Bridge Diffusion Model: Bridge Chinese Text-to-Image Diffusion Model with English Communities

Shanyuan Liu, Bo Cheng, Yuhang Ma, Liebucha Wu
Ao Ma, Xiaoyu Wu, Dawei Leng*, Yuhui Yin
360 AI Research
{liushanyuan, chengbo1, mayuhang, wuliebucha}@360.cn
{maao, wuxiaoyu1, lengdawei, yinyuhui}@360.cn

Abstract

Text-to-Image generation (TTI) technologies are advancing rapidly, especially in the English language communities. However, apart from the user input language barrier problem, English-native TTI models inherently carry biases from their English world centric training data, which creates a dilemma for development of other language-native TTI models. One common choice is to fine-tune the English-native TTI model with translated samples. It falls short of fully addressing the model bias problem. Alternatively, training non-English language native models from scratch can effectively resolve the English world bias, but model trained this way would diverge from the English TTI communities, thus not able to utilize the strides continuously gaining in the English TTI communities any more. To build Chinese TTI model meanwhile keep compatibility with the English TTI communities, we propose a novel model structure referred as "Bridge Diffusion Model" (BDM). The proposed BDM employs a backbone-branch network structure to learn the Chinese semantics while keep the latent space compatible with the English-native TTI backbone, in an end-to-end manner. The unique advantages of the proposed BDM are that it's not only adept at generating images that precisely depict Chinese semantics, but also compatible with various English-native TTI plugins, such as different checkpoints, LoRA, ControlNet, Dreambooth, and Textual Inversion, *etc*. Moreover, BDM can concurrently generate content seamlessly combining both Chinese-native and English-native semantics within a single image, fostering cultural interaction.

Code — https://github.com/360CVGroup/Bridge_Diffusion_Model

Model — https://huggingface.co/qihoo360/BDM1.0

Introduction

The latest advancements in diffusion models have significantly transformed the process of text-to-image generation. With extensive training data and model parameters, diffusion models(Ho, Jain, and Abbeel 2020) can now vividly depict visual scenes based on written prompts, allowing users to effortlessly generate beautiful images using natural language. Among these models, Stable Diffusion(Rombach et al. 2022) has emerged as a widely adopted and community-driven technology. It maps images from pixel space to latent space, resulting in impressive image quality while significantly reducing memory usage. The SD community has made important progress, introducing various LoRA(Hu et al. 2021) models and checkpoints. These additions enhance the base model's capabilities, enabling it to generate more refined or personalized content within specific domains. It's worth noting that the rapid progress of these technologies is mainly centered around English-language communities.

However, current models exhibit language-related biases. As pointed out by (Luccioni et al. 2023; Miller et al. 2023), Text-to-Image (TTI) models designed for English speakers, such as DALL-E 2(Ramesh et al. 2022), Stable Diffusion(Rombach et al. 2022) versions 1.4 and 2, tend to disproportionately emphasize characteristics associated with white individuals and males. These language-related biases are inherent and widespread in current TTI models, primarily because they are predominantly trained on data from the English-speaking world, as exemplified by the commonly used LAION dataset(Schuhmann et al. 2022). Consequently, there is an over-representation of English-speaking figures and an inadequate representation of non-English-speaking counterparts.

The primary focus of this work is to design a TTI model that is compatible with multiple languages. By "compatible", we mean not only supporting non-English prompt input but also capable of generating images that align with common sense in non-English languages. While translation-based methods are straightforward and cost-effective, they can only address the non-English input capability, leaving the inherent model bias untouched. Another approach involves alignment-based strategies, which align the embedding space of different language text encoders with parallel translation text corpus. However, this method is essentially another form of "translation". Taiyi-Stable-Diffusion-1B-Chinese-EN-v0.1(Zhang et al. 2022) took this route, fine-tuning the TTI model with Chinese-native data based on an aligned text encoder. This allows the English-native model to incorporate Chinese-native language semantics at a low cost while maintaining some level of compatibility between the English and Chinese TTI communities, although achieving the right balance is challenging. However, when tasked with capturing intricate nuances of native language seman-

*Corresponding authors.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Figure 1: Selected samples generated by our BDM model, with Chinese prompts and different English plugins from the Stable Diffusion community.

tics or language-specific concepts, the effectiveness of this approach may be notably limited. To address the inherent bias in English-native models, the most radical approach is to train a TTI model from scratch with non-English native data. For example, ERNIE-ViLG 2.0(Feng et al. 2023) and Wukong-Huahua(Gu et al. 2022) are trained with Chinese native data and can generate high-quality images consistent with Chinese language semantics. However, a fundamental challenge with this approach is that it loses compatibility with its ancestral English-native models, preventing the direct utilization of progress from the English-native TTI communities. This could lead to community isolation and development stagnation for the Chinese-native TTI community in the long run.

