AltDiffusion: A Multilingual Text-to-Image Diffusion Model

Fulong Ye^{1,2*†}, Guang liu^{2*‡}, Xinya Wu^{2}, Ledell Wu^{2}

[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] Beijing Academy of Artificial Intelligence
fulong_ye@bupt.edu.cn
{liuguang, yxwu, wuyu}@baai.ac.cn

Abstract

Large Text-to-Image(T2I) diffusion models have shown a remarkable capability to produce photorealistic and diverse images based on text inputs. However, existing works only support limited language input, e.g., English, Chinese, and Japanese, leaving users beyond these languages underserved and blocking the global expansion of T2I models. Therefore, this paper presents AltDiffusion, a novel multilingual T2I diffusion model that supports eighteen different languages. Specifically, we first train a multilingual text encoder based on the knowledge distillation. Then we plug it into a pretrained English-only diffusion model and train the model with a two-stage schema to enhance the multilingual capability, including concept alignment and quality improvement stage on a large-scale multilingual dataset. Furthermore, we introduce a new benchmark, which includes Multilingual-General-18(MG-18) and Multilingual-Cultural-18(MC-18) datasets, to evaluate the capabilities of T2I diffusion models for generating high-quality images and capturing culture-specific concepts in different languages. Experimental results on both MG-18 and MC-18 demonstrate that AltDiffusion outperforms current state-of-the-art T2I models, e.g., Stable Diffusion in multilingual understanding, especially with respect to culture-specific concepts, while still having comparable capability for generating high-quality images. All source code and checkpoints could be found in https://github.com/superhero-7/AltDiffuson.

Introduction

In recent years, there has been an emerging interest in large Text-to-Image(T2I) diffusion models, such as Stable Diffusion(SD)(Rombach et al. 2022), Imagen(Saharia et al. 2022) and DALLE2(Ramesh et al. 2022), due to their remarkable ability to produce photorealistic and diverse images based on text input. A limitation of these large T2I diffusion models is that they only support prompts in English, which is inconvenient for non-English users, e.g., Spanish or French. Non-English users usually utilize T2I diffusion models with the help of translation tools, which may lead to translation error and information loss, especially in some culture-specific concepts. For example, the name of the famous Chinese painter Baishi Qi may be translated as ”white stone” in English. And the translation process is getting more complex when it comes to mixed language prompts. Intuitively, a T2I generative model uses native languages directly without additional translation steps have no such problems. Recently, some scholars have begun to develop multilingual T2I diffusion models. Taiyi-Bilingual(Wang et al. 2022b), ERNIE-ViLG 2.0(Feng et al. 2023) are T2I bilingual models that support both Chinese and English. However, these T2I diffusion models are still limited by the scarcity of language varieties.

To address the problem, we propose a novel multilingual T2I diffusion model, which is capable of processing eighteen languages[1] that cover 46.94% of the world's first-language speakers and 27.64% of the world's second-language speakers, named AltDiffusion(AD), along with an efficient training approach. We first train a multilingual text encoder based on the knowledge distillation(Chen et al. 2022) to enhance the language capability to support eighteen languages. Then, the parameters of the text encoder are frozen and plugged into a pretrained English-only diffusion model. Next, we propose a two-stage training schema to enhance the language capability of the diffusion model. In the first stage, we train the K, V matrix in cross-attention of the UNet on a multilingual dataset LAION 5B(Schuhmann et al. 2022) to align the embedding space between the UNet and the text encoder. In the second stage, all parameters of the UNet are unfrozen to improve the quality of the generated images using LAION Aesthetics(Shing and Sawada 2022a). In addition, a classifier-free guidance training technique is employed to further improve the generation quality. It is worth noting that our training approach possesses strong generality and can support any pretrained T2I diffusion models.

To evaluate our AD model, we introduce a benchmark including two evaluation datasets that focus on two aspects of multilingual generation, respectively. For general

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[*] Equal contribution.

[†] Work done during internship with Beijing Academy of Artificial Intelligence.

[‡] Corresponding Author.

[1] Eighteen languages: English, Chinese, Japanese, Thai, Korean, Hindi, Ukrainian, Arabic, Turkish, Vietnamese, Polish, Dutch, Portuguese, Italian, Spanish, German, French, and Russian.

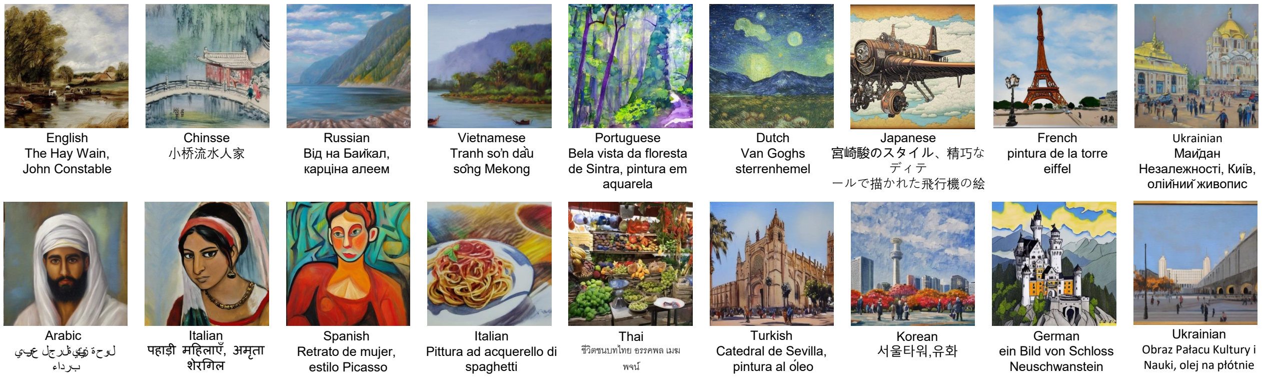


Figure 1: Images generated by AltDiffusion with prompts in various languages. We select prompts with culture-specific concepts in different languages to demenstrate the strong capability of multilingual T2I generation of AltDiffusion.

