Rapid Diffusion: Building Domain-Specific Text-to-Image Synthesizers with Fast Inference Speed

Bingyan Liu[1,2*] and **Weifeng Lin**[1,2*] and **Zhongjie Duan**[3,2] and **Chengyu Wang**[2†]
Ziheng Wu[2] and **Zipeng Zhang**[2] and **Kui Jia**[1†] and **Lianwen Jin**[1]
Cen Chen[3] and **Jun Huang**[2]
[1]South China University of Technology, Guangzhou, China
[2]Alibaba Group, Hangzhou, China
[3]East China Normal University, Shanghai, China
{eeliubingyan, eelinweifeng}@mail.scut.edu.cn, zjduan@stu.ecnu.edu.cn
{chengyu.wcy, zhoulou.wzh, zhangzipeng.zzp}@alibaba-inc.com
kuijia@gmail.com, eelwjin@scut.edu.cn, cenchen@dase.ecnu.edu.cn
huangjun.hj@alibaba-inc.com

Abstract

Text-to-Image Synthesis (TIS) aims to generate images based on textual inputs. Recently, several large pre-trained diffusion models have been released to create high-quality images with pre-trained text encoders and diffusion-based image synthesizers. However, popular diffusion-based models from the open-source community cannot support industrial domain-specific applications due to the lack of entity knowledge and low inference speed. In this paper, we propose *Rapid Diffusion*, a novel framework for training and deploying super-resolution, text-to-image latent diffusion models with rich entity knowledge injected and optimized networks. Furthermore, we employ BladeDISC, an end-to-end Artificial Intelligence (AI) compiler, and FlashAttention techniques to optimize computational graphs of the generated models for online deployment. Experiments verify the effectiveness of our approach in terms of image quality and inference speed. In addition, we present industrial use cases and integrate *Rapid Diffusion* to an AI platform to show its practical values. [1]

1 Introduction

Text-to-Image Synthesis (TIS) is a prevalent multi-modal task that aims to generate realistic images based on textual inputs, which supports real-world applications such as product appearance design and art creation. Apart from Generative Adversarial Network (GAN)-based approaches (Agnese et al., 2020), recently, pre-trained diffusion models (Rombach et al., 2022; Ramesh et al., 2022) have been proposed to create artistic images with qualities comparable to or better than those from humans.

Despite the exciting advancement, for industrial domain-specific applications, we suggest that popular latent diffusion models from the open-source community (such as the Stable Diffusion model series[2]) are incapable of supporting those applications. The reasons are are twofolds. i) For diffusion-based methods, a CLIP-based text encoder (or other similar models) is required to encode the input texts, providing conditional inputs for the U-Net model (Rombach et al., 2022). As entities (or objects) are usually the key elements for generated images, CLIP models pre-trained over text-image pairs collected from the Web may need more abilities of concept understanding and are challenging to capture the specific entity knowledge required for realistic image generation (Ma et al., 2022). ii) For industrial applications, the model inference speed and the computational cost are vital factors to be considered. The cumbersome computation of the iterative diffusion process is often the bottleneck of fast inference (Song et al., 2021). Therefore, obtaining knowledgeable diffusion models to generate high-resolution images with moderate parameter sizes and optimized implementations that support fast online inference is desirable.

To address the above issues, we propose *Rapid Diffusion*, a novel framework for the training and deploying text-to-image diffusion models with rich entity knowledge injected and networks optimized. In *Rapid Diffusion*, a knowledge-enhanced CLIP model is effectively trained for learning entity knowledge from knowledge graphs (KGs). To

*B. Liu and W. Lin contributed equally to this work.
†C. Wang and K. Jia are co-corresponding authors.
[1]The source code is publicly available in the EasyNLP framework (Wang et al., 2022). URL: https://github.com/alibaba/EasyNLP.
[2]https://stability.ai/blog/stable-diffusion-public-release

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 295–304

generate high-resolution images and avoid parameter explosion, we integrate an ESRGAN-based network (Wang et al., 2018) after the diffusion block for image super-resolution, instead of directly leveraging a large-scale hierarchical diffusion model. For online deployment, an efficient inference pipeline is designed with the neural architectures optimized based on FlashAttention (Dao et al., 2022). The Intermediate Representation (IR) of computational graphs built from the generated models are further processed by a recently released Artificial Intelligence (AI) compiler (Zhu et al., 2021).