We propose a new diffusion network structure referred to as the "Bridge Diffusion Model" (BDM) to address the previously mentioned challenge. The unique advantages of the proposed BDM lie in its ability to precisely generate images following non-English language native semantics while also being compatible with English-native TTI communities. Existing techniques from English-native TTI communities, such as different checkpoints, LoRA, ControlNet, Dreambooth, and Textual Inversion, can all be directly applied in BDM. BDM effectively addresses the language-related bias in TTI models while maintaining interoperability between non-English language native TTI communities and English-native communities. This is where the name "Bridge" originates. In this context, we specifically focus on implementing Chinese language native TTI realization, and the method could be applicable for any other non-English language native TTI model.

Our approach involves employing a backbone-branch network architecture, similar to ControlNet(Zhang, Rao, and Agrawala 2023), as illustrated in Fig.2. The backbone remains frozen during training and can be sourced from any pre-trained diffusion TTI model. In our current implementation, we utilize Stable Diffusion 1.5(Rombach et al. 2022). The branch functions as a module for injecting language-native semantics, and its parameters are trained using language-native text-image pairs. In contrast to ControlNet, BDM's branch incorporates a Chinese-native CLIP(Yang et al. 2022) as the text encoder, responsible for processing the non-English language-native text prompts. The English-native text encoder is fed with an empty constant string ("") in our implementation.

For model inference, language-native prompts will be processed through the Chinese text encoder from BDM's branch part. Simultaneously, we can still input an empty constant string ("") into the English text encoder. Since BDM incorporates an entire English-native TTI model as its backbone, existing techniques like LoRA(Hu et al. 2021), ControlNet(Zhang, Rao, and Agrawala 2023), Dreambooth(Ruiz et al. 2023), Textual Inversion(Gal et al. 2023), and even various style fine-tuned checkpoints from English TTI communities (such as Civitai(civitai 2023), Stable Diffusion Online(stablediffusionweb 2023), to name a few) can be seamlessly applied to BDM with minimal cost.

In summary, the primary contributions of this work are as follows: (1) We propose a backbone-branch network architecture referred as BDM. By integrating an English TTI backbone with a Chinese-native semantics injection branch, BDM is able to solve TTI model's English-native bias meanwhile keep the compatibility with its ancestral English-native communities. (2) We proposed a tailored training strategy for BDM. By utilizing the latent space corresponding to the empty constant string ("") of the English backbone, we aligned the Chinese-native latent space with the English latent space, enabling BDM to incorporate English communities plugins. (3) With thorough experiments, we verify that the BDM trained with Chinese-native data can utilize techniques developed for English-native TTI models, such as LoRA, Dreambooth, Textual Inversion, ControlNet, *etc*. Thus bridge the interoperation between non-English and English-native TTI communities.

Related Work

In recent years, the field of TTI generation has experienced remarkable growth. Early endeavors employed Generative Adversarial Networks (Goodfellow et al. 2017) to produce images from textual descriptions. This was achieved by harnessing the adversarial interplay between a generator and a discriminator to yield lifelike images. With the resounding success of the transformer architecture, a subset of research shifted the generation task towards a sequence-to-sequence paradigm, training generators in an autoregressive manner. Noteworthy instances encompass ERNIE-ViLG (Zhang et al. 2021), DALL-E (Ramesh et al. 2021), and Parti (Yu et al. 2022). More recently, diffusion models have garnered acclaim for achieving cutting-edge outcomes in this domain (Ramesh et al. 2022; Rombach et al. 2022; Saharia et al.

2022b). These methods have continually refreshed the metrics in this field by iteratively injecting text conditions during the denoising process. Evident progress can be seen in works like LDM (Rombach et al. 2022), DALL-E 2 (Ramesh et al. 2022), and Imagen (Saharia et al. 2022b). Recently, diffusion models based on the DIT have been emerging in abundance, with some even achieving state-of-the-art performance, such as SD3.5(AI 2024) and Flux(Labs 2024). The primary aim of our research is to leverage the backbone-branch architecture to empower the network in generating images that possess a profound grasp of native language semantic understanding.

Moreover, distinct research endeavors have concentrated on directing the output of diffusion models to finely manage image content and style. Techniques like Dreambooth(Ruiz et al. 2023) and Textual Inversion(Gal et al. 2023) bestow precise control over the attributes of generated images, accomplishing objectives analogous to reference images. LoRA(Hu et al. 2021) facilitates lightweight fine-tuning of the base model to achieve specified objectives or styles. ControlNet(Zhang, Rao, and Agrawala 2023), T2I-Adapter(Mou et al. 2023) and HiCo(Cheng et al. 2024) facilitate meticulous control over diverse conditions by incorporating branches into pretrained base models. Within our study, our focus is centered around seamlessly integrating BDM with the control plugins from the aforementioned English communities.