quality evaluation, we expand the data of XM-3600 by filtering high-quality image-text pairs from WIT and construct a high-quality dataset Multilingual-General-18(**MG-18**) that includes 7,000 images per language to evaluate FID(Heusel et al. 2017), IS(Salimans et al. 2016), and CLIP Sim(Hessel et al. 2021). For culture-specific concepts evaluation, we introduce Multilinguale-Cultural-18(**MC-18**), a culture-specific dataset with 50 text prompts per language about culture-specific concepts of different countries. The MC-18 is the first dataset about culture-specific concepts. This benchmark provides robust and comprehensive evaluation for multilingual T2I generation.

Our experimental results on MG-18 demonstrate that AD is the first multilingual T2I diffusion model that supports eighteen languages and outperforms other multilingual diffusion models, e.g. Taiyi(Wang et al. 2022a) and Japanese SD models(Shing and Sawada 2022b), in FID, IS, and CLIP Sim. In addition, AD surpasses translation-based SD in CLIP Sim of all languages and achieves comparable results in FID and IS, proving that AD is better than SD in multilingual understanding and can generate almost the same high-quality images as SD on general prompts. Experimental results on MC-18 show that AD beats translation-based SD in the culture-specific concepts in all languages.

Our contributions are as follows:

- We introduce AltDiffusion, a novel multilingual diffusion model that supports eighteen languages, which covers 46.94% of the world's first-language speakers and 27.64% of the world's second-language speakers.
- We introduce a benchmark that includes two datasets for evaluating T2I generative model: a general quality evaluation dataset MG-18 and a culture-specific dataset MC-18. This benchmark provides robust and comprehensive evaluation for multilingual T2I generation.
- AltDiffusion outperforms other multilingual diffusion models and performs better than a translation-based Stable Diffusion in multilingual understanding capability, especially in culture-specific concepts.

Related Work

Multilingual Text-to-image Generation Recently, T2I diffusion models ((Rombach et al. 2022), (Saharia et al. 2022), (Ramesh et al. 2022), (Zhang et al. 2021), (Feng et al. 2022), (Ding et al. 2021), (Ding et al. 2022)) achieve remarkable success in generating photorealistic and diverse images. Stable Diffusion(SD) is a prominent open-source framework with a considerable community. SD model consists of three parts: autoencoder is responsible for encoding images and decoding the pictures into and from the latent space; text encoder is accountable for encoding text prompts; Unet(Ronneberger, Fischer, and Brox 2015) is responsible for predicting noise based on the language embedding in latent space. Despite its strong generative capability, the fact that the SD model can only support English input still leads to limitations. Some works, such as CogView((Ding et al. 2021), (Ding et al. 2022)) and ERNIE-ViLG((Zhang et al. 2021), (Feng et al. 2022)), start to explore the T2I diffusion model that can support multilingual text prompt. Only some studies try to extend the applicability of the SD beyond English to other languages, e.g., Taiyi(Wang et al. 2022a) and Japanese SD (Shing and Sawada 2022b). However, these T2I generative models are still limited by the scarcity of language varieties. In this work, We are committed to constructing a multilingual T2I diffusion model which can serve most of the world's population.

Multilingual CLIP CLIP shows a solid capability to provide a robust text-image representation in English. Recently, several works have tried to expand the language capabilities of CLIP to other languages. For instance, previous studies (Aggarwal and Kale 2020), (Carlsson et al. 2022) attempt to create multilingual CLIP models. AltCLIP (Chen et al. 2022) applies knowledge distillation techniques to develop a state-of-the-art multilingual CLIP model by leveraging the XLM-R multilingual model (Conneau et al. 2019). Following AltCLIP, we retrain the text encoder to align the penultimate hidden layer of OpenCLIP(Cherti et al. 2022) with the text encoder used in SD v2 to create a multilingual CLIP model.