In the experiments, we evaluate the effectiveness of *Rapid Diffusion* in terms of the qualities of generated images from multiple application domains and the model inference speed for online deployment. We also provide industrial use cases to show how our framework benefits real-world applications. In addition, we have integrated the proposed training and deployment workflows into an industrial, cloud-native AI platform to facilitate zero-code model training and elastic inference on distributed GPU clusters. In summary, the major contributions of this work are as follows:

- We propose the *Rapid Diffusion* framework for the training and deployment of domain-specific diffusion-based TIS models. Specifically, a new knowledge-enhanced model pipeline is designed for super-resolution TIS. An efficient inference pipeline is further designed to optimize the computational graphs of our model for faster model inference.
- Experiments over multiple domains show the effectiveness of *Rapid Diffusion* in terms of both image quality and inference speed, achieving an average FID score of 21.90 and $\times$1.73 acceleration ratio compared to all the counterparties.
- We demonstrate the industrial use case and the integration of *Rapid Diffusion* to a cloud-native AI platform to show its practical values for real-world applications.

2 Related Work

2.1 Text-to-Image Synthesis (TIS)

TIS is a multi-modal task of converting texts to images with the same semantic meanings. In the early years, traditional methods (Zhu et al., 2007) mainly focused on analyzing the correlations between sentences and images but could not generate new images on the pixel level. Generative Adversarial Network (GAN) (Goodfellow et al., 2014) was proposed in 2014 and became the mainstream approach in the image synthesis field (Agnese et al., 2020). GANs and their variants (Reed et al., 2016; Liu et al., 2022b) have proved their effectiveness in TIS but still lack the ability to generate high-resolution images. Diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015) have attracted the attention of researchers in recent years. Leveraging large-scale text-image datasets, pre-trained diffusion models (Rombach et al., 2022; Ramesh et al., 2022) become competitive with human painters. However, these diffusion models need help with the efficiency problem and more knowledge for the generation process.

2.2 Efficient Methods for Diffusion Models

Diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015) typically add noise to images (or latent tensors generated from images) and then learn to denoise step by step. The number of steps while training may be very large, making the sampling time-consuming. To improve the sampling efficiency, a recent study (Salimans and Ho, 2022) introduces knowledge distillation to diffusion models. This acceleration method can reduce the steps but requires additional training. Another family of methods tries to construct new samplers without further training. For example, DDIM (Song et al., 2021) uses a deterministic generative process to produce images much faster. Some numerical solvers, including forward Euler and linear multistep method (Butcher, 2000), are leveraged to reduce the steps (Karras et al., 2022). By using a pseudo numerical algorithm to solve differential equations on manifolds, PNDM (Liu et al., 2022a) further improves the generation quality within a few given steps. The better implementation of attention algorithms can speedup the process, which requires fewer IO accesses (Dao et al., 2022). Colossal-AI (Bian et al., 2021) accelerates the training speed of diffusion models and reduces the GPU memory usage for deployment. In addition, when diffusion models are deployed online, the amount of computation can be reduced with better-complied computational graphs of these models. For example, TensorRT[3] provides an inference optimizer and runtime that achieves lower

[3] https://github.com/NVIDIA/TensorRT

latency of model inference. In our work, we integrate various techniques from both modeling and engineering aspects to deliver better training and inference experiences for diffusion-based, domain-specific TIS applications.

3 The Proposed Framework

In this section, we formally present the techniques of the proposed *Rapid Diffusion* framework in detail. A brief overview of *Rapid Diffusion* is presented in Figure 1.

3.1 Model Architecture

As seen in Figure 1, our model converts input texts to high-resolution images by modeling the transformation and interaction between three representation spaces: i) knowledge-enhanced text embedding space, ii) latent space, and iii) pixel space.

