ARTIST: A Transformer-based Chinese Text-to-Image Synthesizer Digesting Linguistic and World Knowledge

Tingting Liu[1,2]*, **Chengyu Wang**[2]*, **Xiangru Zhu**[2,3], **Lei Li**[1,2], **Minghui Qiu**[2], **Jun Huang**[2], **Ming Gao**[1,4]†, **Yanghua Xiao**[3]

[1] School of Data Science and Engineering, East China Normal University, Shanghai, China
[2] Alibaba Group, Hangzhou, China
[3] School of Computer Science, Fudan University, Shanghai, China
[4] KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, China
ttliu@stu.ecnu.edu.cn, chengyu.wcy@alibaba-inc.com, xrzhu19@fudan.edu.cn
leili@stu.ecnu.edu.cn, {minghui.qmh,huangjun.hj}@alibaba-inc.com
mgao@dase.ecnu.edu.cn, shawyh@fudan.edu.cn

Abstract

Text-to-Image Synthesis (TIS) is a popular task to convert natural language texts into realistic images. Recently, transformer-based TIS models (such as DALL-E) have been proposed using the encoder-decoder architectures. Yet, these billion-scale TIS models are difficult to tune and deploy in resource-constrained environments. In addition, there is a lack of language-specific TIS benchmarks for Chinese, together with high-performing models with moderate sizes. In this work, we present ARTIST, A tRansformer-based Chinese Text-to-Image SynThesizer for high-quality image generation. In ARTIST, the rich linguistic and relational knowledge facts are injected into the model to ensure better model performance without the usage of ultra-large models. We further establish a large-scale Chinese TIS benchmark with the re-production results of state-of-the-art transformer-based TIS models. Results show ARTIST outperforms previous approaches. [1]

1 Introduction

Text-to-Image Synthesis (TIS) is a popular multi-modality task that aims to convert natural language texts into realistic images (Frolov et al., 2021). For accurate TIS, various methods have been proposed based on Generative Adversarial Networks (GANs) (Xu et al., 2018; Zhang et al., 2021a).

Recently, with the wide popularity of the transformer model architecture (Vaswani et al., 2017), transformer-based TIS has received more attention, which converts raw texts into "pseudo image tokens" by transformers and generates images based on models such as VQGAN (Esser et al., 2021) and VQ-VAE (van den Oord et al., 2017). Notable works include DALL-E (Ramesh et al., 2021), DALL-E 2 (Ramesh et al., 2022), CogView (Ding et al., 2021), CogView2 (Ding et al., 2022), ERNIE-ViLG (Zhang et al., 2021b), M6 (Lin et al., 2021), OFA (Wang et al., 2022b) and a few others.

Despite the remarkable progress, we suggest that the mainstream transformer-based TIS models may have a few drawbacks. i) Most TIS models have billion-scale parameters, making it challenging to fine-tune and deploy them in resource-constrained environments (such as Ramesh et al. (2021); Ding et al. (2021)). This highly limits the applications of these models in real-world applications. ii) The encoder-decoder architecture in the transformer (Vaswani et al., 2017) does not explicitly model the semantics of key elements appearing in texts (i.e., entities or objects), and hence may lack the background knowledge for realistic image generation. iii) Most existing works are benchmarked with English datasets, e.g., MS-COCO (Lin et al., 2014). There is a lack of language-specific benchmarks for other languages (Chinese in our work), together with high-performing TIS models that are suitable for efficient domain-specific fine-tuning and online deployment. In addition, the multi-granularity of word segmentation of the Chinese language (Lai et al., 2021) makes the underlying transformer difficult to understand the true meanings of the input texts, which also causes ambiguity for knowledge injection to the models.