In addition, research has also been devoted to enhancing the multilingual capability of TTI models to support non-English input captions. For example, Altclip(Chen et al. 2023) extended the text encoder of diffusion models using the pretrained multilingual text encoder XLM-R. ERNIE-ViLG 2.0(Feng et al. 2023) embarked on training a Chinese diffusion model from scratch using Chinese image-text pairs. (Li et al. 2023) alleviated the language barrier by translating English captions into other languages through a neural machine translation system. Although they improved the support for multilingual text inputs, they did not achieve the integration of native language and English communities, while our focus is on incorporating native language models into the English communities.

Preliminary

Diffusion model

Diffusion models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Song and Ermon 2019; Vincent 2011) are a class of generative models that have gained significant attention recently due to their advancements in image generation (Dhariwal and Nichol 2021; Kawar, Ganz, and Elad 2022; Song et al. 2020; Vahdat, Kreis, and Kautz 2021), leading to the latest technological developments in several downstream applications. These applications include image restoration (Kawar et al. 2022a; Saharia et al. 2022a), adversarial purification (Blau et al. 2022; Nie et al. 2022), image compression (Theis et al. 2022), image classification (Zimmermann et al. 2021), among others in various fields (Gao et al. 2022; Kawar et al. 2022b; Popov et al. 2021; Sasaki, Willcocks, and Breckon 2021).

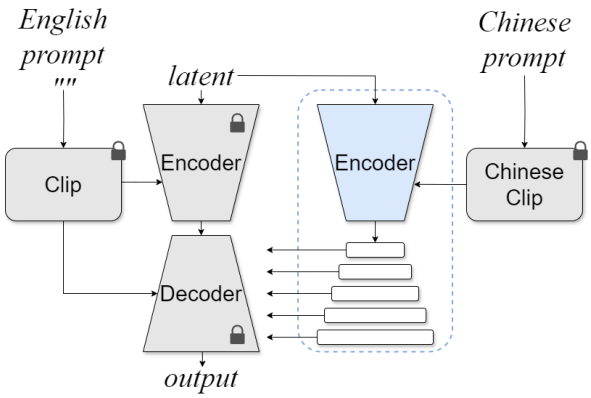


Figure 2: BDM structure.

The fundamental concept of DDPM(Ho, Jain, and Abbeel 2020) involves iteratively applying diagonal Gaussian noise to the initial data sample (denoted as x) and, after T steps, transforming it into an isotropic Gaussian distribution

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon_t, \quad t \in \{1, \ldots, T\} \quad (1)$$

Here, the sequence $\{x_t\}$ starts from $x_0 = x$ and ends with $x_T \sim \mathcal{N}(0, I)$, where each step adds noise $\epsilon_t \sim \mathcal{N}(0, I)$, and $\{\alpha_t\}_{1\ldots T}$ represents the predefined schedule . The training objective of this network is straightforward denoising, aiming to make $\epsilon_\theta(x_t, t) \approx \epsilon_t$. This leads to a highly consistent learned image distribution with the target distribution, resulting in excellent generative performance.

Method

The overall framework of BDM is shown in the Fig.2. It adopts a backbone-branch architecture similar to ControlNet(Zhang, Rao, and Agrawala 2023) and the backbone is responsible for providing the same latent space as the English communities, while the branch is responsible for finding appropriate offsets in the latent space, allowing native language semantic features to be injected into the English latent space.

backbone-branch architecture

For BDM, we embrace a backbone-branch architecture akin to ControlNet(Zhang, Rao, and Agrawala 2023). The backbone employs Stable Diffusion 1.5(Rombach et al. 2022), and the branch consists of a learnable parameter replica derived from the backbone, with convolutional layers responsible for processing conditional image inputs omitted. In addition, the backbone and branch process the same latent space features. For the backbone, text encoding is accomplished via OpenAI CLIP(Radford et al. 2021), while the branch employs Chinese CLIP(Yang et al. 2022) for text encoding. This framework harmoniously marries the latent space and generation capabilities of BDM with the English model. Simultaneously, the branch accommodates native language semantics, thereby facilitating the creation of a native language model that seamlessly aligns with the English communities' dynamics.

training strategy

Image diffusion models learn the process of progressively denoising images to generate new samples. The denoising can take place either in the pixel space or a latent space that is encoded from the training data. Stable Diffusion, for instance, employs latent images as the domain for training. In this context, the terms "image", "pixel", and "denoising" all refer to their corresponding concepts in the "perceptual latent space" (Rombach et al. 2022).