3.1.1 Knowledge-enhanced Text Embedding Space

In this stage, we aim to encode the semantics of input texts to text embeddings. A common practice is to leverage the text encoder of CLIP (Radford et al., 2021), which jointly learns textual and visual representations in a unified space. Yet, CLIP pre-trained over plain text-image pairs may have weak representation power of entities. In our work, we leverage the 100 million text-image pairs from Wukong (Gu et al., 2022) as our multi-modal pre-training corpus, as our real-world applications mostly focus on the Chinese language. For entities, we leverage the largest Chinese KG available to us, i.e., OpenKG[4] (containing over 16 million entities and 140 million relation triples). During the CLIP pre-training process, the input representation of an entity token e appearing in a sentence of the Wukong corpus is augmented by: $\vec{e} = \vec{e}_{txt} + \vec{e}_{kg}$ where $\vec{e}_{txt}$ is the vanilla token embedding of the entity e, and $\vec{e}_{kg}$ is the KG embedding derived by the TransE algorithm (Bordes et al., 2013) due to its effectiveness and simplicity. Note that although we focus on the pre-training of Chinese Knowledge-enhanced CLIP (CKCLIP) models here, our method is language-invariant and can be applied to other languages with minor modifications.

During the fine-tuning process of our domain-specific TIS models, the parameters of the text encoder of our CKCLIP model are set to be trainable to capture more domain-related semantics. We further add some text prompts to the input text according to the application scenario (e.g., "the photo of [object]", "the picture of [object]") and obtain its CLIP representation as the conditional input to the next stage.

3.1.2 Latent Space

According to the given knowledge-enhanced text embedding $\vec{e}$, we use a latent diffusion model to generate image encoding with similar semantic meaning in a latent space. The model architecture is U-Net (Ronneberger et al., 2015) with a cross-attention mechanism capturing the textual conditioning information. In the training stage, the image x is encoded into the latent space and then we add Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$ to obtain x_t, where $t = 1, \ldots, T$ is the step in the diffusion process. The loss function of image reconstruction is formulated as follows:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{x,t\sim\mathcal{U}(1,\ldots,T),\epsilon\sim\mathcal{N}(0,1)} \left(||\epsilon - \epsilon_\theta(x_t, t)||_2^2 \right) \quad (1)$$

The generation process is the reverse of the diffusion process. Starting from the random Gaussian noise x_T, the latent diffusion model gradually denoises the latent tensor and successively calculates $x_{T-1}, x_{T-2}, \ldots, x_1$. To improve the correlation between the generated image and the input prompt, we use the classifier-free guidance (Ho and Salimans, 2021) to generate the corresponding images. Additionally, to avoid the efficiency problem caused by too large step T, we employ PNDM (Liu et al., 2022a) scheduler to reduce the steps. In our work, we also pre-train the latent diffusion model using the Wukong dataset (Gu et al., 2022) and fine-tune the model using every downstream dataset respectively.

3.1.3 Pixel Space

After generating the final latent x_0, a KL-regularized decoder $\mathcal{D}$ reconstructs the image from x_0 in the pixel space. In our design, the generated images are not necessarily in high resolution. Instead, an ESRGAN-based network (Wang et al., 2018) is applied after the decoder such that after a single forward pass, a corresponding high-resolution image can be generated. An alternative design choice is directly generating high-resolution images using the diffusion model. However, this setting can be sub-optimal to satisfy the requirements of moderate model size and fast inference

