AnyTrans: Translate AnyText in the Image with Large Scale Models

Anonymous ACL submission

Abstract

This paper introduces AnyTrans, an all-encompassing framework for the task–Translate AnyText in the Image (TATI), which includes multilingual text translation and text fusion within images. Our framework leverages the strengths of large-scale models, such as Large Language Models (LLMs) and text-guided diffusion models, to incorporate contextual cues from both textual and visual elements during translation. The few-shot learning capability of LLMs allows for the translation of fragmented texts by considering the overall context. Meanwhile, the advanced inpainting and editing abilities of diffusion models make it possible to fuse translated text seamlessly into the original image while preserving its style and realism. Additionally, our framework can be constructed entirely using open-source models and requires no training, making it highly accessible and easily expandable. To encourage advancement in the TATI task, we have meticulously compiled a test dataset called MTIT6, which consists of multilingual text image translation data from six language pairs.

1 Introduction

Translating AnyText in the Image (TATI) has become an essential tool in our daily lives, transforming how we interact with the world. This capability extends to a wide range of applications, from facilitating cross-cultural communication to supporting education, and playing a significant role in global business operations. Falling under the umbrella of multi-modal machine translation (MMT) (Elliott et al., 2016; Calixto et al., 2017; Elliott and Kádár, 2017; Libovický et al., 2018; Sulubacak et al., 2019), the process of translating text in images is commonly known as Text Image Translation (TIT) (Ma et al., 2022; Lan et al., 2023). TIT aims to accurately convert text in source images into desired target languages.

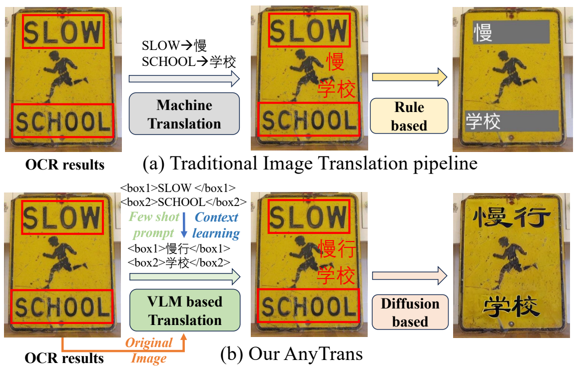

Figure 1: Comparison between traditional image translation pipeline and our AnyTrans. Our AnyTrans combines image information and context for more accurate translation and generates more realistic text.

However, we argue that translated text alone is insufficient. A seamless integration of text and image is crucial for effectively conveying the intended message. Thus, we believe that our proposed task, Translate AnyText in the Image (TATI), better aligns with practical needs. It not only aims to translate textual content within an image but also maintains the visual coherence and intrinsic harmony of text and graphic elements, thereby enhancing the overall comprehensibility of texts in images.

Current popular products, such as Google Image Translation (Translate, b), Microsoft Image Translation (Translate, c), and Apple iOS Image Translation (Translate, a), have made significant progress in translating text within images. However, as illustrated in Figure 1 (a), Microsoft Image Translation, for instance, utilizes traditional machine translation to translate text recognized by OCR models. It then employs a simple rule to insert the translated text back into the original image. Unfortunately, this approach often overlooks the contextual relationship between textual elements within images. This oversight can result in inaccurate translations and visual inconsistencies, thereby compromising the authenticity of the newly generated image.

To address the identified shortcomings, our framework illustrated in Figure 1 (b) significantly diverges from conventional text translation tasks in images. By leveraging the advanced contextual comprehension capabilities of LLMs, our approach achieves superior translation accuracy. Alternatively, the integration of a vision language model (VLM) may allow a dual consideration of both visual and textual contexts within the source images, further enhancing translation quality.

Our methodology unfolds in three consecutive steps. Initially, we utilize the latest PP-OCR (Du et al., 2020) to accurately locate the text within the image and decipher its content. This step is crucial for determining the exact area for text editing and translating the text content precisely. Secondly, once the text is identified, we employ a few-shot prompt learning strategy that enables (visual) language models to maintain the format during contextual translation. This approach ensures that the translation is both contextually appropriate and linguistically accurate. Finally, we apply a modified AnyText (Tuo et al., 2023) to render the translated text back into the original image. In this phase, the translated text is fused into its original location, identified during the initial step. We propose resizing the anticipated text box by considering the length of the detected box, the original source text, and the translated target text. This modification maximizes the preservation of the original image's style and produces a clean, new image. As shown in Figure 1 (b), our method does achieve superior translation quality and visual effects while preserving the image's legibility and aesthetic appeal. The new text seamlessly blends with the original visual context, maintaining both coherence and style.

Our main contributions are as follows:

(1) We present an integrated framework for the task–Translate AnyText in the Image (TATI), consisting of three key steps: source text detection and recognition, text image translation, and target text fusion.

(2) Our method is training-free and can be built entirely on open-source models, yet it delivers results that are comparable to or even surpass those of commercial, proprietary products.

(3) We constructed a multilingual text image translation test dataset called MTIT6, which consists of translation data in six language pairs and is manually sequenced by humans, promoting the field of image translation.

2 Related Works

2.1 Text Image Translation and Multilingual Translation

The field of multimodal machine translation (MMT) (Caglayan et al., 2016; Huang et al., 2016; Libovický and Helcl, 2017; Calixto et al., 2017; Su et al., 2021) has witnessed remarkable advancements in recent years, catalyzing a surge in scholarly and industry interest. The prevailing practical demand for MMT is the translation of text within images, known as text image translation (TIT) (Ma et al., 2022; Mansimov et al., 2020; Jain et al.; Lan et al., 2023). However, TIT leaves the image unchanged, while integrating text translations directly into images is essential for helping users understand the meaning of both text and visuals. Taking these factors into account, we believe that our proposed TATI task is more aligned with practical requirements.