The diffusion algorithms start with an initial image z_0, and then iteratively add noise to produce a series of noisy images z_t, where t indicates the number of times noise is added. As t increases, the image approximates pure noise. In this context, image diffusion algorithms learn a network ϵ_θ to predict the noise to be added to the noisy image z_t, given a set of conditions, including the time step t ,text prompts for the English backbone c_{ent} ,text prompts for the native language branch c_{nlt}

$$\begin{aligned}\mathcal{L} =&\mathbb{E}_{\boldsymbol{z}_0,t,\boldsymbol{c}_{ent},\boldsymbol{c}_{nlt},\epsilon\sim\mathcal{N}(0,1)}\\ &\left[||\epsilon-\epsilon_\theta(z_t,t,\boldsymbol{c}_{ent},\boldsymbol{c}_{nlt})||_2^2\right]\end{aligned} \tag{2}$$

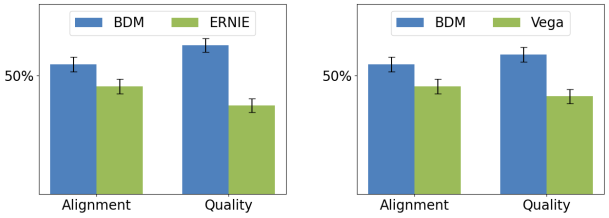
The learning objective $\mathcal{L}$ represents the overall objective of the entire diffusion model and can be directly employed for training BDM.

Our first key finding here is that to successfully train the BDM model, it's important to only inject text prompt information through the non-English language branch, meanwhile leave the text input of the English backbone empty, otherwise the training would not converge. Reasons are two folds: on the one hand, the English backbone here is used only for latent space alignment, *i.e.* for keeping BDM's compatibility with the English SD model, thus there's no need for text input of the English backbone; one the other hand, if the text input of the English backbone is non-empty, it would produce strong effect on the UNet denoising process, thus would significantly impede the effective injection of native language semantics through the branch network. Therefore, throughout the training process, we consistently set c_{ent} as an empty string, ensuring a dependable and aligned latent space for the BDM. The BDM utilizes the latent space of English backbone to find the optimal shift in native language semantics within the English domain, akin to how English community plugins seek specific semantic shifts in the latent space of Stable Diffusion 1.5. It's worth noting that, with a shared latent space, English community plugins seamlessly integrate with each other, and the BDM, following the same principle and latent space, can also seamlessly integrate with these plugins, as verified in subsequent experiments.

inference strategy

Unlike the training strategy mentioned earlier, our second key finding is that, during inference, BDM exhibits additional versatility. By manipulating the text prompt input to different language branches, we can generate images that emphasize Chinese semantics, English semantics, or even a combination of both.

To start, we have the option to set both the prompt and negative prompt of the English backbone to empty strings.



(a) ERNIE-ViLG-2.0 (b) Vega

Figure 3: Comparison of BDM and ERNIE-ViLG-2.0 (referred to as ERNIE in the figure)/Vega on BDM-870 with human evaluation. In the human experiments, BDM utilized realisticVisionV51 as the English backbone to achieve enhanced generation results.

This practice aligns with the training phase, solely utilizing prompts from the native language branch to infuse semantics. We observed that using appropriate positive/negative prompts in the English backbone during the inference stage can enhance image quality or alter the visual style. Therefore, for better generation results or a more suitable style, both Chinese and English prompts, along with negative prompts, can be adjusted based on user preferences or requirements. Moreover, the weight of the native language branch can also be changed according to user needs.

BDM also has the ability to integrate various plugins from the English communities. Incorporating checkpoint or Dreambooth(Ruiz et al. 2023) involves a direct switch from the backbone's Stable Diffusion 1.5 to the desired checkpoint/Dreambooth. When incorporating LoRA(Hu et al. 2021)—be it style LoRA or object LoRA—embedding the LoRA model parameters into the backbone is feasible. If LoRA contains trigger words, it is crucial to input these trigger words into the English backbone. Similarly, when integrating with ControlNet(Zhang, Rao, and Agrawala 2023), the ControlNet branch can seamlessly combine with the backbone, resulting in a dual-branch configuration comprising the BDM native language branch and the original ControlNet branch. Regarding Textual Inversion(Gal et al. 2023) integration, Textual Inversion's embedding can be directly loaded into the backbone's prompt or negative prompt. The combination of these operations can be customized according to specific needs.

In contrast, when it comes to constructing native language models from scratch, it becomes imperative to retrain requisite plugins like LoRA due to the lack of compatibility with English-speaking communities. Nevertheless, BDM stands out by seamlessly incorporating plugins from English-speaking communities, providing an inherent advantage that facilitates effortless compatibility with these communities.

Experiments

This section offers an overview of the experimental setup and showcases the effectiveness of BDM through both qualitative and quantitative demonstrations.