[4] http://openkg.cn/

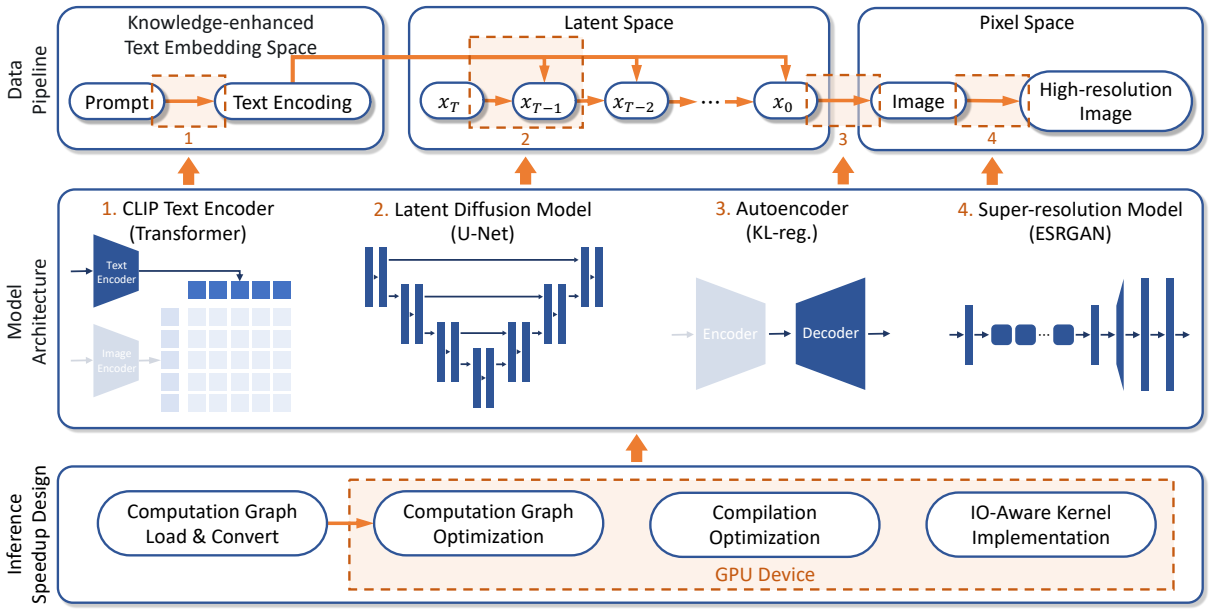


Figure 1: An overview of the *Rapid Diffusion* framework.

speed. Consider a base 256×256 diffusion-based U-Net model that employed for the TIS task, where one with a $256 \times 256 \longrightarrow 1024 \times 1024$ diffusion-based super-resolution U-Net model and the other with an ESRGAN-based model. The former contains more parameters and requires more steps for inference, resulting in its inference time being several times slower than the latter. Therefore, we adopt an ESRGAN-based model to generate high-resolution images efficiently.

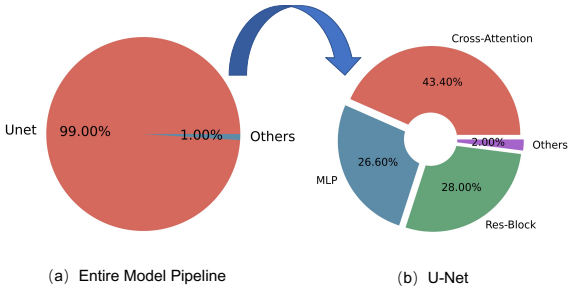


Figure 2: The profiling result of model inference in the percentage of the entire CUDA time.

3.2 Inference Speedup Designs

The inference process of the proposed model in this paper consists of three main components. We profile the inference speed of the original PyTorch model in eager mode and observe that the bottleneck is primarily located in the loop of the U-Net model, where the cross-attention computation dominates the inference time. The profiling result can be seen in Figure 2. To resolve this issue, we incorporate automatic slicing and compilation optimization techniques to optimize the entire pipeline in an end-to-end manner and introduce an IO-aware attention implementation to enhance the inference performance further.

3.2.1 Compilation Optimization

Our algorithm generates various low-level runtime flows for models with dynamic shapes on specific devices. It is achieved by enhancing a set of IR to create a complete dynamic shape representation (Zhu et al., 2021). For the operations with intensive memory access, we fully utilized shared memory to design larger-grained kernel fusion strategies, effectively reducing the CPU/GPU switches (Zheng et al., 2022). Optimal graph partitioning and kernel implementation selection are performed for optimal inference speed. The optimization has been applied throughout the computing module, resulting in a significant improvement in inference speed.

3.2.2 Effective IO-Aware Attention

Based on the automatic compilation optimization, we further utilize the FlashAttention technique (Dao et al., 2022) for the cross-attention operator of U-Net, which is the core of the network's inference bottleneck. The technique is based on the attention IO characteristics and performs tiling operations on the attention calculation to reduce memory read-write computation. We introduce different FlashAttention kernel implementations for

various combinations of computing devices and hardware architectures and dynamic inputs. The technique mentioned in the previous section effectively assists us in automatically finding the optimal implementation. As a result, the cross-attention calculation can be accelerated without deviation, yielding a 1.9× speed-up for the U-Net module.

4 Experiments

4.1 Experimental Settings

We first pre-train the CKCLIP model in the following experiments using the text-image pairs from Wukong (Gu et al., 2022) and the OpenKG. After that, the text encoder of CKCLIP and our diffusion model are pre-trained using the same data source. We fine-tune and evaluate the model over three domain-specific datasets to show the values of *Rapid Diffusion* in real-world applications. Implementation details and parameter settings can be found in the appendix.