We present ARTIST, A tRansformer-based Chinese Text-to-Image SynThesizer with rich linguistic and world knowledge digested. It aims to gen-

* T. Liu and C. Wang contributed equally to this work.
† Corresponding author.
[1] All the benchmark resources and the ARTIST checkpoints will be released to the EasyNLP framework (Wang et al., 2022a). URL: https://github.com/alibaba/EasyNLP

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 881 - 888

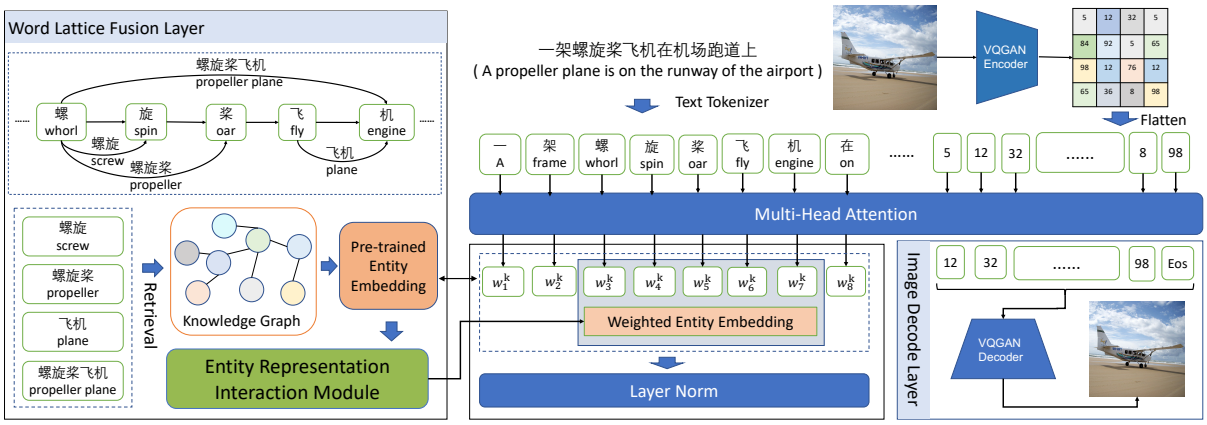

Figure 1: The overall framework of our ARTIST framework.

erate high-quality images with a moderate parameter size. To enhance the understanding abilities of ARTIST, we model the multi-granularity of input texts by linguistics knowledge and attentively inject entity embeddings into the encoder of the transformer model, which is derived from the massive relational facts in knowledge bases. The resulting images are further generated by VQGAN based on "pseudo images tokens" produced by the transformer decoder (Esser et al., 2021).

For evaluation, we establish a large-scale Chinese TIS benchmark over multiple public multimodal datasets and re-produce the results of several popular transformer-based TIS models. The experimental results show that our ARTIST model outperforms previous approaches. In summary, we make the following major contributions in this work:

- We formally propose the ARTIST framework for knowledge-enhanced Chinese TIS.
- In ARTIST, the rich linguistic and relational knowledge facts are injected into the model for better Chinese language understanding.
- We establish a large-scale Chinese TIS benchmark with the re-production results of state-of-the-art transformer-based TIS models. The experimental results also show that ARTIST outperforms previous approaches.

2 The ARTIST Model

2.1 Overview

Figure 1 shows the overview of our ARTIST framework. In the word lattice fusion layer, as these exists multi-granularity of the Chinese language, thus, for an input text, we obtain all possible word segmentation results and generate the word lattice of the corresponding text. The pre-trained entity embeddings are learned from a large-scale knowledge graph and are selectively injected into entity representations by our designed Entity Representation Interaction Module (ERIM). With the fused knowledge, the transformer model auto-regressively generates "pseudo image tokens", where the codebook is obtained from a VQGAN model. Finally, the images are decoded using the same VQGAN model.

2.2 Word Lattice Generation

Previous works for TIS treat tokens in input texts equally, while we suggest that entities described in texts are the critical guide to generating the images that are strongly related to the specific objects. Hence, it is vital to identify the entities and integrate token embeddings with the pre-trained entity embeddings during transformer training. For Chinese, different word segmentations have a great impact on meanings of sentences, which leads to error propagation and language ambiguity. [2] Following Li et al. (2020), we obtain word lattices of the input sentences, containing all possible word segmentations and entities of the texts.

2.3 Entity Representation Interaction

The lattice structure represents all possible entities in the sentence, yet too much knowledge injection may lead to sentence meaning confusion (also

[2]For example, in Figure 1, we find that if we direct match the names of texts with entities in the knowledge graph, "propeller" and "plane" can be detected. The generated images of our model do not necessarily guarantee the existence of the object "propeller plane". In contrast, our method allows the model to "see" all possible entities expressed in texts and learns to "decide" to plot certain objects in the image.

