

UMTIT: Unifying Recognition, Translation, and Generation for Multimodal Text Image Translation

Liqiang Niu, Fandong Meng and Jie Zhou

Pattern Recognition Center, WeChat AI, Tencent Inc, China
{poetniu, fandongmeng, withtomzhou}@tencent.com

Abstract

Prior research in Image Machine Translation (IMT) has focused on either translating the source image solely into the target language text or exclusively into the target image. As a result, the former approach lacked the capacity to generate target images, while the latter was insufficient in producing target text. In this paper, we present a Unified Multimodal Text Image Translation (UMTIT) model that not only translates text images into the target language but also generates consistent target images. The UMTIT model consists of two image-text modality conversion steps: the first step converts images to text to recognize the source text and generate translations, while the second step transforms text to images to create target images based on the translations. Due to the limited availability of public datasets, we have constructed two multimodal image translation datasets. Experimental results show that our UMTIT model is versatile enough to handle tasks across multiple modalities and outperforms previous methods. Notably, UMTIT surpasses the state-of-the-art TrOCR in text recognition tasks, achieving a lower Character Error Rate (CER); it also outperforms cascading methods in text translation tasks, obtaining a higher BLEU score; and, most importantly, UMTIT can generate high-quality target text images.

Keywords: Image Machine Translation, Multimodal, Text Recognition, Text Translation, Image Generation

1. Introduction

Current Image machine translation (IMT) aims to translate an image containing text in the source language into target language text or a new image containing text in a specific target language. IMT can be widely applied to various types of images, such as scanned book photos, mobile reading screenshots, travel-shot road signboards, and photos taken in daily life.

As illustrated in Figure 1(a), traditional IMT relies on a cascaded system that combines Optical Character Recognition (OCR) (Li et al., 2022), Neural Machine Translation (NMT) (Vaswani et al., 2017) and a complex process of rendering the translated text back onto the source image. This approach carries the potential risk of error compounding between components and redundancy in model parameters. Fortunately, end-to-end models have emerged the dominant approach for natural language processing (NLP) and computer vision (CV) tasks, such as speech translation (Jia et al., 2019), text recognition (Li et al., 2022), and object detection (Carion et al., 2020).

Recent advancements in end-to-end methods for IMT include In-Image Neural Machine Translation (IIMT) (Mansimov et al., 2020; Tian et al., 2023), ItNet (Jain et al., 2021), Text Image Translation (TIT) (Ma et al., 2022), and Document Image Translation (DIT) (Zhang et al., 2023). As depicted in Figure 1(b), Mansimov et al. (2020) pioneered an end-to-end model that directly translates text from single-line source images into images of the target language. However, this model faces challenges due

to its reliance on convolutional networks and pixel-space operations, resulting in translated images with incomplete sentences and indistinct characters. Moreover, it cannot directly translate into the target text; instead, it requires post-processing OCR to extract text from the translated images. Similarly, the new end-to-end IIMT model introduced by Tian et al. (2023) is limited to single-line images, which is not representative of the more complex multi-line images encountered in real-world applications. Figures 1(c) and (d) showcase the end-to-end TIT frameworks proposed by Jain et al. (2021) and Ma et al. (2022), which aim to supplant traditional cascaded approaches that combine OCR with NMT. The primary distinction between the two frameworks is that the former accommodates multi-line text images, whereas the latter is restricted to single-line text images. Compared to earlier TIT models, the recent LayoutDIT model introduced by Zhang et al. (2023) incorporates the complex visual layout of document images to enhance understanding and translation. Although both TIT and DIT models are capable of translating source text images into text of the target language with high translation quality, they still lack the fundamental capability to generate images of the target text.

In this work, we propose UMTIT, a Unified Multimodal Text Image Translation model, as illustrated in Figure 1(e), which integrates text translation and image generation into a single model. Given a text image as input, UMTIT can translate it into the target language without the need for additional OCR and NMT models. Moreover, to address the shortcomings of ItNet and TIT, UMTIT is capable of

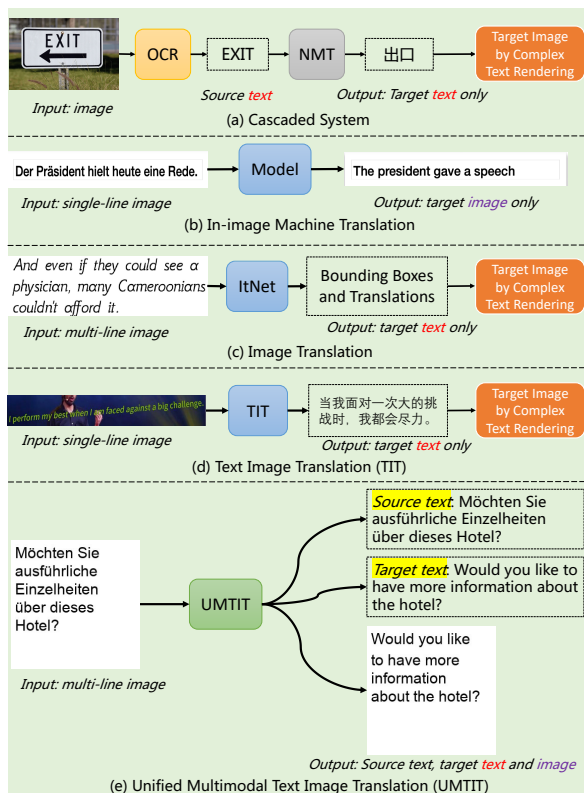


Figure 1: Examples of Image Machine Translation (IMT), including traditional Cascaded System (a), In-image Machine Translation (Mansimov et al., 2020; Tian et al., 2023) (b), Image Translation (Jain et al., 2021) (c), Text Image Translation (TIT) (Ma et al., 2022) (d), and our Multimodal Text Image Translation (e).

generating multi-line target images with consistent style and font. During the text translation process, UMTIT first encodes the image using a vision transformer (Dosovitskiy et al., 2021; Liu et al., 2021) and then decodes the translation with a text transformer (Vaswani et al., 2017; Liu et al., 2020) in an autoregressive manner. UMTIT is also flexible and can be easily extended to recognize the text in the source image.