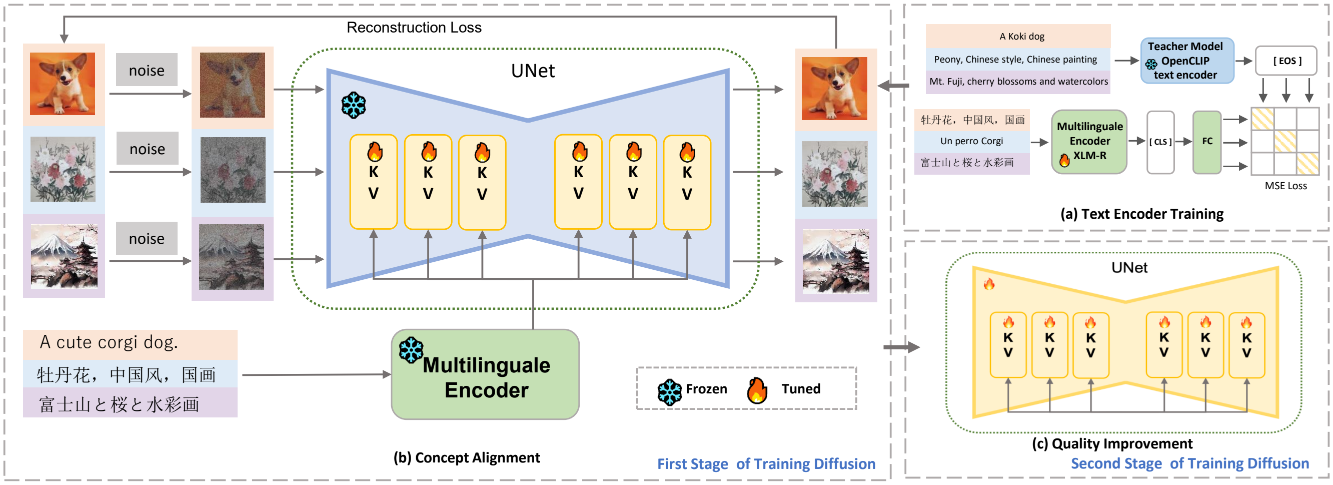


Figure 2: Illustration of the training approach. First, we train a multilingual text encoder. Then in the concept alignment stage, we only unfreeze the k and v parameters in cross-attention. In the quality improvement stage, all parameters of the UNet are unfrozen. Both stages are trained in 18 languages(Here only illustrate English, Chinese, and Japanese).

Multilingual Image Caption Datasets Image caption datasets are widely used for multimodal tasks, which are mainly accessible in English, e.g., Flickr30k(Plummer et al. 2017), MS COCO(Lin et al. 2014). These monolingual datasets are limited by language-linguistic diversity. Thus some works have focused on image caption datasets in different languages. Multi30K(Elliott et al. 2016) is a dataset that supports German, while Wikimedia Commons(Schamoni, Hitschler, and Riezler 2018) can support German, French, and Russian. Recently, datasets that support multiple languages such as WIT(Srinivasan et al. 2021), XM 3600(Thapliyal et al. 2022) are proposed. WIT is collected by gathering diverse textual content linked to an image from Wikipedia, which has 37.6 million image-text pairs in 100+ languages. XM 3600 is a manually annotated dataset with 3,600 image-text in 36 languages. Based on XM3600 and WIT datasets, we build a dataset MG-18 to evaluate AD.

Method

The current Large T2I diffusion models usually consist of a text encoder and a UNet model. To train a multilingual T2I diffusion model, we first enhance the language capability of the text encoder and then align it with the UNet to enhance the language capability of UNet.

Enhance Language Capability of the Text Encoder

Following AltCLIP(Chen et al. 2022), we retrain the text encoder to support 18 languages based on the knowledge distillation. As shown in Figure 2(a), the text encoder from OpenCLIP(Cherti et al. 2022) is the teacher model, and XLM-R(Conneau et al. 2019) is the student model. Given parallel prompts $(text_{english}, text_{otherlanguage})$, the $text_{english}$ input is fed into the teacher model, and the $text_{otherlanguage}$ input is fed into the student model. We minimize the Mean Squared Error(MSE) between $[TOS]$ embedding of the teacher model and $[CLS]$ embedding of the student model. A fully connected network maps the outputs of XLM-R and the OpenCLIP text encoder penultimate layer to the same dimensionality. After training, we obtain a multilingual text encoder whose embedding space is close to the original OpenCLIP text encoder.

Enhance Language Capability of the UNet

After training the text encoder, the parameters of the text encoder are frozen and plugged into an off-the-shelf pretrained English-only diffusion model(here, we use SD, but our method can be extended to other diffusion models with a text encoder). Then we use two-stage training schema, including concept alignment and quality improvement stage, to transform the English-only diffusion model into a multilingual one.

Concept Alignment Stage This stage aims to re-establish the relationship between the text and the images by aligning the embedding space between the text encoder and UNet. The training dataset is LAION(Schuhmann et al. 2022), which is a large-scale dataset with a multilingual corpus(detailed introduction is in Dataset section). Preliminary analysis(Gandikota et al. 2023) reveals that the cross-attention mechanism of the diffusion model plays a crucial role in matching the text and images. Therefore, as illustrated in Figure 2(b), we freeze the multilingual text encoder, the autoencoder, and most of the parameters of the UNet, then train the K, V matrix of the cross-attention module using the denoising diffusion objective(Ho, Jain, and Abbeel 2020):

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x),t,c,\epsilon\sim N(0,1)}\left[\left\|\epsilon - \epsilon_O\left(z_t, c, t\right)\right\|_2^2\right] \qquad (1)$$

where $t \sim$ Uniform $[1, T]$, z_t is a noisy version of latent embedding z of input image x (i.e. $z = \mathcal{E}(x)$), obtained using $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. θ is the parameter of UNet to predict the noisy $\epsilon_\theta\left(z_t, c, t\right)$ condition on z_t, multilingual text condition embedding c and t.

	Painting	Literature	Festival	Clothe
ch	A landscape painting by BaishiQi	Jade-like adornment forms a tall tree, with ten thousand branches hanging down like green silk threads.	On the eve of ChineseNew Year, every household hangs bright red lanterns.	The girl wearing a qipao performed on stage
ja	Miyazaki Hayao's style, the fantastic scene of the night where the spirits shining in the woods dance	In the cauliflower field, the moon faces west and the sun faces east	At the night of Japan's Star Festival, the city lights up and decorates special decorations.	The embroidery of the kimono is delicate embroidery.
es	The Mother of Picasso": A touching portrait of his mother."	In a deep canyon, the rocky walls rise impressively, creating a spectacle of natural shapes and colors.	Carnival in Spain is an explosion of joy and color that fills the streets with fun and festivity.	The "trajede baturro" from Aragón consists of a vest, sash, and trousers.