Meanwhile, Large Language Models (LLMs) (Gao et al., 2024; Vilar et al., 2022; Zeng et al., 2023; Wu et al., 2021) have shown impressive multilingual translation proficiency. Integrating multilingual translation (Dong et al., 2015; Firat et al., 2016; Neubig and Hu, 2018; Chen et al., 2017, 2022; Cheng, 2019) with image-to-image translation opens vast opportunities and has wide-ranging applications, such as in cross-border e-commerce platforms, among others.

2.2 Text Editing in Images

Recent advancements in image processing have seen a burgeoning interest in text editing (Yang et al., 2018b; Wu et al., 2019; He et al., 2023; Zhu et al.; Ma et al., 2023; Chen et al., a, 2023; Couairon et al., 2022; Tuo et al., 2023) within images. Numerous methods leveraging Generative Adversarial Networks (GANs) have emerged for scene text editing, aiming to transform the text within a scene image to a specified target while retaining the authentic style. Despite their innovations, GAN-based approaches (Wu et al., 2019; Goodfellow et al., 2017; Mirza and Osindero, 2014; Zhu et al., 2017; Yang et al., 2018a; Azadi et al., 2018) struggle to edit images featuring intricate scenes or a multitude of diverse elements. The recent development of diffusion models (Saharia et al., 2022; Rombach et al., 2022; Chung et al., 2022; Zhang et al., 2023a; Nichol et al.; Avrahami et al., 2022; Yang et al., 2022; Zhang et al., 2023b; Mou et al., 2023) allows for the generation of images of exceptional quality and diversity.

Frameworks such as ControlNet (Zhang et al., 2023b) and T2IAdapter (Mou et al., 2023) have harnessed auxiliary cues like color maps, and segmentation maps to steer the image generation process, achieving remarkable levels of control and image quality. Galvanized by these advances, a series of text-centric image editing techniques (Zhu et al.; Ma et al., 2023; Chen et al., a, 2023; Couairon et al., 2022; Tuo et al., 2023) have been introduced based on diffusion models. Among these, AnyText (Tuo et al., 2023) stands out for its proficient multilingual text editing capabilities, producing impressive results in text rendering and manipulation. The advancements of these technologies seamlessly enable the realization of TATI task, facilitating a more intuitive and efficient process.

3 Methodology

In this section, we will detail each component of our AnyTrans. Following the module order shown in Figure 3, we begin by introducing the detection and recognition of text in the image. Following this, we introduce how to leverage (vision) LLMs for translation. Lastly, we describe the text editing process informed by the translation outcomes.

3.1 Text Detection and Recognition

As illustrated in the *Text Detection & Recognition* section of Figure 3, to accomplish our image-to-image translation task, we first need to detect the position of the text in the image and recognize its content. Essentially, this procedure involves text detection (He et al., 2021; Liao et al., 2020; Lyu et al., 2018; Ma et al., 2018; Zhou et al., 2017) and recognition (Bautista and Atienza, 2022; Li et al., 2021; Shi et al., 2017; Chen et al., b; Yu et al., 2023), which embodies a classic OCR endeavour. Although VLM also has a certain degree of OCR capability, its capability lags far behind traditional OCR models (Liu et al., 2023). So we harness the capabilities of the pre-trained OCR model, which excels in both text detection and recognition. Subsequently, the outcomes of OCR are fed into subsequent modules for translation and text editing.

3.2 Beyond Box-level Text Translation

Building on the recognition outcomes obtained from the OCR module, our next step involves translating the textual content into the desired target language. It is important to note that the OCR system processes and retrieves text content sequentially, which means the extracted sequence may not always reflect the true semantic order. This presents significant challenges for traditional translation models, which often struggle to accurately interpret the broader context and semantic connections between individual text segments. For instance, as illustrated in Figure 1 (a), the word “SLOW” in an image should convey the meaning “slow down for passing students”. However, traditional translation pipelines only translate the text within each isolated box, failing to grasp the context and leading to poor translations.

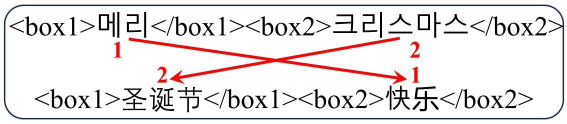

Figure 2: A prompt example from Korean to Chinese. In Chinese, the order of the two words should be switched.

Fortunately, the landscape of translation has undergone a seismic shift with the emergence of large language models (LLMs), which exhibit a markedly enhanced ability to understand context and generate coherent translations. With their powerful multilingual and instruction-following capabilities, LLMs can be seamlessly integrated into our multilingual image translation framework without additional training. By employing a few-shot prompt strategy, we can enable the translation of multiple text segments in a more coherent manner.

Therefore, we integrated the LLM into the core of our proposed framework. Particularly, as shown in Figure 2, for texts within an image identified by OCR, we concatenate them into a long text sequence using HTML-style tags <boxidx></boxidx> to retain the positional information of the detected text. The translated sentence should be organized in the same format, but with the word order adjusted accordingly. In practice, we use five-shot demonstrations for each language pair in the instruction prompt to help the LLM understand our designed format.