Model	Data distribution	FID ↓
ERNIE-ViLG (Zhang et al. 2021)	English and Chinese	14.70
LDM (Rombach et al. 2022)	English	12.61
GLIDE (Nichol et al. 2022)	English	12.24
DALL-E 2 (Ramesh et al. 2022)	English	10.39
Stable Diffusion (Rombach et al. 2022)	English	**8.59**
eDiff-I (Balaji et al. 2022)	English	6.95
ERNIE-ViLG 2.0(Feng et al. 2023)	English and Chinese	6.75
RAPHAEL (Xue et al. 2023)	English	6.61
BDM	Chinese	**9.93**

Table 1: Comparison of BDM and representative text-to-image generation models on MS-COCO 256 × 256 with zero-shot FID-30k. We use classifier-free guidance scale 3.7 and Chinese language branch weight 1.5 for our diffusion model. Here we choose SD1.5 as the English backbone for BDM

class	human		architecture	
culture	Chinese	Caucasian	Chinese	Caucasian
BDM(RealV51)	**23.48**	22.05	**28.59**	27.81
RealV51	22.86	**23.24**	26.19	**28.08**
class	food		festival	
culture	Chinese	Caucasian	Chinese	Caucasian
BDM(RealV51)	**24.77**	24.02	**23.75**	20.77
RealV51	23.65	**24.81**	19.20	**24.03**

Table 2: CLIP score↑ to measure Chinese cultural inclination

Experimental Setup

The training dataset comprises around one billion image-text pairs, which includes various internal Chinese datasets and parts of publicly available English datasets. To ensure the generation of meaningful Chinese concepts and minimize the impact of English concepts, our dataset is predominantly composed of Chinese data. We exclude data with watermarks, low aesthetic quality, or irrelevant image-text content.

Utilizing the latent space of Stable Diffusion 1.5(Rombach et al. 2022), BDM employs the same Variational Autoencoder (VAE) to facilitate transformations between images in pixel space and the latent space. The entire model is built using PyTorch and we use the AdamW(Loshchilov and Hutter 2019) optimizer for training, setting a learning rate of 1e-5 and a batch size of 3200. The training process spans two months on 80 NVIDIA A800 GPUs.

Quantitative Evaluation

Human Evaluation on BDM-870. We collected 870 diverse Chinese prompts from real users as benchmark for human evaluation of BDM model, and named it BDM-870. The BDM-870 are intentionally diversified, containing 25 categories. These 25 categories are evenly distributed, and more detailed data content is available in this link[1]. Based on BDM-870, we conducted back-to-back evaluations on image quality and text-image alignment, at 512×512 resolution, against recent state-of-the-art models, such as ERNIE-ViLG 2.0 and Vega AI[2](vegaai 2023). Seven evaluators were presented with two sets of images generated by BDM and the competing model, but unaware of which. They are required to pick out the better image with respect to either image quality or text-image alignment. For image quality, aesthetics, object integrity, rationality of details such as face, fingers and limbs are mainly considered. Results are illustrated in Fig.3, and human raters prefer BDM over ERNIE-ViLG 2.0 by 54.57%±3.1% and 62.65%±2.9% and over Vega AI by 54.62%±3.1% and 58.79%±3.0% quantitatively. For Chinese-native text-to-image generation, BDM outperformed all other models considered in human evaluation.

Evaluation on COCO. In line with previous research(Saharia et al. 2022b; Feng et al. 2023; Balaji et al. 2022; Rombach et al. 2022), we conducted an evaluation of BDM on the COCO 256 × 256 dataset using the zero-shot Frechet Inception Distance (FID), a metric that assesses image quality and diversity. We randomly selected 30,000 images from the validation set for assessment and translate the English captions into Chinese automatically. During training, BDM predominantly employs Chinese text-image data, resulting in a significant disparity between

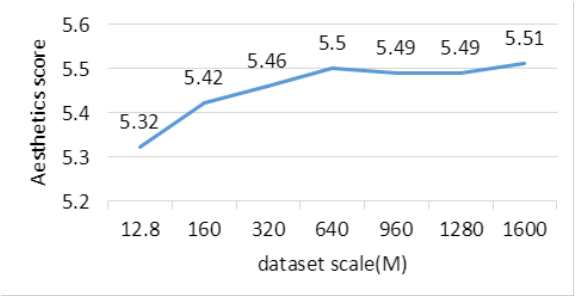


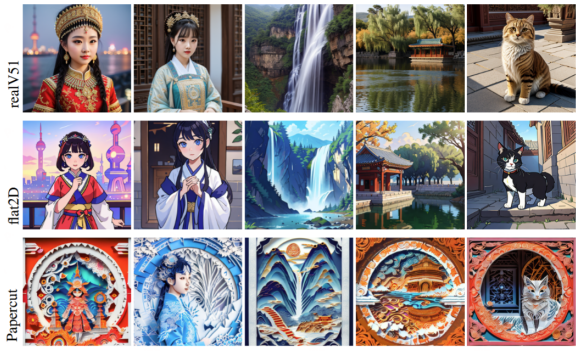
Figure 4: Training data scale.

(1)广州早茶Cantonese morning tea(2)锅包肉Guobaorou
(3)红糖糍粑Brown Sugar Glutinous Rice Cake(4)鹤crane
(5)起重机crane

Figure 5: Native language semantics.

[1]https://github.com/360CVGroup/Bridge_Diffusion_Model

[2]Vega AI is a leading Chinese text-to-image creation platform.