Table 1: Data samples of MC-18 in Chinese(ch), Japanese(ja) and Spanish(es). For the convenience of reading, the examples shown here are translated into English. For more and more complete examples, please visit our github homepage.

Previous observation(Rombach et al. 2022) indicates that reducing the dimensions of images from 512×512 to 256×256 results in minimal damage to semantic information, only eliminating some imperceptible details. In line with our objective of aligning the semantic information between modalities, we utilize the lower image resolution of 256×256 during this stage to facilitate faster training while minimizing computational costs.

Quality Improvement Stage In the second stage, as illustrated in Figure 2(c), we apply a continuous learning strategy by loading the first stage checkpoint and subsequently finetuning all the parameters of the UNet using the same objective function as defined in Equation 1. We train our model on high-quality multilingual aesthetic datasets with a resolution of 512x512 to improve the generative quality. Furthermore, we also drop 10% of text inputs following the setting as SD(Rombach et al. 2022) for classifier-free guidance(Ho and Salimans 2022), which is helpful when calculating the final noisy score during inference. $\tilde{\epsilon}_\theta\left(z_t, c, t\right)$ is obtained by a combination of conditioned score $\epsilon_\theta\left(z_t, c, t\right)$ and unconditioned score $\epsilon_\theta\left(z_t, t\right)$, which is formalized in the following equation:

$$\tilde{\epsilon}_\theta\left(z_t, c, t\right)=\epsilon_\theta\left(z_t, t\right)+\alpha\left(\epsilon_\theta\left(z_t, c, t\right)-\epsilon_\theta\left(z_t, t\right)\right) \quad (2)$$

where $\alpha > 1$ is a scale weight of condition. With the completion of the second stage, we finally obtain a multilingual T2I diffusion model that meet the needs of users across different linguistic backgrounds.

Dataset

Training Data

All the image-text pairs we use to train AD come from LAION (Schuhmann et al. 2022).

LAION 5B LAION 5B includes three sub-datasets: LAION2B-en, LAION2B-multi and LAION1B-nolang. LAION2B-en contains 2.32 billion image-text pairs in English. LAION2B-multi contains 2.26 billions image-text pairs and the text comes from 100+ languages beyond English. In the first training stage, we filter 1.8 billions data in eighteen languages from LAION2B-multi and combine it with LAION2B-en.

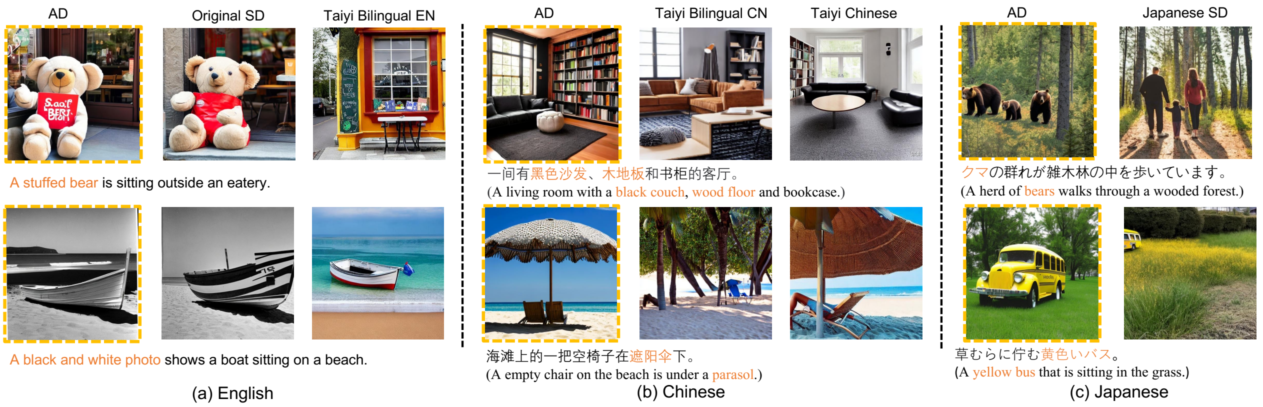
LAION Aesthetics LAION Aesthetics contains several collections of subsets from LAION 5B with high quality. An Aesthetics Predictor is trained using LAION to predict the aesthetics score of images on a scale of 1 to 10, with higher aesthetics score being better. Then the Aesthetics Predictor is used for filtering the data. To conduct the second training stage, we filter eighteen languages from the LAION Aesthetics and the LAION Aesthetics V1-multi dataset with the predicted aesthetics score higher than seven.

Evaluation Benchmark

To evaluate the capability of AD to generate images and capture culture-specific concepts of different languages, we introduce two datasets: Multilingual-General-18(MG-18) for generation quality evaluation and Multilingual-Cultural-18(MC-18) for culture-specific concepts evaluation.

Multilinguale-General-18(MG-18) We construct a large and high-quality dataset MG-18 which contains 7,000 image-text pairs in 18 languages, by expanding XM 3600 with high-quality images from WIT in two steps. In the first step, we use an Optical Character Recognition(Du et al. 2020) system to filter out images with more than five words, considering images with excessive text tend to be document images, which are unsuitable for evaluating the generation capabilities of the T2I model. Next, we use AltCLIP to calculate the similarity score between the image and the caption, and then keep those with a score higher than 0.2.