Additionally, while multiple translation options may exist for a given text, the entire text sequences alone may not fully disambiguate the meaning. Therefore, incorporating visual information from images is also crucial. To address this, we have explored the supportive role of using a vision LLM in text translation. This method leverages the comprehensive visual information contained in images to refine the quality of the translation.

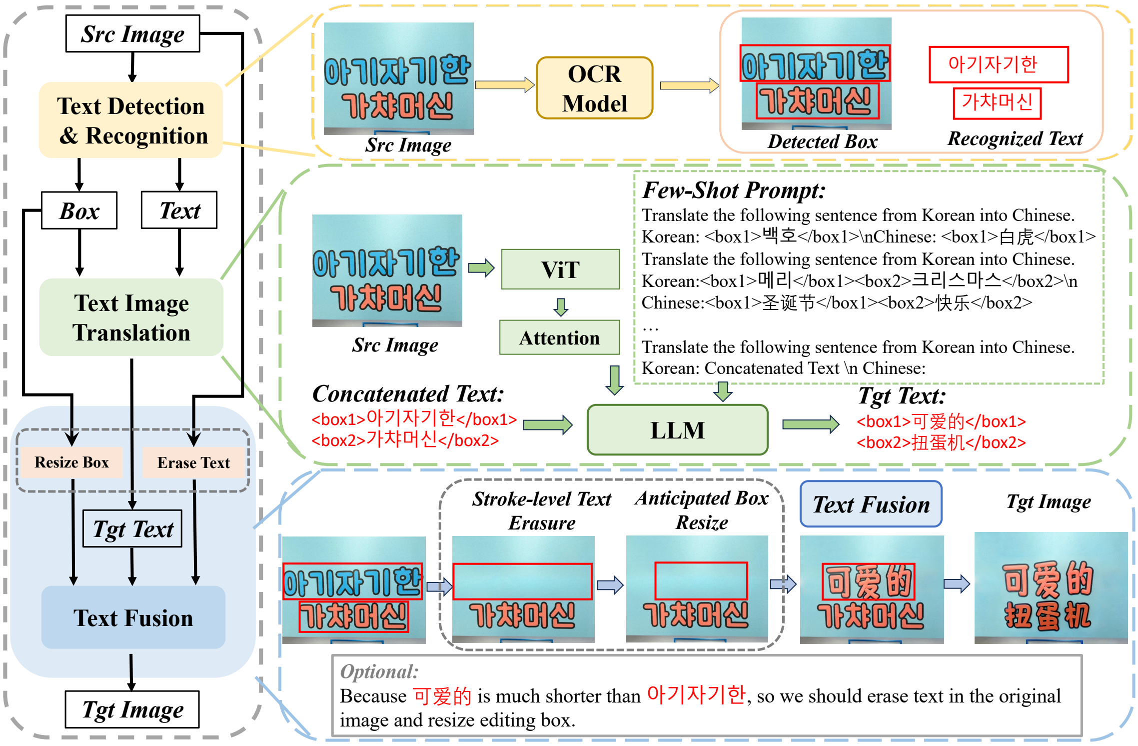

Figure 3: An overview of AnyTrans. Our translation framework is built around three key components: firstly, Text Detection and Recognition utilizing an offline OCR model; secondly, Text Image Translation using (vision) LLMs; and finally, Text Fusion using the modified AnyText.

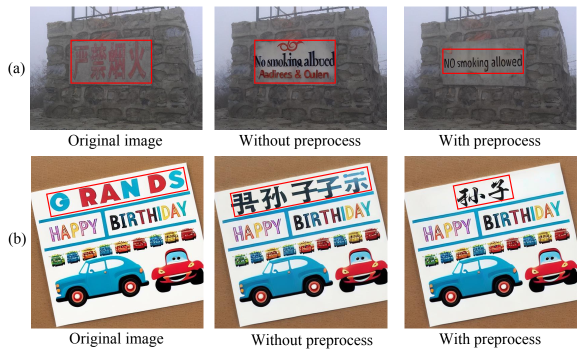

Figure 4: Preprocessing for AnyText is crucial for producing accurate and authentic text, especially in scenarios where there is a significant disparity in text length before and after translation.

3.3 Text Fusion in Image

The final module in our framework involves generating a new image with the translated texts. To achieve a cohesive visual effect, we propose integrating the translated texts into the original image, placing them precisely where the original text appeared. This ensures that the translated text not only communicates the intended message but also harmonizes with the visual context of the image.

Traditional rule-based algorithms for fusing text into images exhibit several significant drawbacks, including compromising the integrity of the image background, limiting outputs to a singular font style, and resulting in a final appearance that often lacks realism. Instead, we adopt the technique of diffusion model, which enables natural text editing within images. Specifically, for our text editing process, we propose a multilingual text editing method built on Anytext (Tuo et al., 2023).

In the original Anytext, the areas designated for editing are the detection boxes identified by OCR, and the input text is the translated sentence. However, Anytext is particularly sensitive to the length of the input text designated for rendering. As shown in Figure 4, the quality of the generated text is significantly impacted by the length ratio between the detected box and the input text. When this ratio deviates too far from 1, the vacant area tends to be filled with irrelevant content, significantly compromising both the visual effect and the translation quality.

Stroke-level Text Erasure To address this issue, as illustrated in the *Text Fusion* section of Figure 3, we first apply stroke-level text erasure (Li et al., 2023). Unlike the end-to-end text editing approach used in Anytext, we decompose the process into two sub-steps. The first step involves applying a fine-grained inpainting method specifically designed to remove the strokes of characters or letters in the original texts. This method can successfully remove multi-line texts with minimal line spacing, resulting in a cleaner visual effect.

