early stages of training, thereby quickly increasing training costs, while too small an increase can cause the model to explore the diversity of different samples too slowly. To balance the training costs and the diversity of samples explored by the model, it is necessary to set reasonable values for R_S and N_E. Due to the high cost of searching for the most reasonable parameter combination (fully training the model under each combination requires 240 A100 GPU days), we only explored a limited number of parameter combinations and stopped the training as long as we observed the trend in data growth.

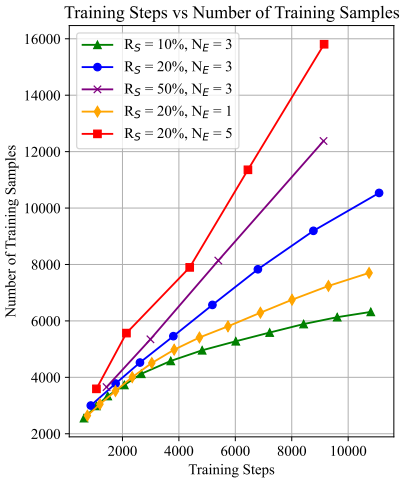


Figure B: The effect of values of hyper-parameters on training data growth rate.

As shown in Fig. B, five kinds of combination of select ratio R_S and number of extensions N_E are evaluated, we can observe that if the R_S is too large, the volume of training data increased rapidly with increasing cost too fast. Despite a larger number of extensions N_E encouraging the model to explore new training samples greedily, it also increases the training cost rapidly. In practice, we found that setting it to 3 is enough to explore new samples. Finally, we select the combination of $R_S = 20\%$ and $N_E = 3$ to achieve a good balance of training cost and exploration of diverse samples. Besides, since the mutation rate R_M does not affect the growth rate of the training set significantly, we simply set it to 10% as an augmentation of the exploration of samples.

F Implementation Details

The model is trained by the AdamW optimizer with a learning rate of $2e-5$ and with a gradient clip of 1.0. A constant learning rate schedule with 1000 warm-up steps is used. The gradient checkpointing is used to save the VRAM. For generating 512px images, we train the base model with 100K training steps, and for 1024px image generation, we train the model with 20K training steps.

The default number of training samples at the beginning of training is 20K, which will gradually grow to the upper bound. For the ablation studies in Sec. 4.3, we set the upper bound to 100K to explore the extreme performance of EvolveDirector. The initial text prompts are sampled from those captioned for the SAM dataset [9]. For approaching multiple advanced models in the Sec.4.4, we set