Multilinguale-Cultural-18(MC-18) One of the important capabilities of multilingual T2I models is to understand culture-specific concepts of different languages. To evaluate this, we construct MC-18, a dataset that contains culture-specific concepts in painting, literature, festival, food, clothe, and landmark. First, we select the representatives of each language in the above six aspects. Then we use ChatGPT to generate prompts and ask the crowdsourcing personnels to select suitable prompts. We create 50 prompts for each of the 18 languages. Some samples of MC-18 are shown in Table 1



A stuffed bear is sitting outside an eatery.

A black and white photo shows a boat sitting on a beach.

(a) English

一间有黑色沙发、木地板和书柜的客厅。
(A living room with a black couch, wood floor and bookcase.)

海滩上的一把空椅子在遮阳伞下。
(A empty chair on the beach is under a parasol.)

(b) Chinese

クマの群れが雑木林の中を歩いています。
(A herd of bears walks through a wooded forest.)

草むらに佇む黄色いバス。
(A yellow bus that is sitting in the grass.)

(c) Japanese

Figure 3: Comparison of generated results with original English SD and other multilingual diffusion models on MG-18. Results of AD are framed with yellow dashed boxes. Obvious differences in generation are highlighted in orange in the prompts.

Experiments

Implement Details

The optimizer is AdamW(Loshchilov and Hutter 2017). The learning rate is 1e-4, with 10,000 warmup steps on 64 NVIDIA A100-SXM4-40GB GPUs.

Follow AltCLIP(Chen et al. 2022), we use knowledge distillation to retrain the multilingual text encoder. Through the training process, the text encoder remains frozen. In addition, we adopt a continuous learning strategy for model training. In the concept align stage, we use the SD v2.1 512-base-ema checkpoint to initialize all parameters except the text encoder, with a batch size of 3,072 and a resolution of 256x256. The training process on LAION2B-en and LAION2B-multi for 330,000 steps takes approximately eight days. In the quality improvement stage, the training starts at the 330,000-step checkpoint, with a batch size of 3,840 on LAION Aesthetics V1-en and V1-multi and 270,000-steps with a resolution of 512x512, which takes around seven days. After that, a new round of training continues from the 270,000-step checkpoint for another 150,000 steps, with 10% of the text randomly discarded for classifier-free guidance learning, taking approximately four days. The teacher model using in knowledge distillation is OpenCLIP ViT-H-14. We also use Xformer and Efficient Attention to save memory use and speed up training. The decay of EMA is 0.9999.

Results on MG-18

We evaluate the general multilingual T2I generative capability of AD on MG-18. We compare AD with two kinds of models. The first is translation-based SD, which requires translating prompts in other languages into English before generation. The second is multilingual baseline diffusion models that beyond English. The inference resolution of all models is 512×512, using 50 DDIM steps and 9.0 classifier-free guidance scale.

Metrics We use FID and IS for evaluating the generation quality, and use Multilingual CLIP(Radford et al. 2021) to calculate the cosine similarity score to evaluate the consistency of generated images with multilingual text.

	AltDiffusion(AD)			Stable Diffusion(SD)		
	FID	IS	C-Sim	FID	IS	C-Sim
English	19.02	26.97	**0.324**	**18.02**	**27.45**	0.322
Chinese	20.32	**29.46**	**0.350**	**18.51**	28.30	0.317
Japanese	18.90	28.13	**0.356**	**18.09**	**29.39**	0.328
Thai	19.94	**27.63**	**0.353**	**19.82**	25.61	0.240
Korean	20.54	27.63	**0.338**	**18.69**	**28.79**	0.284
Hindi	20.92	25.90	**0.338**	**18.52**	**26.30**	0.311
Ukrainian	19.27	28.18	**0.346**	**17.36**	**28.54**	0.314
Arabic	20.32	28.71	**0.346**	**18.34**	**28.90**	0.298
Turkey	19.54	28.53	**0.347**	**17.40**	**28.54**	0.315
Vietnamese	19.02	29.22	**0.346**	**17.02**	**30.86**	0.312
Polish	19.67	29.11	**0.347**	**18.36**	**30.41**	0.327
Dutch	20.14	27.64	**0.350**	**17.91**	**29.48**	0.329
Portuguese	20.78	**28.59**	**0.352**	**18.82**	28.56	0.302
Italian	19.77	27.19	**0.352**	**17.38**	**28.53**	0.317
Spanish	**20.15**	**27.64**	**0.357**	20.41	25.31	0.260
German	18.74	27.58	**0.359**	**17.06**	**28.66**	0.347
French	18.99	28.34	**0.357**	**17.09**	**29.71**	0.341
Russian	19.18	28.26	**0.347**	**17.49**	**29.42**	0.322

Table 2: Comparison of evaluation results with translation-based SD on MG-18. C-Sim means clip similarity.

Compare with Translation-based SD v2.1 We use the original multilingual prompts directly as the input of AltDiffuison. Considering that SD only supports English inputs, we first use the state-of-the-art opensource translation model NLLB-3B[2] to translate other languages into English and then feed them into SD.

As shown in Table 2, AD surpasses translation-based SD in CLIP Sim of all languages and achieves comparable results in FID and IS, proving that AD is better than SD in multilingual understanding and can generate almost the same high-quality images as SD on general prompts.

Compare with Other Baselines We compare AD with other multilingual baseline diffusion models, including Taiyi Chinese, Taiyi Bilingual ,and Japanese SD.