Anticipated Box Resize To address the length ratio issue and further avoid conflicts between adjacent lines, we propose an OCR box resizing preprocessing step for the anticipated target box. Specifically, if the word count ratio between the pre and post-translation text exceeds 1.2 or is less than 0.8, we will adjust the length or width of the anticipated box based on the ratio. This process requires some customization depending on the language pair. For example, in zh-en translations, we assume the length of a Chinese character to be 2.5 times that of an English letter, given the fact that larger size for a single Chinese character. In the end, the fusion of target text is applied to the erased area.

4 Experiments

4.1 Dataset

Image	Locations	Source Texts	Target Translations	Order
	(121.0, 185.0), (380.0, 158.0), (384.0, 203.0), (126.0, 231.0)	NEW MEXICO	新墨西哥	1;3,2
	(106.0, 228.0), (342.0, 221.0), (344.0, 271.0), (108.0, 278.0)	LAND OF	之地	新墨西哥;魅力之地
	(110.0, 278.0), (419.0, 287.0), (417.0, 341.0), (108.0, 332.0)	ENCHAN TMENT	魅力	

Figure 5: An example of our MTIT6 dataset, which contains position information of the text in the image, corresponding translation information, and corrected translation order.

We present MTIT6, a comprehensive multilingual text image translation test dataset, assembled from ICAR 19-MLT(Nayef et al., 2019), OCRMT30K(Lan et al., 2023), along with a selection of high-quality images curated by our team. Our dataset encompasses six language pairs: English-to-Chinese, Japanese-to-Chinese, Korean-to-Chinese, Chinese-to-English, Chinese-to-Japanese and Chinese-to-Korean, each pair features about 200 images. In creating this dataset, we employed the lightweight PP-OCR tool for initial OCR recognition, and then the OCR outputs were further refined and translated by language experts. Furthermore, considering differences in word order across different languages, our language experts meticulously annotated the sequences of text identified by OCR within each image. This approach enabled us to maintain semantic integrity by rearranging the text into coherent sequences, based on their annotated order. Figure 5 presents an example of our MTIT6 dataset.

4.2 Comparison Results

4.2.1 Quantitative Results

For evaluation, we choose the **BLEU** (Papineni et al., 2001) and **COMET** (Rei et al., 2020) metrics. We evaluate the image-to-text (**I2T**) intermediate translation results and image-to-image (**I2I**) final translation results. We have integrated a wide range of models into our AnyTrans, which included classic encoder-decoder models (Costa-jussà et al., 2022; Fan et al., 2021), widely accessible open-source LLMs (qwen-chat1.5-7B,14B and 110B), and commercially advanced close-source LLM (qwen-max) and VLM (Bai et al., 2023) (qwen-vl-max), affirming our approach's comprehensive reliability and easy scalability. We also validate the model(Lan et al., 2023) specifically designed for the TIT task in our test dataset. To more accurately evaluate the translation quality of the final image, we use the paid BaiduOCR[1] to recognize the text in the **I2I** stage.

As shown in Table 1 and Table 2, we observed that the performance of the qwen-1.5 series models gradually improved with the increase of the model's parameters. We discovered that the enhancement in performance is attributed not only to the improved quality of translations but also to the bolstered ability to follow instructions. This is particularly evident in the 7B model, which initially exhibited a weaker capacity for instruction adherence. During qwen-7B model's translation process, there is around a 10% chance that the `<boxidx></boxidx>` symbol, employed to demarcate positions, might be inaccurately translated. Another interesting finding is that the performance of qwen1.5-110B is very close to or even exceeds qwen-max in multiple language pairs. This may be because the qwen1.5 series used more new high-quality corpora and adopted technologies such as DPO(Rafailov et al.) and PPO(Schulman et al., 2017) during training. The results demonstrate that while enlarging the model's parameters significantly boosts its capability to adhere to instructions, honing the model's translation skills may rely more heavily on the quality of the corpus and the refinement of training methodologies. Moreover, VLMs improved translation performance, indicating that integrating image data can further augment translation accuracy. This advancement confirms that VLMs represent a key developmental trajectory for future research endeavours in image translation.

[1] https://cloud.baidu.com/product/ocr/general

Methods	zh→ en			zh→ ko			zh→ ja		
	I2T		I2I	I2T		I2I	I2T		I2I
	BLEU	COMET	BLEU	BLEU	COMET	BLEU	BLEU	COMET	BLEU
nllb-200(3.3B)	29.7	66.3	22.2	20.1	72.5	11.4	24.9	77.20	13.1
m2m100(1.2B)	33.1	66.1	23.8	18.1	71.1	10.9	29.4	79.60	14.8
mc-tit	41.6	70.5							
qwen1.5-7B-chat	37.4	73.4	26.5	11.4	70.4	5.5	31.2	80.9	20.6
qwen1.5-14B-chat	38.8	74.6	28.0	16.1	72.9	8.3	30.7	79.3	19.8
qwen1.5-110B-chat	43.8	76.3	30.6	17.1	74.3	9.3	**35.4**	**83.1**	**21.9**
qwen-max	44.0	77.2	31.2	23.5	75.1	15.1	33.5	81.3	20.9
qwen-vl-max	**48.7**	**78.0**	**31.9**	**25.0**	**75.3**	**15.8**	34.2	81.9	21.4

Table 1: Experiments on multilingual TATI task encompass translating Chinese into English, Korean, and Japanese.