Paradigm	Model	Size	COCO-CN		MUGE		Flickr8k-CN		Flickr30k-CN	
			FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
Zero-shot	CogView	4B	102.30	11.81±0.84	29.08	10.71±0.40	102.01	11.58±0.66	103.34	10.50±0.35
	DALL-E	209M	89.73	10.32±0.64	40.28	9.90±0.48	77.84	10.57±0.37	77.08	10.03±0.60
Fine-tuning	DALL-E	209M	84.73	11.08±0.89	22.42	10.28±0.44	72.17	9.89±0.41	68.75	9.86±0.40
	OFA	180M	82.14	12.53±1.10	12.60	13.08±0.50	58.06	13.10±0.51	54.23	**13.25±0.55**
	ARTIST	202M	**68.75**	**13.92±1.18**	**11.84**	**13.42±0.11**	**50.08**	**13.54±0.75**	**52.45**	11.12±0.15

Table 1: The overall experimental results of baseline methods and ARTIST over four benchmark datasets. The best results are printed in bold, with the second best underlined.

called knowledge noise (Liu et al., 2020)). To avoid interplay among representations of multiple entities in the same position, we design the Entity Representation Interaction Module (ERIM) to selectively fuse the knowledge from the lattice into the transformer model. Denote $\mathbf{h}^{(k)} = \{\mathbf{h}_1^{(k)}, \cdots, \mathbf{h}_N^{(k)}\} \in \mathbb{R}^{N \times d}$ as the token embeddings of layer k where $\mathbf{h}_i^{(k)}$ denotes the i-th token embedding, N is the the sequence length, and d is the dimension of hidden representation. Let M be the collection of all possible entities of a given sentence appearing in the lattice. We further denote $\mathbf{e}_m$ as the pre-trained entity embedding of the m-th entity in M, and $\mathbf{e}_{i,m}$ as the entity embedding to be injected into the i-th token based on the knowledge of $\mathbf{e}_m$. Clearly if the i-th token overlaps with the m-th entity, we have $\mathbf{e}_{i,m} = \mathbf{e}_m$ and $\mathbf{e}_{i,m} = \mathbf{0}$ otherwise. In our work, we obtain the pre-trained entity embeddings by TransE (Bordes et al., 2013) from a large-scale Chinese knowledge graph CN-DBPedia that contains more than 9 million entities and 67 million triples of relationships (Xu et al., 2017).[3] The mutual knowledge injection process is then computed as follows:

$$w_{i,m}^{(k)} = [\mathbf{h}_i^{(k)}]^T \cdot \mathbf{e}_{i,m} \tag{1}$$

$$\widetilde{\mathbf{h}}_i^{(k)} = \mathbf{h}_i^{(k)} + \sum_{m=1}^{M} w_{i,m}^{(k)} \cdot \mathbf{e}_{i,m} \tag{2}$$

where $w_{i,m}^{(k)}$ is the weight of the m-th entity embedding for $\mathbf{h}_i^{(k)}$, and $\widetilde{\mathbf{h}}_i^{(k)}$ represents the knowledge-enhanced hidden token embedding, after the knowledge of multiple entity embeddings is selectively injected. The entire sequence embeddings are further denoted as $\widetilde{\mathbf{h}}^{(k)}$. We build up the transformer layers by:

$$\mathbf{g}^{(k)} = \widetilde{\mathbf{h}}^{(k)} + ATTN(LN(\widetilde{\mathbf{h}}^{(k)})) \tag{3}$$

$$\mathbf{h}^{(k+1)} = \mathbf{g}^{(k)} + \mathbf{W}_2 \cdot \sigma(\mathbf{W}_1 LN(\mathbf{g}^{(k)})) \tag{4}$$

where $ATTN$ and LN denote attention and layer norm, with $\mathbf{W}_1, \mathbf{W}_2$ to be learnable parameters.

2.4 Realistic Image Generation

For image generation, we employ the above auto-regressive transformer that allows ARTIST to generate a sequence of "pseudo image tokens" based on the knowledge-enhanced text embeddings. Specifically, the "pseudo image tokens" are the codebook indices encoded by the pre-trained VQGAN model (Esser et al., 2021), denoted as $\mathbf{v} = \{v_1, v_2, \cdots, v_G\}$, where G is the sequence length of image tokens. Given the text tokens $\mathbf{w}$ and the image tokens $\mathbf{v}$, we model $p(v)$ as:

$$p(v) = \prod_{i=1}^{G} p_\Theta(v_i | v_1, \cdots, v_{i-1}, \mathbf{w}) \tag{5}$$

The loss function of the ARTIST model is:

$$L = \mathbb{E}[-\log p(v)] \tag{6}$$

where Θ is the collection of model parameters. Finally, the images are decoded from "pseudo image tokens" to image pixels. The parameters of VQGAN are fixed during model training.