As shown in Table 3, AD outperforms other multilingual baseline diffusion models in all metrics, especially in CLIP

[2] https://huggingface.co/facebook/nllb-200-3.3B

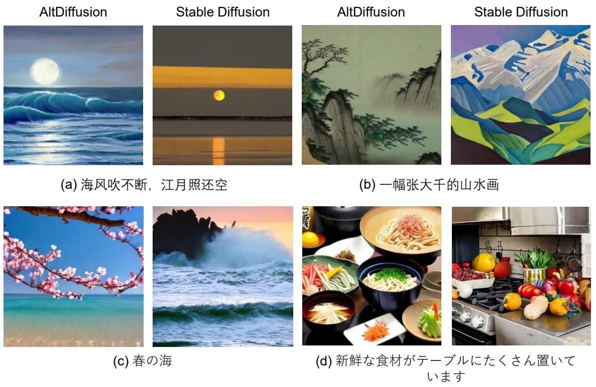
Language	Model	FID(↓)	IS(↑)	CLIP Sim(↑)
English	Taiyi-Bilingual	25.76	26.54	0.261
	AD	**19.02**	**27.45**	**0.324**
Chinese	Taiyi-CN	20.72	28.91	0.276
	Taiyi-Bilingual	23.87	26.96	0.259
	AD	**20.32**	**29.46**	**0.350**
Japanese	Japanese SD	22.78	**30.57**	0.278
	AD	**18.90**	28.13	**0.356**

Table 3: Comparison of zero-shot evaluation results with other multilingual baselines on MG-18.

Sim, AD has achieved 24.1%, 26.8% and 28.1% percent improvement on English, Chinese and Japanese respectively. It shows that AD has stronger image generation capability and multilingual understanding capability than other multilingual baseline models.

To demonstrate the strong capability of AD in multilingual T2I generation, we provide generated images in various languages in Figure 1. As shown in Figure 3(b) and (c), AD can generate images that are more consistent with multilingual text, while other models often ignore or make a mistake in concepts. For example, "black couch", "wood floor" and "parasol" in Chinese, and "bears" and "yellow bus" in Japanese. AD can generate results comparable to SD and better than Taiyi Bilingual in English, as shown in Figure 3(a).

Results on MC-18



(a) 海风吹不断，江月照还空

(b) 一幅张大千的山水画

(c) 春の海

(d) 新鮮な食材がテーブルにたくさん置いています

Figure 4: Comparison of generated results with translation-based SD on MC-18.

Annotators familiar with culture of different countries are asked to conduct a human evaluation for the understanding capability of the model in culture-specific concepts.

Evaluation Setting We assign three annotators to each language. Annotators see two images generated by the same prompt generated by AD and SD, respectively. Then they are asked to score from two dimensions: Culture Consistency, and Image-Text Consistency, scoring from 1-5. After scoring, the annotators select the final result from ["Alt is better", "SD is better", "Same"]. Then we calculate the total score according to the following formula:

$$(Total_{Alt}, Total_{SD}) = \left(\frac{|A| + 0.5 \cdot |C|}{N}, \frac{|B| + 0.5 \cdot |C|}{N} \right) \quad (3)$$

where $|A|$, $|B|$ and $|B|$ is the count of "Alt is better","SD is better" and "Same", $N = |A| + |B| + |C|$. The results for each language are the average of 3 annotator scores.

Evaluation Results As shown in Table 4, AD beats SD for the final total scores in all languages and outstands in Cultural and Image-Text Consistency, showing that AD performs better in multilingual understanding. The evaluation results indicate that through training with large-scale multilingual data, culture-specific concepts of different languages can be injected into the model.

	AltDiffusion(AD)			Stable Diffusion(SD)		
	Culture Cons	T-I Cons	Total	Culture Cons	T-I Cons	Total
Chinese	**4.125**	**3.775**	**0.769**	3.356	3.350	0.231
Japanese	**4.507**	**4.487**	**0.757**	3.301	3.253	0.243
Thai	**4.027**	**4.044**	**0.648**	3.383	3.579	0.352
Korean	**3.236**	**3.371**	**0.607**	3.135	3.287	0.393
Hindi	**4.824**	**4.301**	**0.523**	4.784	4.261	0.477
Ukrainian	**4.422**	**3.955**	**0.641**	4.322	3.905	0.359
Arabic	**4.824**	**4.401**	**0.609**	4.647	4.314	0.391
Turkey	**3.376**	**3.293**	**0.510**	3.328	3.241	0.490
Vietnamese	**3.637**	**3.511**	**0.533**	3.467	3.396	0.467
Polish	**4.243**	**3.735**	**0.546**	4.130	3.676	0.454
Dutch	**4.495**	**4.465**	**0.527**	4.195	4.215	0.473
Portuguese	**3.757**	**3.627**	**0.639**	3.639	3.509	0.361
Italian	**3.586**	**3.449**	**0.565**	3.512	3.375	0.435
Spanish	**4.202**	**4.113**	**0.554**	4.138	4.064	0.446
German	**3.981**	**3.916**	**0.698**	3.825	3.688	0.302
French	**4.655**	**4.582**	**0.527**	4.652	4.579	0.473
Russian	**3.258**	**3.086**	**0.556**	2.868	3.002	0.444

Table 4: Comparison of human evaluation results with translation-based SD on MC-18. T-I cons means Text-Image consistency.

We illustrate the performance difference between AD and SD on MC-18 in Figure 4. AD has a better understanding capability in culture-specific concepts. For example, in Figure 4(a), the prompt is actually a Chinese poem, but SD generates a realistic image. In Figure 4(b), models are asked to generate a landscape painting by Chinese traditional painter Daqian Zhang. The image of AD shows the characteristics of Chinese traditional painting, while the image of SD is an oil painting. As for "the spring sea" in Japanese in In Figure 4(c), AD generates cherry blossoms more suitable for artistic concept.