Methods	en→ zh			ko→ zh			ja→ zh		
	I2T		I2I	I2T		I2I	I2T		I2I
	BLEU	COMET	BLEU	BLEU	COMET	BLEU	BLEU	COMET	BLEU
nllb-200(3.3B)	21.5	73.3	15.1	9.1	65.3	8.7	7.4	61.3	7.2
m2m100(1.2B)	24.2	76.9	18.9	14.8	67.8	13.1	24.3	74.5	22.7
qwen1.5-7B-chat	27.6	80.7	21.4	20.9	75.72	18.2	30.0	78.7	27.5
qwen1.5-14B-chat	34.5	81.3	26.8	27.7	77.8	23.6	38.4	81.3	28.6
qwen1.5-110B-chat	**37.9**	84.2	27.0	32.6	80.5	31.4	38.2	80.7	30.9
qwen-max	34.7	84.1	24.1	33.1	81.0	29.8	32.2	80.4	27.1
qwen-vl-max	36.3	**84.3**	**27.8**	**35.4**	**81.7**	**31.6**	**54.2**	**83.8**	**44.3**

Table 2: Experiments on multilingual TATI tasks encompass translating English, Korean, and Japanese into Chinese.

4.2.2 Qualitative Results

To the best of our knowledge, this is the first paper to research the task of TATI, so there is no open-source model to compare with, so we can only compare with commercial closed-source image translation products, including Google Image Translation (Translate, b), Microsoft Image Translation (Translate, c) and Apple IOS Image Translation (Translate, a). As shown in the cases in Figure 8, Microsoft and Apple Image Translation generate translations in rectangular areas based on rules and then paste them back to the original image. However, these rectangular areas' colours fail to match those of the original image. Consequently, directly integrating the text from these areas into the original image significantly disrupts its visual harmony. Google Image Translation exhibits some improvement. It first erases the original text and then returns the translated text to the original image. However, this process leaves noticeable erasure marks, and the text, being rule-based, appears overly uniform and fails to harmonize with the original image's aesthetics. In contrast, our AnyTrans seamlessly integrates the translated text into the original image and even manages to preserve the font colour and style to a notable degree. Therefore, it is clear that our AnyTrans significantly surpasses image translation products in maintaining visual continuity.

4.2.3 Human and GPT Evaluation

To evaluate the authenticity and style consistency of translated images, we randomly selected 50 images from six language pairs, totalling 300 images. We then assessed the translation results from Google Image Translation, Microsoft Image Translation, Apple Image Translation, and AnyTrans. Each image was scored based on our evaluation criteria by three assessors and GPT4o, and the detailed evaluation criteria can be found in the appendix. As shown in Figure 6, whether it is the human evaluation or GPT4o automatic evaluation, our method significantly outperforms Microsoft and Apple Image Translation in terms of authenticity and style consistency and achieves comparable scores to Google. We also verify the correlation between GPT4o evaluation results and human preference scores in Figure 7. By calculating Spearman's correlation coefficient for each language pair, we observe a strong correlation between the two evaluation methods. The consistency further demonstrates the superiority of our approach.

Upon analyzing the cases with lower scores than Google, we found most instances are due to the limited performance of AnyTrans in generating text on small fonts. In contrast, Google Image Translation, being based on rule-based generation of text, has a clear advantage in translating texts of small font sizes. Nevertheless, based on the

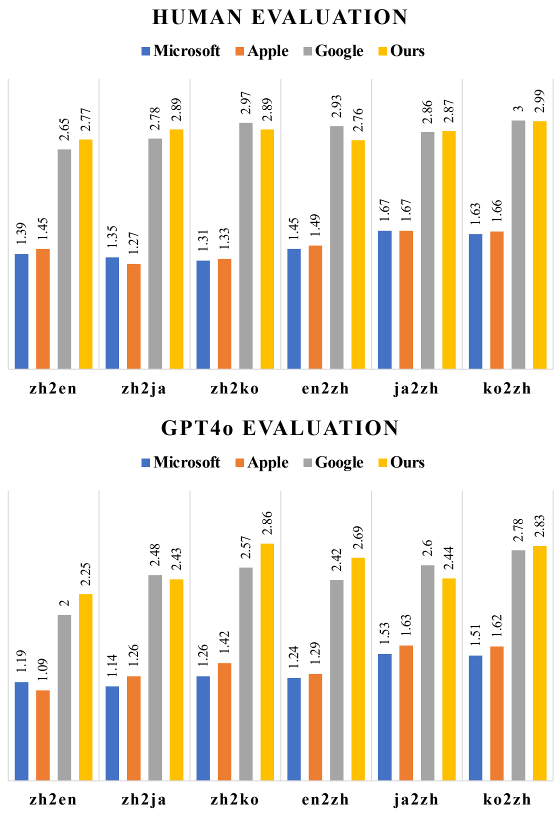

Figure 6: Overall human evaluation and GPT4o results of image translation performance for different methods. Our method significantly outperforms Microsoft and Apple and achieves comparable results to Google.

Methods	Average	
	BLEU	COMET
qwen1.5-7B-chat(box)	25.9	75.7
qwen1.5-7B-chat(context)	26.5	76.3
qwen1.5-14B-chat(box)	30.6	76.9
qwen1.5-14B-chat(context)	31.0	77.9
qwen1.5-110B-chat(box)	32.2	78.1
qwen1.5-110B-chat(context)	**33.2**	**79.1**

Table 3: Ablation experiments on translation strategies and model categories on multilingual TIT tasks.

advantages of authenticity and style consistency, our AnyTrans still achieved scores comparable to Google Image Translation.