4.2 Results of Three Application Scenarios

We report the performance of *Rapid Diffusion* over three domain-specific scenarios (i.e., E-commerce[5], Chinese Painting (Li et al., 2021) and Cuisine, which are closely related to our applications) in terms of Frechet Inception Distance (FID) (Heusel et al., 2017) score. Details of the three datasets, together with the training/validation/testing splits, are given in Table 4 in the appendix. We compare our model with three popular open-source diffusion models, namely Stable diffusion[6], Stable diffusion 2[7], and Taiyi Diffusion[8] (which is the largest Chinese diffusion model available so far). Note that Stable diffusion and Stable diffusion 2 mainly support English text inputs. Hence, we leverage the Chinese-English translation model (Wei et al., 2022) to translate our texts to English. The results are shown in Table 1. It can be seen that *Rapid Diffusion* outperforms all counterparties over the three datasets, achieving the average FID score at 21.90. The results indicate that our knowledge-enhanced models over domain-specific scenarios understand domain knowledge better and can generate more realistic and varied images.

[5] https://tianchi.aliyun.com/muge

[6] https://huggingface.co/CompVis/stable-diffusion-v-1-4-original

[7] https://huggingface.co/stabilityai/stable-diffusion-2

[8] https://huggingface.co/IDEA-CCNL/Taiyi-Stable-Diffusion-1B-Chinese-v0.1

Model	E-commerce	CP	Cuisine	Avg.
Stable Diffusion	48.32	70.31	26.89	48.51
Stable Diffusion 2	59.65	60.21	29.79	49.88
Taiyi Diffusion	42.43	59.56	24.08	42.02
Rapid Diffusion	**22.72**	**29.79**	**13.20**	**21.90**

Table 1: Performance of *Rapid Diffusion* and baselines over the testing sets of three application scenarios in terms of FID score. CP denotes "Chinese Painting".

4.3 Effectiveness of Knowledge-enhanced Chinese CLIP

As CLIP models aim to learn cross-modal representations, we first intrinsically evaluate our model by text-image retrieval. We compare the vanilla Chinese CLIP model and our CKCLIP model using the same pre-training text-image corpus. Pre-training details can also be found in the appendix. For evaluation, we employ the standard split of Flickr30K-CN (Lan et al., 2017), and then fine-tune both models. Table 2 reports the text-to-image and text-to-image retrieval results over the testing set. Our CKCLIP model improves retrieval performance by significant margin (especially for R@1 metric), showing its ability to learn cross-modal representations. In addition, we provide some qualitative results from the Cuisine dataset to show how more entity knowledge can lead to better representation and generation of the key objects in images, as shown in Figure 3.

Model	Text-to-image			Image-to-text		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	83.3	97.3	99.5	70.1	91.9	96.4
CKCLIP (ours)	**90.0**	**98.7**	**99.7**	**75.0**	**93.6**	**96.5**

Table 2: Performance of the knowledge-enhanced CLIP for text-image retrieval in terms of Recall@1/5/10.

Strawberry Ham Pizza | Strawberry Mixed Noodles | Strawberry Hamburger

Figure 3: Qualitative results of generated images with entity knowledge injected during CLIP pre-training. Note that the presented cuisines may not be existent in the real world. "Strawberry" is the target entity.

4.4 Results of Inference Speedup

For the implementation of compilation optimization, we employ BladeDISC (Zhu et al., 2021) as

our underlying AI compiler, which is an end-to-end dynamic shape compiler for machine learning workloads. Results of the comparison between our implementation and Torch Native (eager mode) are displayed in Table 3. According to the results, U-Net inference takes the longest in the entire process (in 1900ms), while the text encoder and the image decoder take only 0.012ms and 0.04ms, respectively. However, with the optimization by BladeDISC, we speed up the inference time of the text encoder, U-Net and the image decoder by $\times 3$, $\times 1.91$, and $\times 1.9$ times compared to the eager mode. Additionally, FlashAttention assists us in further optimizing U-Net, which decreases the inference time from 994ms to 759ms. Finally, we are able to generate images more quickly. Note that the underlying GPU used in the experiments is NVIDIA A100 (80GB), and the scheduler runs 50 in steps.