Application

Genernal Application Figure 6 shows the general capacities of AD, such as Image to Image(Img2Img) and Inpainting. AD supports users to to use Image to Image or Inpaint function directly use languages beyond English, e.g., Chinese.

Compatibility with Downstream Tools Although large T2I models achieve impressive performance, they still lack controllability in specific commercial applications. Recently, some methods, such as ControlNet(Zhang and Agrawala

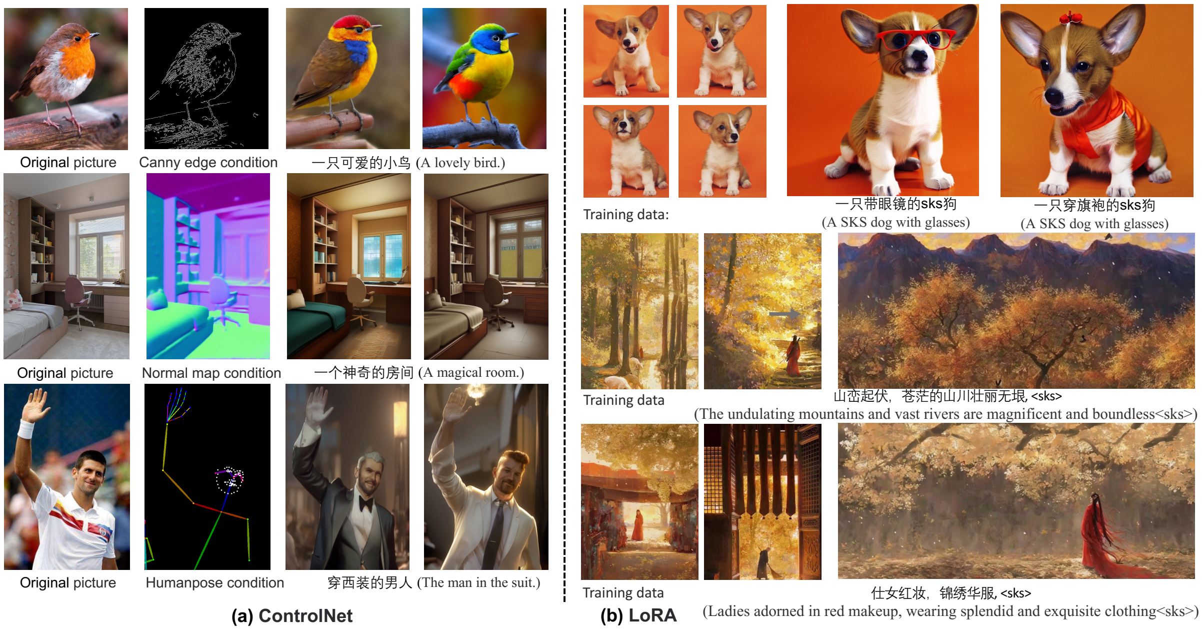

Original picture
Canny edge condition
一只可爱的小鸟 (A lovely bird.)
Original picture
Normal map condition
一个神奇的房间 (A magical room.)
Original picture
Humanpose condition
穿西装的男人 (The man in the suit.)
(a) ControlNet
Training data:
一只带眼镜的sks狗
(A SKS dog with glasses)
一只穿旗袍的sks狗
(A SKS dog with glasses)
Training data
山峦起伏，苍茫的山川壮丽无垠, <sks>
(The undulating mountains and vast rivers are magnificent and boundless<sks>)
Training data
仕女红妆，锦绣华服, <sks>
(Ladies adorned in red makeup, wearing splendid and exquisite clothing<sks>)
(b) LoRA

Figure 5: AD has strong compatibility with downstream T2I tools such as ControlNet and LoRA.

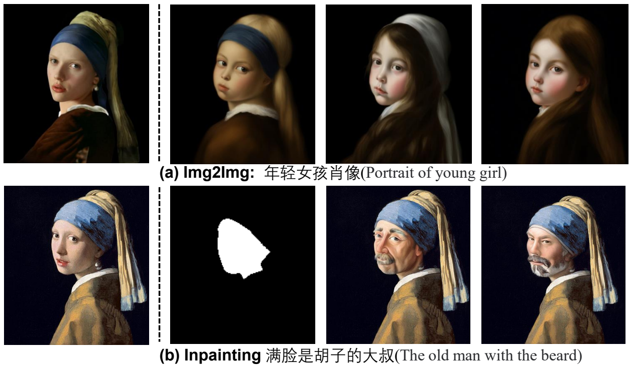

(a) Img2Img: 年轻女孩肖像(Portrait of young girl)
(b) Inpainting 满脸是胡子的大叔(The old man with the beard)

Figure 6: The images generated by Img2Img and Inpainting function using AD .

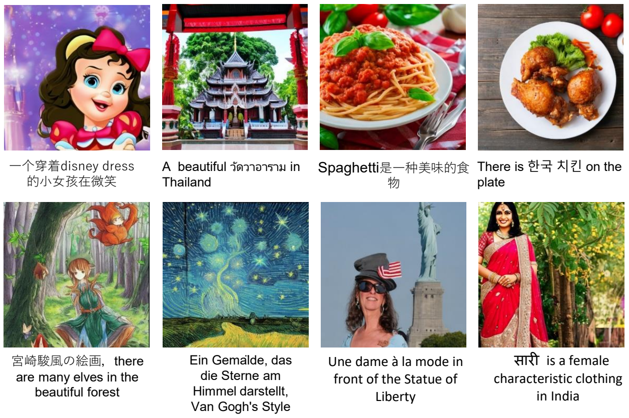

一个穿着disney dress 的小女孩在微笑
A beautiful วัดวาอาราม in Thailand
Spaghetti是一种美味的食物
There is 한국 치킨 on the plate
宮崎駿風の絵画，there are many elves in the beautiful forest
Ein Gemälde, das die Sterne am Himmel darstellt, Van Gogh's Style
Une dame à la mode in front of the Statue of Liberty
साड़ी is a female characteristic clothing in India

Figure 7: Images generated by mixed languages using AD.