https://doi.org/10.52202/079017-3880

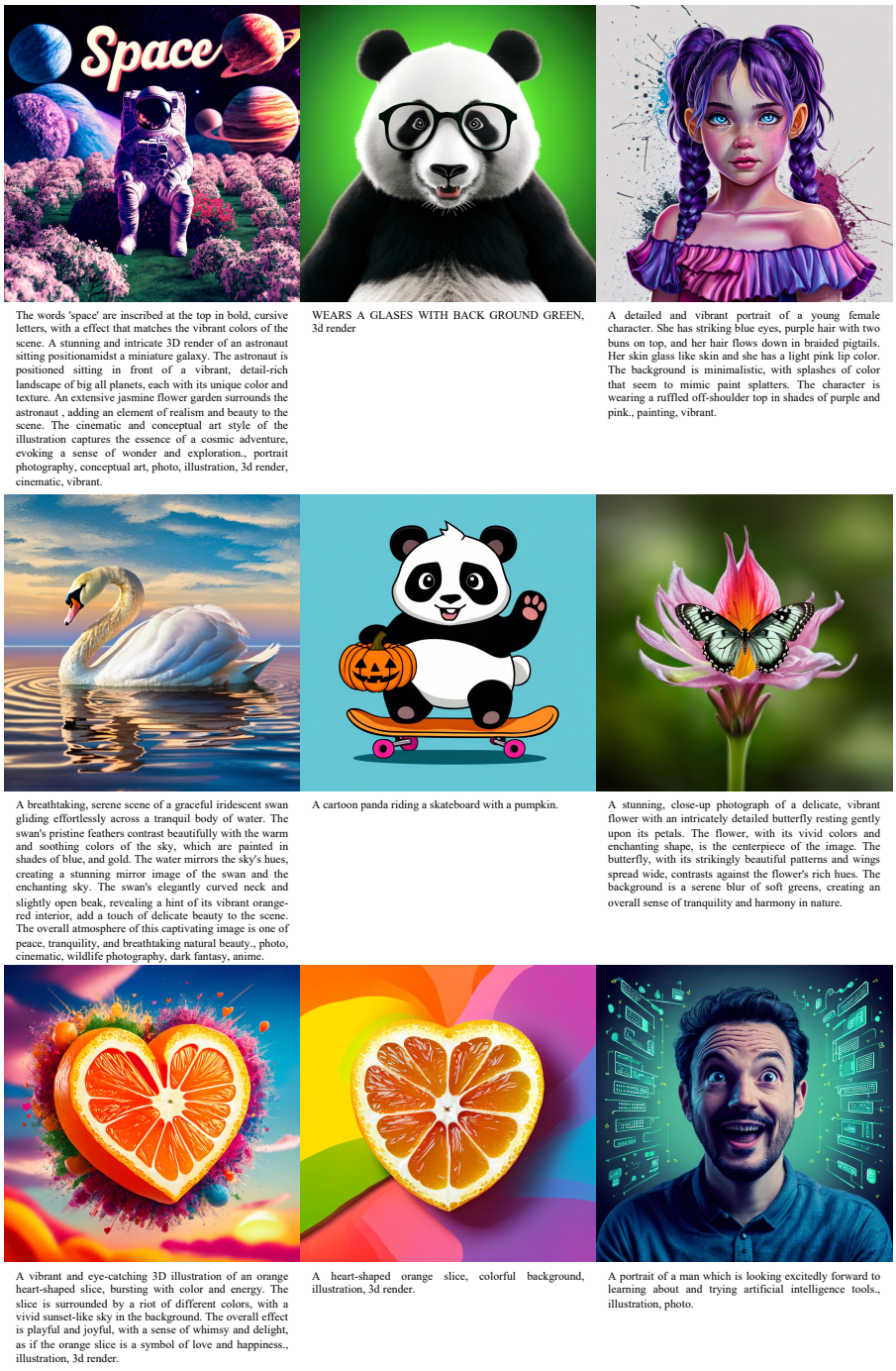


Figure C: Additional results 1/3. Images generated by the Edgen.

the upper bound to 300K to ensure a more sufficient learning of much powerful generation abilities of the most advanced models. The initial text prompts are randomly selected from both the SAM captioning dataset and the community-sourced text prompts.

G Additional Results

A strange high-definition photo of a cute orange cat, wearing red-rimmed glasses and engrossed in his smartphone, sitting on the toilet. The bathroom was decorated with funny posters, including one of a man laughing and another of a woman. The golden hour sun casts a warm, inviting light through the open windows, illuminating the view. Various items such as books, toilet brushes and soap scattered around add to the unique atmosphere. Potted plants on the floor and natural light streaming in bring a touch of freshness to this cheerful 8k image.

A colorful comic-style linocut illustration of a captivating pop art-inspired composition, a female model walking on a pink carpeted runway. She is wearing a unique, off-shoulder dress with a draped design that exposes parts of her body. The dress is in a light blue color and has a sheer, transparent section at the bottom. The model has a short, blonde hairstyle and is looking directly at the camera. The background is a mesmerizing blend of multi-colored abstract shapes and mandala patterns, creating a harmonious and visually striking scene.

The head of a lion with majesty and elegance. Chin high and pride. Her hair is long and voluminous, and at its ends you can see trails of light and fire coming out of her hair. In the background you can see the Savannah with a few animals in the distance such as giraffes or birds.

Illustrate a delicate scene where a small, intricately detailed hummingbird with a speckled blue-green throat hovers near a cluster of large, soft pink flowers. The setting is dreamlike, with a misty, orange-toned background suggesting either dawn or dusk. Focus on the light reflecting off the bird's iridescent feathers and the gentle sway of the flower petals., painting.

A charming and captivating illustration of a Maltese puppy wearing a colorful, enchanting witch's hat, attend a glamorous garden party under the moonlight. The puppy is smiling and enjoying the warm atmosphere, surrounded by vibrant, glowing flowers. The background is filled with colorful bokeh, and the scene is bathed in a soft glow. Sharp details and a shallow depth of field create a sense of depth and dimension. The overall atmosphere is cozy and cinematic, with a touch of fairy tale enchantment.

An amusing and whimsical image of four multicolored gummy bears gathered under a small umbrella. A large spoon hovers above them, pouring a shower of sugar over the gummy bears, creating a sweet and playful scene. The background is a bright and colorful dreamscape, filled with an assortment of candies, pastel clouds, and delightful surprises.

A charming illustration of a dog with its mouth open, showing off its playful and friendly personality. The dog has thick fur and a mischievous glint in its eyes, with a wagging tail that exudes joy. The background features a cozy home with a warm fireplace, a sunny yard, and a colorful garden., illustration.

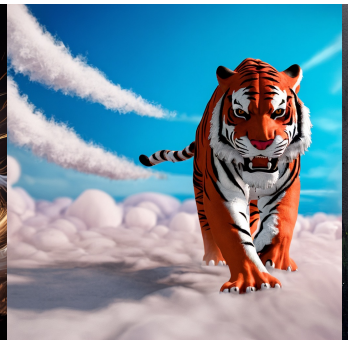
A breathtaking photo-realistic image of a confident and alluring old classy gentleman, striding down a bustling city street. Dressed in an elegant XIX century suit."