(1) 一个美丽女孩的肖像，苗族服饰，在上海外滩，背景是东方明珠
The portrait of a beautiful girl, Miao costumes,is on the Bund of Shanghai, with the Oriental Pearl TV Tower in the background
(2) 一个美丽女孩的肖像，身穿蓝色汉服，在南锣鼓巷
Portrait of a beautiful girl wearing a blue Hanfu in Nanluogu Lane
(3) 飞流直下三千尺，疑是银河落九天。Flying down three thousand feet, it is suspected that the Milky Way falls nine days
(4) 颐和园，昆明湖。The Summer Palace, Kunming Lake
(5) 中华田园猫在北京四合院玩耍
Chinese Pastoral Cat Playing in Beijing Siheyuan

Figure 6: Checkpoint.

(1) 一个女孩在故宫里唱京剧
A girl is singing Beijing Opera in the Forbidden City
(2) 一个女孩在亭子里弹古筝
A girl is playing a guzheng in the pavilion
(3) 一个女孩身穿汉服，头戴发饰，手里拿着月饼
A girl dressed in Hanfu, wearing hair accessories, and holding mooncakes in her hand

Figure 7: LoRA.

一个女孩头戴凤冠，身披霞帔
A girl wearing a phoenix crown and a Xia Pei

Figure 8: ControlNet.

the world concepts in the output of BDM and COCO. Furthermore, there is a loss of accuracy in the process of automatically translating COCO captions into Chinese. In this context, BDM still manages to achieve FID of 9.93, maintaining a performance that closely aligns with SD1.5's 8.59. It even outperforms some English models, as shown in Table1. This illustrates that BDM can effectively preserve the performance of English backbone with minimal loss while accommodating the generation of Chinese concepts. Moreover, to demonstrate the effectiveness of BDM's structure and strategies, no additional optimizations were applied to enhance the generation quality for both the English backbone and Chinese language branch. This highlights BDM's ability to easily and effectively realize a non-English language native model that is compatible with the English-speaking community.

Chinese cultural inclination. To validate the Chinese cultural inclination of BDM, we generated 25 general prompts about concepts such as race, traditional architecture, food, and festivals using GPT. Then, we used BDM(RealisticVisionV51) and RealisticVisionV51 to generate 750 images. We calculated the CLIP score between the generated images and the concepts of "Chinese" and "Caucasian" to measure the model's cultural inclination, results are present in Table 2. The table clearly shows that the CLIP score between BDM (RealV51) and "Chinese" consistently surpasses the CLIP score between BDM (RealV51) and "Caucasian", suggesting a stronger inclination towards generating Chinese cultural semantics. Conversely, the English model exhibits the opposite trend. Furthermore, the CLIP score between BDM (RealV51) and "Chinese" surpasses that between RealV51 and "Chinese", while the opposite trend is observed for "Caucasian", indicating that BDM demonstrates a stronger inclination towards Chinese cultural context compared to the English model.

Training data scale. To show how the performance of BDM scales with data size, we present aesthetic metric records along BDM's training, as shown in Fig.4. BDM's performance continuously improves with more training data.

Qualitative Results

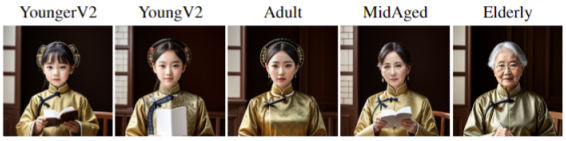
In this section, the semantic information for all images comes exclusively from the Chinese language branch which means that the English backbone receives only general descriptions such as "high definition", style descriptors like "CG", "anime", or trigger words, lacking any significant semantic input. All the plugins are obtained from Civitai.

Native language semantics. In Fig.5, we demonstrate BDM's capability to generate native language semantic images. The original results of SD1.5 and the two versions of BDM are showcased: BDM(SD1.5) with Stable Diffusion 1.5(Rombach et al. 2022) as the backbone, and BDM(RealisticVisionV51) with RealisticVisionV51(SG_161222 2023) as the backbone. As shown by the results of prompts 1-3, the Chinese semantics generated by the English model are all incorrect, while BDM can generate accurately. Additionally, as shown by the results of prompts 4-5, due to the inherent translation issues of polysemy, the English model cannot generate the correct target, while BDM can avoid the impact of translation.

Moreover, it's worth mentioning that the image quality of the generated images by RealisticVisionV51 is significantly

{马修·麦康纳,克里斯蒂安·贝尔,亨利·卡维尔,
斯嘉丽·约翰逊,比莉·艾利什,朱迪·科默}穿中式婚礼礼服
{Matthew McConaughey,Christian Bale,Henry Cavill,
Scarlett Johansson,Billie Eilish,Jodie Comer} wears a Chinese wedding dress

Figure 9: Dreambooth.