2023) and LoRA(Hu et al. 2021), have garnered widespread attention for enhancing model controllability. Compatibility with these downstream tools is essential for a Large T2I. As shown in Figure 5, AD is totally compatible with ControlNet and LoRA. Thus users can use their imagination to create images easily.

Mixed Language Generation As shown in Figure 7, AD supports mixed language input. It will be very troublesome if the model can only support English because users need to translate various languages into English and then concatenate them as input. AD can freely combine different languages, such as Thai and English, Japanese and English, Chinese and Korean, etc.

Conclusion

This paper introduces AltDiffusion(AD), a multilingual T2I diffusion model that supports eighteen languages. We train a multilingual text encoder and plug it into pretrained diffusion model, and then train the diffusion model using a two-stage training schema. In addition, we introduce a benchmark to evaluate AD, including two datasets focusing on general and culture-specific evaluation: MG-18 and MC-18. Experimental results show AltDiffusion outperforms current state-of-the-art T2I models, e.g., Stable Diffusion in multilingual understanding, especially with respect to culture-specific concepts, while still having comparable capability for generating high-quality images. Meanwhile, as a large multilingual T2I diffusion model, AD is compatible with all downstream T2I tools, e.g., ControlNet and LoRA, which may promote research and application in multilingual T2I.

Acknowledgements

The work was supported by the National Key R&D Program of China (2022ZD0116312). We extend our gratitude to LAION for open-sourcing the dataset, to StabilityAI for open-sourcing the base models, and to team of stable diffusion researchers for providing helpful suggestions. Lastly, we acknowledge the data and infrastructure team at BAAI for their support throughout the project.

References

Aggarwal, P.; and Kale, A. 2020. Towards zero-shot Cross-lingual Image retrieval. *arXiv preprint arXiv:2012.05107.*

Carlsson, F.; Eisen, P.; Rekathati, F.; and Sahlgren, M. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6848–6854.

Chen, Z.; Liu, G.; Zhang, B.-W.; Ye, F.; Yang, Q.; and Wu, L. 2022. AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities. *arXiv preprint arXiv:2211.06679.*

Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; and Jitsev, J. 2022. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143.*

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116.*

Ding, M.; Yang, Z.; Hong, W.; Zheng, W.; Zhou, C.; Yin, D.; Lin, J.; Zou, X.; Shao, Z.; Yang, H.; and Tang, J. 2021. CogView: Mastering Text-to-Image Generation via Transformers. *arXiv preprint arXiv:2105.13290.*

Ding, M.; Zheng, W.; Hong, W.; and Tang, J. 2022. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. *arXiv preprint arXiv:2204.14217.*

Du, Y.; Li, C.; Guo, R.; Yin, X.; Liu, W.; Zhou, J.; Bai, Y.; Yu, Z.; Yang, Y.; Dang, Q.; et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941.*

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30K: Multilingual English-German Image Descriptions. *Cornell University - arXiv,Cornell University - arXiv.*

Feng, Z.; Zhang, Z.; Yu, X.; Fang, Y.; Li, L.; Chen, X.; Lu, Y.; Liu, J.; Yin, W.; Feng, S.; et al. 2022. ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model with Knowledge-Enhanced Mixture-of-Denoising-Experts. *arXiv preprint arXiv:2210.15257.*

Feng, Z.; Zhang, Z.; Yu, X.; Fang, Y.; Li, L.; Chen, X.; Lu, Y.; Liu, J.; Yin, W.; Feng, S.; et al. 2023. ERNIE-ViLG 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10135–10145.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing Concepts from Diffusion Models. *arXiv preprint arXiv:2303.07345.*

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718.*

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598.*

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685.*

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

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101.*

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2017. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*, 74–93.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML.*

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125.*

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 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*, 234–241. Springer.

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

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Schamoni, S.; Hitschler, J.; and Riezler, S. 2018. A Dataset and Reranking Method for Multimodal MT of User-Generated Image Captions. *Conference of the Association for Machine Translation in the Americas,Conference of the Association for Machine Translation in the Americas*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

Shing, M.; and Sawada, K. 2022a. Japanese Stable Diffusion. https://github.com/rinnakk/japanese-stable-diffusion.

Shing, M.; and Sawada, K. 2022b. Japanese Stable Diffusion. https://github.com/rinnakk/japanese-stable-diffusion.

Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; and Najork, M. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2443–2449.

Thapliyal, A. V.; Pont-Tuset, J.; Chen, X.; and Soricut, R. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*.

Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; Huang, Y.; Li, X.; Wu, Y.; Lu, J.; Zhu, X.; Chen, W.; Han, T.; Pan, K.; Wang, R.; Wang, H.; Wu, X.; Zeng, Z.; Chen, C.; Gan, R.; and Zhang, J. 2022a. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; et al. 2022b. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv preprint arXiv:2209.02970*.

Zhang, H.; Yin, W.; Fang, Y.; Li, L.; Duan, B.; Wu, Z.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. ERNIE-ViLG: Unified generative pre-training for bidirectional vision-language generation. *arXiv preprint arXiv:2112.15283*.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.