Inference Setting	CLIP (ms)	U-Net (ms)	Decoder (ms)	ESRGAN (ms)	Total (ms)
Torch Native	0.012	1900	0.04	54.5	3129
Ours (w/o. FA)	0.004	994	0.02	-	2042
Ours (w/ FA)	-	759	-	-	1807
Acceleration ratio	**×3**	**×1.91**	**×1.90**	-	**×1.73**

Table 3: Inference speedup results of *Rapid Diffusion*. FA denotes "FlashAttention".

4.5 Results of Super-resolution

For image super-resolution, the ESRGAN-based network can be efficiently leveraged to achieve upscaling results. We can directly use the ESRGAN-based network following the latent diffusion model because it has been pre-trained on common-used image datasets such as DIV2K (Agustsson and Timofte, 2017). However, considering the uniqueness and consistency of domain-specific images, we conjecture that fine-tuning enables the model to perform better. Experiments show that after fine-tuning, the model beats the pre-trained model according to our qualitative and quantitative results, which achieves 23.1 in terms of Peak Signal to Noise Ratio[9], while the pre-trained model achieves only 22.7. Figure 4 further compares images with and without our pre-trained/fine-tuned models.

4.6 Case Studies

We provide more cases from each application domain to show how much our model outperforms previous ones. Refer to Figure 5 in the appendix.

[9] https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

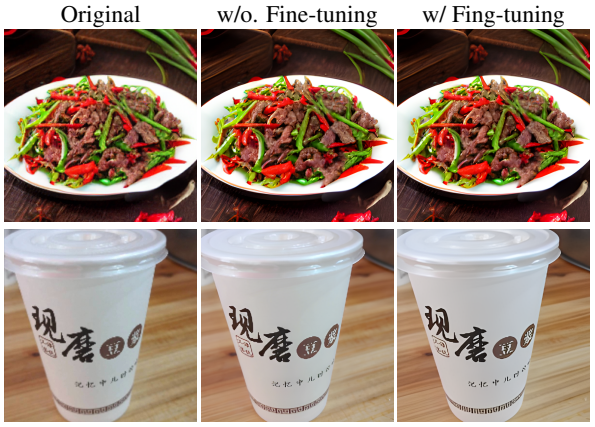


Figure 4: $256 \times 256 \rightarrow 1024 \times 1024$ super-resolution on Chinese cuisine images. The first line is the sample for the image generated from our model based on user-defined text prompts and the second line is a sample from the validation set. (Best viewed zooming-in.)

5 Applications

In this section, we demonstrate the practical values of *Rapid Diffusion* by industrial use case and the integration to Alibaba Cloud PAI (Machine Learning Platform for AI).

5.1 Industrial Use Case

Here, we briefly discuss two real-world use cases. The first is a fashion design for e-commerce manufacturers. The inputs to our system consist of keywords for multiple elements, such as trend, fabric, color and style. An automatic prompt generation process is called to provide TIS models natural-language-like inputs. For a single request of fashion design, a handful of prompts can be generated, each associated with multiple generated images. The images are then regarded as materials for designers. The cuisine dataset described previously is from our online food delivery and local life service platform. Our diffusion model for cuisine generation provides the inspiration functionality to help service providers to create innovative menus where users can select or freely enter all kinds of food-related keywords to generate images. Note that the images will be marked as "AI-generated" before they are sent to our applications.

5.2 Integration to AI Platform

To allow users to create their models, we have integrated *Rapid Diffusion* into a cloud-native AI platform to facilitate zero-code model training and elastic inference. For model training, after uploading training/validation datasets and checking hyper-

parameters, a training job is automatically submitted to our deep learning container, where the training command and the docker image have already been prepared. After the job is completed, the resulting model is available for deployment. Based on Query Per Second (QPS) requirements, our prediction service can scale to an adjustable number of machines in the cloud. We can call the TIS service via a RESTful API by HTTP requests.

6 Conclusion and Future Work

We present the *Rapid Diffusion* framework for the training and deploying knowledge-enhanced, domain-specific, high-resolution, diffusion-based TIS models. Experimental results show the effectiveness of *Rapid Diffusion* in both image quality and inference speed, achieving an average FID score of 21.90 and $\times 1.73$ acceleration ratio compared to all the counterparties. We further show its practical values through industrial use cases and the integration into an AI platform. In the future, we will extend the functionality of *Rapid Diffusion* and further increase the inference speed by advanced compilation optimization techniques.