一个美丽女人身穿相声马褂在坐着学习
A beautiful woman wearing a cross
talk jacket is sitting and studying

Figure 10: Textual Inversion.

better than that of Stable Diffusion 1.5, despite BDM being trained on the latter. This observation emphasizes the idea that as the performance of the English backbone improves, BDM also sees improvements. Consequently, options such as refining the backbone or incorporating a more advanced English backbone can be considered to further enhance BDM's capabilities.

Checkpoint. In Fig.6, we present BDM's capability in effectively integrating with various English communities checkpoints. The backbone is alternately set to realisticVisionV51(SG_161222 2023)(realV51 for short),flat2DAnimerge(bigbeanboiler 2023)(flat2D for short) and midjourneyPapercut(shadowxshinigami 2023)(Papercut for short). By using the same native language text as input across all configurations, it becomes evident that BDM, with various English backbones, can generate images that are not only semantically consistent within the native language context but also accurately reflect backbone's unique styles.

LoRA. In Fig.7, we demonstrate BDM's ability to smoothly integrate with LoRA(Hu et al. 2021) within English communities. We select three different variants of LoRA, namely 3DMM(LONGD 2023), blindbox(samecorner 2023), and animeoutline(CyberAIchemist 2023), and find that BDM, in combination with each variant, is capable of generating Chinese semantic images that align with the respective styles of LoRA.

ControlNet. In Fig.8, We showcase BDM's capability to work with ControlNet(Zhang, Rao, and Agrawala 2023). We select (Zhang 2023) as the control model and use depth maps to generate controlled images. The images produced feature individuals of East Asian descent, aligning with the distribution of BDM's training data, and their attire distinctly reflects Chinese elements.

Dreambooth. In Fig.9, we demonstrate BDM's successful integration with Dreambooth(Ruiz et al. 2023). We chose Dreambooth models representing six well-known figures from the Famous People checkpoint(malcolmrey 2023) and then produce images depicting these figures in traditional Chinese wedding attire. Clearly, the generated images accurately portray the chosen individuals, seamlessly incorporating relevant Chinese cultural elements.

桌子上的鼠标mouse on table

Figure 11: Ablation Study.

Textual Inversion. In Fig.10, we showcase BDM's ability to work with Textual Inversion(Gal et al. 2023). We select English Textual Inversion embeddings capable of adjusting age, named Age Slider(Zovya 2023), and carry out image generation using the same random seed and identical native language text descriptions. It is clear that within the native language context, the age of the depicted individuals can be varied by adjusting the embeddings.

Ablation Study

BDM's structure can be considered as a diffusion UNet with two different language encoders and one shared decoder. As mentioned in , by controlling text prompt input to different language branch, we can generate images emphasizing different language semantics. To further study BDM's capability in capturing different language's semantics, we conduct experiments by weighting the Chinese language branch, similar to the experiments on weighting the branches in ControlNet(Zhang, Rao, and Agrawala 2023) and LoRA(Hu et al. 2021). The weight increases from 0 to 1 with step 0.2. The Chinese language branch receives Chinese prompt meanwhile the English backbone receives corresponding English prompt (translated from Chinese prompt) as text input. Random seed is kept fixed during the experiment. We use realisticVisionV51(SG_161222 2023) as the backbone model.

To amplify the study effect, we deliberately choose a case where the translated English prompt contains polysemous word "mouse". The Chinese prompt means "computer mouse on table" whereas the translated English prompt "mouse on table" can be considered as an animal mouse. As shown in Fig.11, as the weight increases for BDM's Chinese language branch, the generated image gradually changes from an animal mouse to a computer mouse, aligning more closely with the Chinese semantics from BDM's Chinese language branch.

Conclusions

We introduce the "Bridge Diffusion Model" (BDM) as a new structure to develop Chinese native TTI model which keeps compatibility with the English TTI communities. The English backbone within BDM is kept frozen thus maintains compatibility with its English ancestor model, meanwhile the non-English language branch is responsible for expressing native language meanings. Experiments show BDM can produce high-quality images from Chinese prompts and easily integrate with English TTI plugins.

References

AI, S. 2024. SD3.5. https://github.com/Stability-AI/sd3.5. 2024-12-15.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

bigbeanboiler. 2023. https://civitai.com/models/35960/flat-2d-animerge.

Blau, T.; Ganz, R.; Kawar, B.; Bronstein, A. M.; and Elad, M. 2022. Threat Model-Agnostic Adversarial Defense using Diffusion Models. *CoRR*, abs/2207.08089.

Chen, Z.; Liu, G.; Zhang, B.; Yang, Q.; and Wu, L. 2023. AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, 8666–8682. Association for Computational Linguistics.

Cheng, B.; Ma, Y.; Wu, L.; Liu, S.; Ma, A.; Wu, X.; Leng, D.; and Yin, Y. 2024. HiCo: Hierarchical Controllable Diffusion Model for Layout-to-image Generation. *CoRR*, abs/2410.14324.

civitai. 2023. https://civitai.com/.