4.3 Ablation Study

We performed detailed ablation studies to explore the efficacy of two translation strategies: translating the contents within detection boxes individually versus translating all recognized text in the image as a whole. Specifically, for the latter translation method, we concatenate recognized texts from an entire image using ‘<boxidx></boxidx>‘ tags. These are then merged with few-shot prompts into a lengthy sentence, which is subsequently inputted into LLMs for translation. We tested on the qwen1.5-7B, 14B and 110B models and calculated the average of the test results for all language pairs. As depicted in Table 3, our strategy of translation as a whole significantly improves translation performance across all three parameter sizes of qwen1.5 models. This enhancement underscores the importance of LLM's advanced contextual understanding in boosting translation performance. We also conducted an ablation experiment on the resize editing area strategy. As shown in Table 4, in the zh2en translation, without the OCR box resizing step, the final I2I translation result dropped by 2.9 points, proving the effectiveness of this strategy.

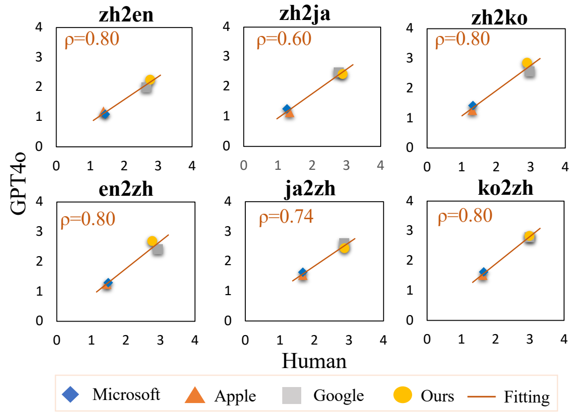

Figure 7: Our experiments show that GPT4o evaluations across all language pairs closely match human perceptions. In each plot, a dot represents the human preference evaluation score (horizontal axis) and GPT4o evaluation score (vertical axis). We linearly fit a straight line to visualize the correlation and calculate Spearman's correlation coefficient (ρ) for each language pair.

Methods	zh→ en		
	I2T		I2I
	BLEU	COMET	BLEU
qwen1.5-110B-chat	43.8	76.27	**30.6**
Wo-resize	43.8	76.27	27.7(**-2.9**)

Table 4: Ablation experiment on resizing editing area.

5 Discussions

As the first paper to introduce (vision) LLMs and diffusion model into the Translate AnyText in the Image (TATI) task, significant opportunities exist for further improvement. Below, we enumerate several potential directions for future advancements:

(1) Integration of OCR and Translation Processes: Our current methodology bifurcates the process into OCR text recognition and translation as distinct steps. While VLMs currently fail to achieve the OCR accuracy of smaller models tailor-made