Ethical Considerations

The techniques for inference speedup presented in this work are fully methodological. Hence, there are no direct negative social impacts. However, as the models automatically generate the images, they may have some negative impacts, such as the generation of toxic content and the existence of social biases. We suggest that the produced models should not be used to generate offensive or inappropriate images for people intentionally. Users should carefully deal with the potential risks by filtering out these images when the models are deployed online.

Acknowledgements

This work was partially supported by Alibaba Innovative Research Foundation (No. D8200510), NSFC (Grant No. 61936003), the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (No. 2017ZT07X183), Zhuhai Industry Core and Key Technology Research Project (No. 2220004002350), and by Alibaba Cloud Group through Research Talent Program with South China University of Technology.

References

Jorge Agnese, Jonathan Herrera, Haicheng Tao, and Xingquan Zhu. 2020. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(4):e1345.

Eirikur Agustsson and Radu Timofte. 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135.

Zhengda Bian, Hongxin Liu, Boxiang Wang, Haichen Huang, Yongbin Li, Chuanrui Wang, Fan Cui, and Yang You. 2021. Colossal-ai: A unified deep learning system for large-scale parallel training. *CoRR*, abs/2110.14883.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pages 2787–2795.

John C Butcher. 2000. Numerical methods for ordinary differential equations in the 20th century. *Journal of Computational and Applied Mathematics*, 125(1-2):1–29.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *CoRR*, abs/2205.14135.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Minzhe Niu, Hang Xu, Xiaodan Liang, Wei Zhang, Xin Jiang, and Chunjing Xu. 2022. Wukong: 100 million large-scale chinese cross-modal pre-training dataset and A foundation framework. *CoRR*, abs/2202.06767.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*.

Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1549–1557.

Dan Li, Shuai Wang, Jie Zou, Chang Tian, Elisha Nieuwburg, Fengyuan Sun, and Evangelos Kanoulas. 2021. Paint4poem: A dataset for artistic visualization of classical chinese poems. *arXiv preprint arXiv:2109.11682*.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. 2022a. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*.

Tingting Liu, Chengyu Wang, Xiangru Zhu, Lei Li, Minghui Qiu, Jun Huang, Ming Gao, and Yanghua Xiao. 2022b. ARTIST: A transformer-based chinese text-to-image synthesizer digesting linguistic and world knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 881–888.

Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. 2022. EI-CLIP: entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR.

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.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. 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.

Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Chengyu Wang, Minghui Qiu, Taolin Zhang, Tingting Liu, Lei Li, Jianing Wang, Ming Wang, Jun Huang, and Wei Lin. 2022. Easynlp: A comprehensive and easy-to-use toolkit for natural language processing. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations*, pages 22–29. Association for Computational Linguistics.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. ESRGAN: enhanced super-resolution generative adversarial networks. In *Computer Vision - ECCV 2018 Workshops*, pages 63–79.

Xiangpeng Wei, Heng Yu, Yue Hu, Rongxiang Weng, Weihua Luo, and Rong Jin. 2022. Learning to generalize to more: Continuous semantic augmentation for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7930–7944.

Zhen Zheng, Xuanda Yang, Pengzhan Zhao, Guoping Long, Kai Zhu, Feiwen Zhu, Wenyi Zhao, Xiaoyong Liu, Jun Yang, Jidong Zhai, et al. 2022. Astitch: enabling a new multi-dimensional optimization space for memory-intensive ml training and inference on modern simt architectures. In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 359–373.

Kai Zhu, WY Zhao, Zhen Zheng, TY Guo, PZ Zhao, JJ Bai, Jun Yang, XY Liu, LS Diao, and Wei Lin. 2021. Disc: A dynamic shape compiler for machine learning workloads. In *Proceedings of the 1st Workshop on Machine Learning and Systems*, pages 89–95.

Xiaojin Zhu, Andrew B Goldberg, Mohamed Eldawy, Charles R Dyer, and Bradley Strock. 2007. A text-to-picture synthesis system for augmenting communication. In *AAAI*, volume 7, pages 1590–1595.

A Generated Results of Case Study

We show more generated images from our model and baselines, presented in Figure 5.

B Data Statistics

Table 4 shows the data statistics of our experiments. We divide the E-commerce and Chinese Painting datasets into training, validation and testing sets according to the ratio of 80%, 10%, and 10%. For the Cuisine dataset, we divide 10% for validation, 10 thousand images for testing and the rest for training. Among these datasets, E-commerce and Chinese Painting are public datasets, while Cuisine is an in-house dataset provided by our online food delivery and local life service platform.