CyberAIchemist. 2023. https://civitai.com/models/16014/anime-lineart-manga-like-style.

Dhariwal, P.; and Nichol, A. 2021. Diffusion Models Beat GANs on Image Synthesis. *Neural Information Processing Systems,Neural Information Processing Systems*.

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.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Gao, J.; Zhang, J.; Liu, X.; Darrell, T.; Shelhamer, E.; and Wang, D. 2022. Back to the Source: Diffusion-Driven Test-Time Adaptation. *CoRR*, abs/2207.03442.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2017. Generative Adversarial Nets. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 177–177.

Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35: 26418–26431.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Neural Information Processing Systems,Neural Information Processing Systems*.

Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv: Computation and Language,arXiv: Computation and Language*.

Kawar, B.; Elad, M.; Ermon, S.; and Song, J. 2022a. Denoising Diffusion Restoration Models. In *NeurIPS*.

Kawar, B.; Ganz, R.; and Elad, M. 2022. Enhancing diffusion-based image synthesis with robust classifier guidance. *arXiv preprint arXiv:2208.08664*.

Kawar, B.; Song, J.; Ermon, S.; and Elad, M. 2022b. JPEG Artifact Correction using Denoising Diffusion Restoration Models. *CoRR*, abs/2209.11888.

Labs, B. F. 2024. FLUX. https://blackforestlabs.ai/. 2024-12-15.

Li, Y.; Chang, C.; Rawls, S.; Vulic, I.; and Korhonen, A. 2023. Translation-Enhanced Multilingual Text-to-Image Generation. In Rogers, A.; Boyd-Graber, J. L.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, 9174–9193. Association for Computational Linguistics.

LONGD. 2023. https://civitai.com/models/73756/3d-rendering-style.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Luccioni, A. S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. *CoRR*, abs/2303.11408.

malcolmrey. 2023. https://civitai.com/models/59622/famous-people.

Miller, E. J.; Steward, B. A.; Witkower, Z.; Sutherland, C. A.; Krumhuber, E. G.; and Dawel, A. 2023. AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*, 09567976231207095.

Mou, C.; Wang, X.; Xie, L.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 16784–16804. PMLR.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 16805–16827. PMLR.

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. *Cornell University - arXiv,Cornell University - arXiv*.

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 Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR*, abs/2204.06125.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. *International Conference on Machine Learning,International Conference on Machine Learning*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-Image Diffusion Models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*.

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. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

samecorner. 2023. https://civitai.com/models/25995/blindbox.

Sasaki, H.; Willcocks, C.; and Breckon, T. 2021. UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models. *Cornell University - arXiv,Cornell University - arXiv*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

SG_161222. 2023. https://civitai.com/models/4201/realistic-vision-v51.

shadowxshinigami. 2023. https://civitai.com/models/80/midjourney-papercut.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In Bach, F. R.; and Blei, D. M., eds., *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, 2256–2265. JMLR.org.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. *Cornell University - arXiv,Cornell University - arXiv*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. *Cornell University - arXiv,Cornell University - arXiv*.

stablediffusionweb. 2023. https://stablediffusionweb.com/.

Theis, L.; Salimans, T.; Hoffman, M. D.; and Mentzer, F. 2022. Lossy Compression with Gaussian Diffusion. *CoRR*, abs/2206.08889.

Vahdat, A.; Kreis, K.; and Kautz, J. 2021. Score-based Generative Modeling in Latent Space. *Neural Information Processing Systems,Neural Information Processing Systems*.

vegaai. 2023. https://www.vegaai.net/.

Vincent, P. 2011. A connection between score matching and denoising autoencoders. *Neural Computation*, 1661–1674.

Xue, Z.; Song, G.; Guo, Q.; Liu, B.; Zong, Z.; Liu, Y.; and Luo, P. 2023. RAPHAEL: Text-to-Image Generation via Large Mixture of Diffusion Paths. *CoRR*, abs/2305.18295.

Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, Y.; Zhou, J.; and Zhou, C. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; Hutchinson, B.; Han, W.; Parekh, Z.; Li, X.; Zhang, H.; Baldridge, J.; and Wu, Y. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *Trans. Mach. Learn. Res.*, 2022.

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. *CoRR*, abs/2112.15283.

Zhang, J.; Gan, R.; 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.; and Chen, C. 2022. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *CoRR*, abs/2209.02970.

Zhang, L. 2023. control_v11f1p_sd15_depth, https://huggingface.co/lllyasviel/control_v11f1p_sd15_depth.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zimmermann, R. S.; Schott, L.; Song, Y.; Dunn, B. A.; and Klindt, D. A. 2021. Score-Based Generative Classifiers. *CoRR*, abs/2110.00473.

Zovya. 2023. https://civitai.com/models/65214/age-slider.