A captivating, cinematic photorealistic image of a group of people standing on a dramatic cliff, their silhouettes accentuated by the soft glow of the full moon. The ocean below is a raging sea with towering waves, creating a sense of awe and danger. The sky is a rich blend of deep blues and purples, with the moon casting a shimmering reflection on the water. The overall atmosphere is eerie yet mysterious, as if the group is witnessing a rare lunar event., photo, cinematic.

Figure D: Additional results 2/3. Images generated by the Edgen.

https://doi.org/10.52202/079017-3880

Majestic white-maned stallion, powerful stance, intricate golden plated armor, mythical aura, flowing mane and tail, regal posture, detailed armor engravings, elegant and fierce, dynamic pose, dramatic lighting casting shadows, vibrant contrast between brown and gold, fantasy setting, epic fantasy illustration, high-resolution digital art, Artstation showcase.



A striking and dynamic 3D render of Mount Pleasant High School's tiger mascot, walking determinedly towards the viewer on the right side of the screen. The tiger, has a focused, fierce expression as it strides forward, giving off an air of strength and resilience. The tiger is realistic. The background showcases a vibrant blue sky with soft, cotton-like clouds gently drifting by, creating a serene and motivational atmosphere. The tiger appears to be walking on clouds, giving the image a surreal and uplifting feel., photo, 3d render.

Close-up portrait, the watercolor loose dreamy young boy with platinum blonde hair, adorned with a wreath of dandelions. His serene, contemplative expression conveys a sense of introspection and melancholy. The boys simple sleeveless clothe contrasts with the somber background, creating a striking visual effect., lets nature take its course, soft, light coloring, side view. Damien Hirst, creative Commons attribution, painted in a dainty neutral style with soft light colors and few details.Art by Quentin Blake, Alberto Vargas, Zdzislaw Beksinski, painting, dark fantasy, illustration, anime, conceptual art, architecture.

A stunning beach art print capturing a serene coastal scene. In the foreground, a meticulously painted seagull is perched on a rock, its wings outstretched and eyes fixed on the horizon. The golden sand stretches out, dotted with footprints that seem to disappear into the distance. A vibrant sun is setting on the right, casting a warm, golden glow on the waves that gently lap at the shore., painting.

A serene and romantic image of a young woman with long brown hair and glasses, seated on a striped beach chair. She wears a yellow off-shoulder dress that complements her calm demeanor. The starry night sky above her is dotted with countless twinkling stars, and a distant cityscape adds a soft glow to the scene. The reflection of city lights on the water creates a tranquil and intimate atmosphere, perfect for contemplation.

A meticulously carved wooden relief depicting a little cute greydoodle puppie and a cute beagle.

A cityscape built on the back of a giant turtle.

2.5D CG, Charming, Rococo angel: Imagine a children's book illustration of a sweet angel in a flowing Rococo dress. Delicate wings flutter softly amidst a vibrant garden of blooming flowers. Her warm smile radiates as she strolls through a magical heaven, straight out of a fairytale. seed# 33663366, illustration, conceptual art.

In a chic loft apartment, a modern kitchen with stainless steel appliances and a marble island, an open living space with a plush sofa and eclectic artworks, floor-to-ceiling windows overlooking the city, and a spiral staircase leading to the bedroom.

Figure E: Additional results 3/3. Images generated by the Edgen.

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: Claims in the abstract and introduction 1 accurately reflect our contributions and scope.

   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 discuss the limitations of our work in Sec. A.

   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/10.52202/079017-3880

Justification: We do not include theoretical results.

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 provide the clear framework structure in Sec. 3 and training and implementation details in Sec. 4.1 and Sec. F.

   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?

Answer: [Yes]

Justification: The code and model weights are open-sourced, and the link is attached at the end of the abstract.

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 provide the detailed experimental setting in the front of each subsection of the experiment section, *i.e.* Sec. 4.2, Sec. 4.3 and Sec. 4.4.

   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: The experiments we do do not need error bars.

   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/10.52202/079017-3880

- 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 information on the computer resources in Sec. 4.1

   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: We have carefully reviewed the Code of Ethics and strictly adhered to it in our work.

   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 claim social impact in the appendix Sec. B.

    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.

- 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: The safeguards are checked by the advanced models internally.

    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: The code, data, and models are all properly credited and their license and terms of use are properly respected in our work.

    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/10.52202/079017-3880

- 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: [NA]

    Justification: We do not introduce new assets.

    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: We do not involve crowdsourcing nor 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: We do not involve crowdsourcing or 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.