Domain	#Train	#Valid	#Test	Sum
E-commerce	75973	9497	9497	94967
CP	71362	8921	8921	89204
Cuisine	804305	89367	10000	903672

Table 4: The statistics of three datasets used in the experiments. CP denotes "Chinese Painting"

C Hyper-parameters Settings

For knowledge-enhanced CLIP pre-training, we follow the hyper-parameter settings in (Gu et al., 2022). For training the latent diffusion model, we set the learning rate as 5×10^{-5}, the batch size as 80, and the image size as 256×256. The latent dimension of the auto-encoder is 32×32. The hidden dimension of the text-encoder is 768.

For fine-tuning the super-resolution model, we obtain low-resolution images by down-sampling high-resolution images using the bi-cubic kernel function. Different from the original two-stage training process, we directly employ the pre-trained ESRGAN model as an initialization for the generator and the discriminator. We use Adam with $\beta_1 = 0.9, \beta_2 = 0.999$. The batch size is set to 16 and the learning rate is set to 1×10^{-4} and halved at $[50k, 100k, 200k, 300k]$ iterations.

During model training, all the experiments are conducted on a single server with 8 NVIDIA A100 GPUs (80G).

C.1 Hyper-parameters of Model Architectures

Table 5 shows the model sizes of all the experiment models, including Stable Diffusion, Stable Diffusion 2, Taiyi Diffusion and our *Rapid Diffusion* model. Compared with the other three baselines, *Rapid diffusion* is the most compact model with better performance in our scenarios.

Model	#Params
Stable Diffusion	1.37B
Stable Diffusion 2	1.29B
Taiyi Diffusion	1.35B
Rapid Diffusion	1.06B

Table 5: The numbers of parameters of all the experiment models.

We further provide the detailed settings of the entire *Rapid Diffusion* model pipelines in Table 6.

#Params	Value
CLIP Text Encoder	
context length	32
vocab size	21128
embedding dimension	768
layers	12
width	768
heads	12
Autoencoder	
z-channel	4
resolution	256
in-channels	3
out-channels	3
channels	128
channel multiplier	1,2,4,4
U-Net	
image size	32
in-channels	4
out-channels	4
model channels	320
attention resolutions	4,2,1
channel multiplier	1,2,4,4
context dimension	768
number heads	8
transformer depth	1
ESRGAN-Generator	
type	RRDBNet
in-channels	3
out-channels	3
hidden features	64
number blocks	23
grow channels	32
ESRGAN-Discriminator	
type	UNetDiscriminatorSN
in-channels	3
hidden features	64
skip connection	True

Table 6: Detailed parameter settings of *Rapid Diffusion*.

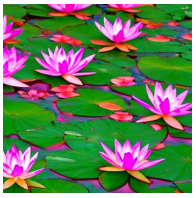
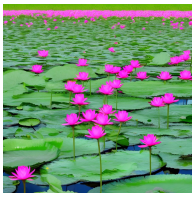
		Stable Diffusion	Stable Diffusion 2	Taiyi Diffusion	*Rapid Diffusion*
E-commerce	爆款冬季女士羽绒服 Best selling women's winter down jackets				
	18K玫瑰金女款时尚黄金项链 18K rose gold women's fashion golden necklace				
	夏季新款运动帆布鞋 New summer sports canvas shoes				
Chinese Painting	停车坐爱枫林晚，霜叶红于二月花 Stop the coach to enjoy the maple woods; frosty leaves are redder than the February flowers. (ancient Chinese poem)				
	千山鸟飞绝，万径人踪灭 From hill to hill no bird in flight, from path to path no man in sight. (ancient Chinese poem)				
	接天莲叶无穷碧，映日荷花别样红 Green lotus leaves outspread as far as boundless sky, pink lotus blossoms take from sunshine a new dye. (ancient Chinese poem)				
Cuisine	小炒黄牛肉 Stir-fried yellow beef				
	鱼香肉丝米饭 Yuxiang shredded pork and rice				
	大杯烧仙草奶茶 Big cup of milk tea with grass jelly				

Figure 5: Some examples of generated images from *Rapid Diffusion* and baseline models.