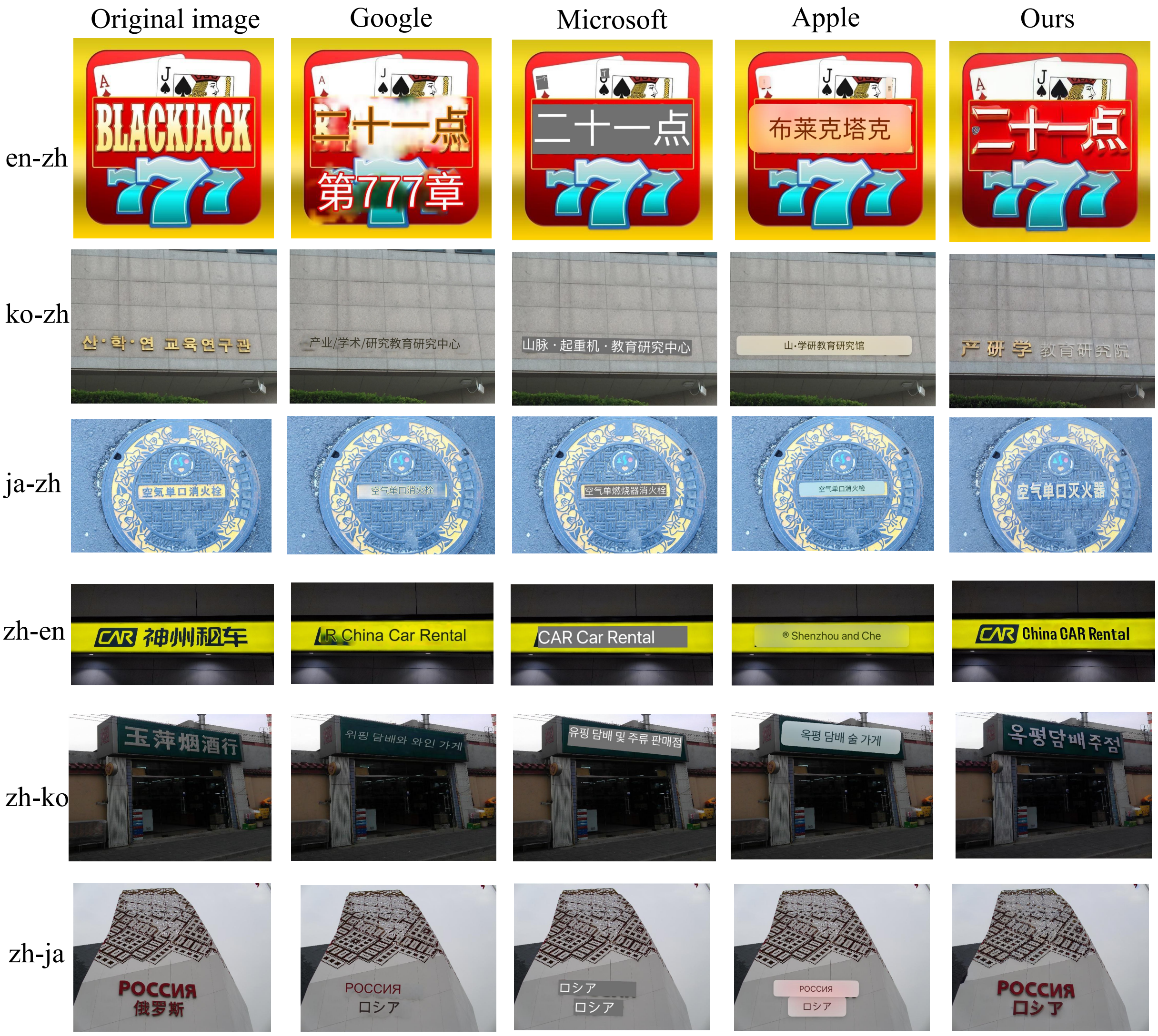

Figure 8: Qualitative comparison of our framework with Google, Microsoft and Apple Image Translation results. Our AnyTrans has obvious advantages in font style preservation and authenticity.

for OCR tasks, further development and OCR-targeted training could potentially elevate VLMs to achieve formidable OCR prowess. This evolution could potentially consolidate text recognition and translation into a seamless, singular step, enhancing efficiency and accuracy.

(2) Text editing model adapted to translation: Due to AnyText (Tuo et al., 2023) being trained on datasets where character size perfectly matches the image size, it needs the text length to be well-matched with the dimensions of the editing area. However, when translating, the length of the translated text inevitably varies across different languages, leading to challenges for Anytext to generate translations that fit the original text area perfectly. The Anticipated Box Resizment strategy helps mitigate the issue but does not fully resolve it. Future efforts could focus on training a text editing model capable of dynamically adjusting font sizes. This would eliminate the necessity for altering the editing area, allowing for modifications that preserve the aesthetic appeal and structural harmony of the original image more faithfully.

6 Conclusion

We introduce a novel framework named AnyTrans designed for Translate AnyText in the Image (TATI). Distinguished from existing closed-source products, our AnyTrans can be built upon open-source models and is training-free. Uniquely, we integrate (vision) LLMs and diffusion models into TATI task for the first time, achieving both accurate translations and authentic translated images. Furthermore, we have curated a multilingual text image translation dataset MTIT6 to promote development in this field.

7 Limitations

(1) Owing to inherent restrictions in Anytext (Tuo et al., 2023), it is unable to produce outputs exceeding 20 letters or characters at a time. Consequently, this limitation extends to our AnyTrans, affecting its ability to effectively translate longer texts.

(2) Given that Anytext's text editing proficiency is confined to Chinese, English, Korean, and Japanese, it lacks the capability to generate text in other languages, such as Arabic. As a result, the range of languages that AnyTrans is capable of translating is similarly restricted.

References

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Samaneh Azadi, Matthew Fisher, Vladimir Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. 2018. Multi-content gan for few-shot font style transfer. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *Cornell University - arXiv,Cornell University - arXiv*.

Iacer Calixto, Qun Li, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *Cornell University - arXiv,Cornell University - arXiv*.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023. Textdiffuser-2: Unleashing the power of language models for text rendering.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, Furu Wei, and Hkust Hkust. a. Textdiffuser: Diffusion models as text painters.

Jingye Chen, Haiyang Yu, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, Bin Li, and Xiangyang Xue. b. Benchmarking chinese text recognition: Datasets, baselines, and an empirical study.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Yun Chen, Yang Liu, and Victor Li. 2022. Zero-resource neural machine translation with multi-agent communication game. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yong Cheng. 2019. *Joint Training for Pivot-Based Neural Machine Translation*, page 41–54.

Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. 2022. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. 2022. Diffedit: Diffusion-based semantic image editing with mask guidance.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. 2020. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *Cornell University - arXiv,Cornell University - arXiv*.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. *Cornell University - arXiv,Cornell University - arXiv*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings* of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024. Towards boosting many-to-many multilingual machine translation with large language models.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2017. Gangenerative adversarial nets. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, page 177–177.

Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Jingdong Sun, Wangmeng Xiang, Xianhui Lin, Xiaoyang Kang, Zengke Jin, Yusen Hu, Bin Luo, et al. 2023. Wordart designer: User-driven artistic typography synthesis using large language models. *arXiv preprint arXiv:2310.18332*.

Minghang He, Minghui Liao, Zhibo Yang, Humen Zhong, Jun Tang, Wenqing Cheng, Cong Yao, Yongpan Wang, and Xiang Bai. 2021. Most: A multi-oriented scene text detector with localization refinement. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*.

Puneet Jain, Orhan Firat, Qi Ge, and Sihang Liang. Image translation network.

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook.

Lei Li, Kai Fan, and Chun Yuan. 2023. Strokenet: Stroke assisted and hierarchical graph reasoning networks. *IEEE Transactions on Multimedia*, 25:5614–5625.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Zhang Cha, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *Cornell University - arXiv,Cornell University - arXiv*.

Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 11474–11481.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Pengyuan Lyu, Cong Yao, Wenhao Wu, Shuicheng Yan, and Xiang Bai. 2018. Multi-oriented scene text detection via corner localization and region segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Cong Ma, Yaping Zhang, Mei Tu, Xu Han, Linghui Wu, Yang Zhao, and Yu Zhou. 2022. Improving end-to-end text image translation from the auxiliary text translation task.

Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. 2023. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently.

Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, page 3111–3122.

Elman Mansimov, Mitchell Stern, Mia Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*.

Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. *Cornell University - arXiv,Cornell University - arXiv*.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. 2023. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models.

Nibal Nayef, Cheng-Lin Liu, Jean-Marc Ogier, Yash Patel, Michal Busta, ChowdhuryPinaki Nath, Karatzas Dimosthenis, Wafa Khlif, Jiri Matas, Umapada Pal, and Burie Jean-Christophe. 2019. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition – rrc-mlt-2019. *Cornell University - arXiv,Cornell University - arXiv*.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *Cornell University - arXiv,Cornell University - arXiv*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, ChristopherD Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model.

Ricardo Rei, Craig Stewart, AnaC Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv: Computation and Language,arXiv: Computation and Language*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv: Learning,arXiv: Learning*.

Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 2298–2304.

Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences*, page 47–60.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2019. Multimodal machine translation through visuals and speech. *CERN European Organization for Nuclear Research - Zenodo,CERN European Organization for Nuclear Research - Zenodo*.

Apple IOS Image Translate. a. `https://support.apple.com/zh-cn/guide/iphone/iphea8b95631/ios`.

Google Translate. b. `https://translate.google.com`.

Microsft Image Translate. c. `https://www.microsoft.com/en-us/translator/apps/`.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. 2023. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance.

Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. 2019. Editing text in the wild. In *Proceedings of the 27th ACM International Conference on Multimedia*.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. *Cornell University - arXiv,Cornell University - arXiv*.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2022. Paint by example: Exemplar-based image editing with diffusion models.

Shuai Yang, Jiaying Liu, Wenjing Wang, and Zong-Ming Guo. 2018a. Tet-gan: Text effects transfer via stylization and destylization.

Shuai Yang, Jiaying Liu, Wenhan Yang, and Zongming Guo. 2018b. Context-aware unsupervised text stylization. In *Proceedings of the 26th ACM international conference on Multimedia*.

Haiyang Yu, Xiaocong Wang, Bin Li, and Xiangyang Xue. 2023. Chinese text recognition with a pre-trained clip-like model through image-ids aligning.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison.

Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi Jaakkola, and Shiyu Chang. 2023a. Towards coherent image inpainting using denoising diffusion implicit models.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: An efficient and accurate scene text detector. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

Yun Zhu, Yaoke Wang, Haizhou Shi, Zhenshuo Zhang, Dian Jiao, and Siliang Tang. Graphcontrol: Adding conditional control to universal graph pre-trained models for graph domain transfer learning.

A Appendix

A.1 Dataset annotation details

We engaged six professional translators for a week-long annotation task, with each translator tasked to annotate 30 images daily to mitigate fatigue. Upon being presented with an image containing source texts detected by PPOCR-v4, translators were tasked to render accurate and fluid translations into the target language. Furthermore, they meticulously annotated the sequences of text recognized by OCR within each image, reordering the text to ensure coherent sequences. For quality assurance, we also employed a professional translator to sample and review the annotated instances. In total, we annotated 1,199 images, averaging around 200 instances per language pair.

A.2 Human and GPT evaluation details

We meticulously selected a sample of 50 images for each of the six languages, summing up to a total of 300 images. To objectively and accurately assess the **authenticity** of translated images along with the **maintenance of font styles**, we utilize a combination of human evaluation and GPT-4o evaluation.

For human evaluation, we enlisted the help of three annotators. For each image assessed, the annotators were provided with the original image alongside the translation outputs from Google, Microsoft, Apple Image Translations, and our AnyTrans. They then scored each translation based on predetermined criteria, with the final score for each image being the average of the three annotators' scores.

For the evaluation involving GPT-4o, to minimize biases associated with the order in which translations are presented, the evaluation is conducted on a one-to-one basis: compare the source image with the translated image from one of the four different methods. This approach was adopted to impartially assess the effectiveness of the four image translation methodologies.

For both human and GPT-4o powered evaluations, detailed results are provided in the supplementary materials, which include the specific images evaluated and the resulting scores. The detailed evaluation criteria are outlined as follows:

(1) 1 point Very low authenticity: The translated text looks completely unnatural and clearly distinguished from the background of the image as if it was added randomly. Inconsistent style: Ignoring the font, size, color and position of the original text, the inconsistency in style makes the entire translated image feel unreal or abrupt.

(2) 2 points Low authenticity: The translated text is slightly stiff in the image and lacks a sense of integration. It can be clearly seen that it was added later. Partially coordinated style: The translated text tries to imitate the original style to a certain extent, but the overall effect is not good, and the sense of style is more obvious.

(3) 3 points General authenticity: The translated text is relatively natural and can be integrated into the image to a certain extent, but there are still recognizable inconsistencies. Partially coordinated style: The translated text partially echoes the style of the original image and contains the correct elements (such as font, size, color), but still lacks some overall harmony.

(4) 4 points High authenticity: The translated text is well integrated into the image, giving people a more natural feeling, and only small flaws may be found when looking closely. Generally coordinated style: The style of the text matches the original image to a large extent. Small details can be optimized, but the overall look and feel is close to the same.

(5) 5 points High authenticity: The translated text blends perfectly with the image background, and it is almost impossible to tell that the text was added later. Completely coordinated style: The style is completely consistent with the original text, including font, size, color, position and shadow effects, and the overall effect is coordinated and very professional.

In actual evaluation, these two aspects can be considered comprehensively based on the overall effect of the translated image on the score.