3 Benchmark and Experimental Results

3.1 Benchmark

To our knowledge, there are no Chinese TIS benchmarks publicly available for us. Thus, we seek to construct a benchmark for the research community. The evaluation datasets include COCO-CN (Li et al., 2019), MUGE[4], Flickr8k-CN (Li et al., 2016) and Flickr30k-CN (Lan et al., 2017), containing a large number of high-quality Chinese text-image pairs. The detailed statistics of the data splits can be found in the appendix (Table 4). Following previous works on TIS, we employ Fréchet Inception Distance (FID) and Inception Score (IS) as metrics (Zhu et al., 2019). A higher IS and a

[3]To trade-off the performance and efficiency, we use the TransE model to learn entity representations.

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

Model	COCO-CN		MUGE		Flickr8k-CN		Flickr30k-CN	
	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
Full Implement.	68.75	13.92±1.18	11.84	13.42±0.11	50.08	13.54±0.75	52.45	11.12±0.15
w/o. all knowledge	76.89	11.65±0.89	13.31	11.91±0.36	55.56	12.54±0.48	55.66	10.19±0.30
w/o. word lattice	74.16	12.47±0.64	13.15	11.86±0.36	54.32	12.15±0.63	55.41	10.21±0.42

Table 2: Results of knowledge ablation. "w/o. all knowledge" means no knowledge is injected, and "w/o. word lattice" means directly injecting entity embeddings at the corresponding locations without word lattice and ERIM.

lower FID indicate that the qualities of generated images are better.

For baselines, CogView (Ding et al., 2021)[5] releases a Chinese model checkpoint that supports zero-shot learning. We also compare ARTIST against other methods with public codes and model checkpoints, including DALL-E (Ramesh et al., 2021)[6] and OFA (Wang et al., 2022b)[7]. As the models of DALL-E and OFA are for English only, while the English translations of captions from COCO-CN, Flickr-8k and Flickr-30k are already available, those of MUGE are translated into English through a commercial translation service. There exist some other recent works; however, their codes and checkpoints are not available at the time of writing.

3.2 Model Configurations of ARTIST

We have pre-trained and released two versions of ARTIST (base and large), with 202M and 433M parameters, respectively, with details further shown in the appendix (Table 5). Both models are pre-trained over a subset of the Wukong corpus (Gu et al., 2022), which contains 100M Chinese pre-training text-image pairs collected from the Web. After that, the models are fine-tuned over the four datasets. During training, we fix the batch size and the learning rate to be 16 and $4.5e-6$, respectively. The sequence length is 288, 32 for text and 256 for image. The vocabulary size is 37,512, containing 21,128 text tokens and 16,384 image tokens. Other hyper-parameters are tuned on development sets. To save computational resources, we also pre-train a Chinese CLIP model (Radford et al., 2021) over the same Wukong corpus to rank the 10 generated images in order to select the best one. We implement ARTIST in PyTorch and conduct experiments on a server with 8 Tesla V100 GPUs (32GB).

[5]https://github.com/THUDM/CogView

[6]DALL-E models are not available in their official repository. We seek to reproduce the zero-shot and fine-tuning results based on https://github.com/lucidrains/DALLE-pytorch.

[7]OFA supports fine-tuning only, with no zero-shot learning functionalities provided. See https://github.com/OFA-Sys/OFA.

Model	COCO-CN		Flickr8k-CN	
	FID↓	IS↑	FID↓	IS↑
OFA	70.82	14.60±1.03	58.56	13.03±0.58
ARTIST	**66.66**	**14.71±1.13**	**49.42**	**15.01±0.64**

Table 3: The performance comparison of larger models (ARTIST-large and OFA-large) over two datasets.

3.3 Overall Performance

Table 1 summarizes the TIS results on all benchmark datasets. As seen, ARTIST shows superiority over both zero-shot generation and fine-tuned methods. Overall, the qualities of images generated by zero-shot learning are not as good as fine-tuned models, regardless of the model scale. Compared with DALL-E and OFA's fine-tuned models, ARTIST achieves new state-of-the-arts over FID on all four datasets at comparable model sizes and has competitive results for IS on datasets other than Flickr30k-CN. In summary, the qualities of images generated by ARTIST are of great advantage by effective knowledge injection.

3.4 Detailed Analysis

Knowledge Ablation. We further conduct an ablation study to verify the impact of knowledge injection. In Table 2, the ablation of either all knowledge or word lattice leads to a substantial drop in performance. Injecting knowledge based on word lattice and ERIM has a more significant improvement. In average, directly injecting entity embeddings reduces FID by 1.1 (from 50.36 to 49.26) and increases IS by 0.1 (from 11.57 to 11.67) compared to no knowledge injection, while injecting knowledge based on our approach reduces FID by 4.58 (from 50.36 to 45.78) and improves IS by 1.43 (from 11.57 to 13.00).

Learning with Larger Models. We also increase the model size of ARITIST to 433M, and compare it against OFA-large (470M). The ARTIST-large model further improves the performance compared to the base model. As shown in Table 3, in average, ARTIST-large reduces 6.65 in FID and improves 1.05 in IS compared to OFA-large.

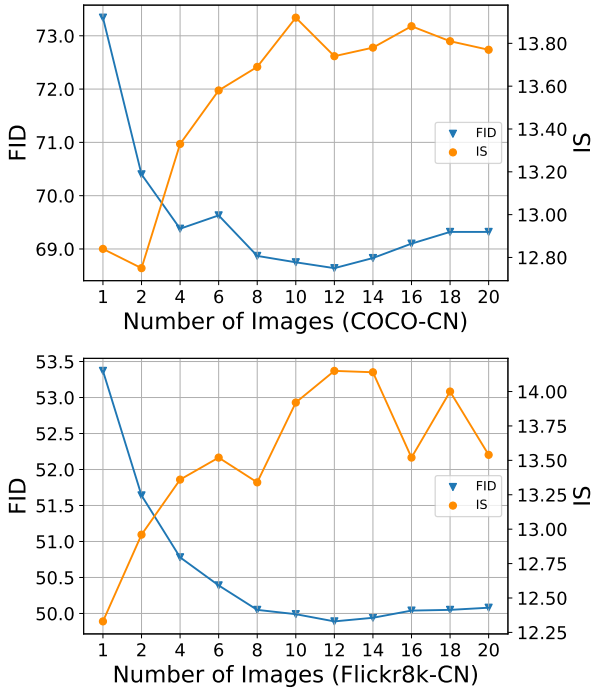

Figure 2: The performance trend of ARTIST when the number of generated images varies.

CLIP. We generate multiple images for each query text and select the best one by the pre-trained CLIP model. As shown in Figure 2, the number of generated images has an impact on the performance. We recommend generating 10 images with ARTIST to balance performance and efficiency. In the literature, DALL-E generates 512 images for each caption and selects the best one, and CogView generates 60. Our work is much more efficient with better performance.

4 Conclusion and Future Work

In this paper, we present the ARTIST framework for knowledge-enhanced Chinese TIS. The rich linguistic and relational knowledge facts are injected into the model for better Chinese language understanding. For evaluation, we establish a large-scale Chinese TIS benchmark and show that the proposed ARTIST models outperform previous approaches. We will release our models and benchmark to the public and extend our work to other languages in the future.

Limitations

Our work focuses on transformer-based TIS models for the Chinese language, where the rich relational knowledge facts and the linguistic characteristics are fused into the models for better performance. It is natural to extend our work to other languages (such as English) by considering the linguistic characteristics of these languages as well, which will be addressed in the future work.

Ethical Considerations

Our contribution in this work is fully methodological, namely a new framework to train Chinese TIS models with rich knowledge injected. Hence, there are no direct negative social impacts of this contribution. However, as transformer-based models may have some negative impacts, such as the generation of toxic contents by machines, the produced TIS models produced by our algorithms would unavoidably suffer from these issues, which can have the possibilities of generating inappropriate images. We suggest that users should carefully deal with the potential risks by filtering out these images when the TIS models are deployed online.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China under Grant No. U1911203, Alibaba Group through the Alibaba Innovation Research Program, the National Natural Science Foundation of China under Grant No. 61877018, the Research Project of Shanghai Science and Technology Commission (20dz2260300) and the Fundamental Research Funds for the Central Universities.

References

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. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. 2021. Cogview: Mastering text-to-image generation via transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pages 19822–19835.

Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *CoRR*, abs/2204.14217.

Patrick Esser, Robin Rombach, and Björn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12873–12883. Computer Vision Foundation / IEEE.

Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Adversarial text-to-image synthesis: A review. *Neural Networks*, 144:187–209.

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.

Yuxuan Lai, Yijia Liu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2021. Lattice-bert: Leveraging multi-granularity representations in chinese pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1716–1731. Association for Computational Linguistics.

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

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6836–6842. Association for Computational Linguistics.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275. ACM.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. COCO-CN for cross-lingual image tagging, captioning, and retrieval. *IEEE Trans. Multim.*, 21(9):2347–2360.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: A chinese multimodal pretrainer. *CoRR*, abs/2103.00823.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014*, volume 8693, pages 740–755. Springer.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 6306–6315.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008.

Chengyu Wang, Minghui Qiu, Taolin Zhang, Tingting Liu, Lei Li, Jianing Wang, Ming Wang, Jun Huang, and Wei Lin. 2022a. Easynlp: A comprehensive and easy-to-use toolkit for natural language processing. *CoRR*, abs/2205.00258.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *Advances in Artificial Intelligence:*

From Theory to Practice - 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, volume 10351 of *Lecture Notes in Computer Science*, pages 428–438. Springer.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324. Computer Vision Foundation / IEEE Computer Society.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021a. Cross-modal contrastive learning for text-to-image generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 833–842. Computer Vision Foundation / IEEE.

Han Zhang, Weichong Yin, Yewei Fang, Lanxin Li, Boqiang Duan, Zhihua Wu, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021b. Ernie-vilg: Unified generative pre-training for bidirectional vision-language generation. *CoRR*, abs/2112.15283.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5802–5810. Computer Vision Foundation / IEEE.

A Dataset Statistics

The statistics of the four benchmark datasets are summarized in Table 4.

Dataset	# Images / # Texts		
	#Training	#Validation	#Testing
COCO-CN	18341/20065	1000/1100	1000/1053
MUGE	89970/89970	4997/4997	4999/4999
Flickr8k-CN	6000/30000	1000/5000	1000/5000
Flickr30k-CN	29783/148915	1000/5000	1000/5000

Table 4: Dataset statistics.

B Configurations of ARTIST Models

The detailed configurations of the ARTIST-base and ARTIST-large models are summarized in Table 5.

Model	ARTIST-base	ARTIST-large
Layers	12	24
Attention Heads	12	16
Hidden Size	768	1024
Text Length	32	32
Image Length	16 ×16	16 × 16
Image Size	256 × 256	256 × 256
Codebook Size	16384	16384

Table 5: Detailed model configurations.

Dataset	VQGAN		ARTIST-base	
	FID↓	IS↑	FID↓	IS↑
COCO-CN	40.46	17.56±1.69	68.75	13.92±1.18
MUGE	5.67	13.54±0.48	11.84	13.42±0.11
Flickr8k-CN	40.82	13.71±1.55	50.08	13.54±0.75
Flickr30k-CN	42.08	16.34±0.53	52.45	11.12±0.15

Table 6: The reconstruction results of VQGAN model, together with the performance of ARTIST-base.

C The Performance of VQGAN

During the training process of ARTIST, we fix the parameters of the VQGAN model. Hence, the qualities of the generated images are constrained by the VQGAN model, which can be viewed as the upper bound. We compare the results of VQGAN reconstruction and ARTIST-base in Table 6. As seen, our proposed ARTIST model is closer to the VQGAN reconstruction results than other baselines in most cases, which shows the superiority of our method.

D Case Studies

Figure 3 and Figure 4 show some qualitative results of ARTIST and other open-sourced models in e-commerce and natural scenes, respectively. In general, our approach generates images with more vivid details in most cases.

Figure 3: Qualitative comparison of images generated from the MUGE dataset (e-commerce products).

Figure 4: Qualitative comparison of images generated from the COCO-CN and Flickr8k-CN datasets (natural scene).