Inspired by the success of text-to-image generation models like DALL-E (Ramesh et al., 2021), MUSE (Chang et al., 2023), and Stable Diffusion (Rombach et al., 2021), we adopt image tokenizers (Esser et al., 2021; Yu et al., 2021) to learn discrete representations of text images. This allow us to convert images into sequences of image tokens, or vice versa. The sequence of image tokens can be more conveniently integrated and modeled with transformer architecture and is faster compared to pixel space. In the process of image generation, UMTIT first encodes the translation using a character-level transformer, and then generates a sequence of image tokens autoregressively. Based on the generated image image tokens, the image tokenizer

can decode to generate a new target text image.

Due to the lack of publicly available datasets for multimodal text image translation, we constructed two synthetic image datasets with multi-line text. One contains over 100,000 English-Germany images built with the WMT14 dataset, while the other one contains 30,000 Germany-English images built with the Multi30K dataset. Our experiments show that UMTIT outperforms the traditional cascaded systems and achieves higher BLEU scores for text translation, even with fewer model parameters. We also found that UMTIT outperforms the state-of-the-art TrOCR (Li et al., 2022) and achieves a lower Character Error Rate (CER) for text recognition task. Finally, UMTIT can generate high-quality multi-line target images with sharp character details.

We summarize our contributions as follows:

- We propose the UMTIT model, the first model that can recognize text in images, translate the source image to the target language, and most importantly, generate high-quality multi-line target text images.
- To the best of our knowledge, UMTIT is the first model to apply image tokenizers from natural scene images to text images for image translation.
- We construct two synthesized datasets containing over 100,000 images for multimodal text image translation.

2. Related Work

2.1. Multimodal Machine Translation

Multimodal Machine Translation (MMT), extends conventional text-to-text NMT by using an auxiliary image modality on the source side to translate texts into a target language. Elliott et al. (2016) propose a multimodal Multi30K dataset, enabling the development of MMT models. Yao and Wan (2020) introduces a multimodal transformer architecture. Tang et al. (2022) proposes image retrieval methods to collect descriptive images for bilingual parallel corpora using search engines. Peng et al. (2022) proposes a novel framework to support image-free inference. Existing methods focus on improving text-only translation by leveraging additional vision information on the source side, but MMT inherently cannot deal with text images.

2.2. Image Machine Translation

Image Machine Translation (IMT) aims to translate text within an image from the source language into an equivalent image containing the translated text

in the target language. Traditionally, IMT has depended on a cascaded system involving text recognition, text-to-text translation, and a complex rendering process. This approach suffers from several drawbacks, including error propagation, parameter redundancy, and increased latency. To overcome these challenges, various end-to-end methods have been introduced. As an early effort, [Mansimov et al. \(2020\)](#) introduced an end-to-end in-image NMT model with a convolutional encoder-decoder architecture. This model, designed for simplicity, processes *single-line* text images and operates directly in pixel space. However, experiments have shown that the model struggles to produce complete target images and achieves an extremely low BLEU score. Echoing [Mansimov et al. \(2020\)](#), the new end-to-end IIMT model by [Tian et al. \(2023\)](#) is also limited to *single-line* text images, which does not reflect the complexity of multi-line images commonly found in real-world settings. Given the challenges associated with image-to-image translation models, some researchers have proposed Text Image Translation (TIT) methods as a replacement for the cascaded approach that combines OCR and NMT. For instance, [Jain et al. \(2021\)](#) introduced ItNet, an end-to-end neural network designed to translate text within images from one language to another. ItNet has demonstrated superior performance to cascaded systems in side-by-side human evaluations on a synthetic dataset. Additionally, [Ma et al. \(2022\)](#) presented a multi-task training framework, showing that integrating OCR and TIT tasks can further enhance translation performance. Moreover, LayoutDIT, proposed by [Zhang et al. \(2023\)](#), accounts for the complex visual layout of document images to improve comprehension and translation. Although current TIT and DIT models surpass cascaded systems in text translation quality, the challenge of generating target language images remains an open issue.

2.3. Image Synthesis Models

Image synthesis is an exciting computer vision field with significant recent developments and has been applied to various tasks, such as text-to-image generation, class-conditional synthesis, inpainting, and super-resolution. Previously proposed models like Generative Adversarial Networks (GAN) ([Goodfellow et al., 2014](#)), Variational Autoencoders (VAE) ([Kingma and Welling, 2013](#)), flow-based models ([Dinh et al., 2014](#)) and autoregressive models (ARM) ([Chen et al., 2020](#)) have their specific difficulties in training, sample quality, pixel space modeling, and high resolution synthesis. To address these shortcomings, recent two-stage approaches first learn compressed latent representations of images and then model a discrete image space instead of raw pixels. For example, DALL-E

([Ramesh et al., 2021](#)) builds a text-to-image model with a discrete VAE to compress images into image tokens and an autoregressive transformer to model the joint distribution over the text and image tokens. To better represent images, [Esser et al. \(2021\)](#) and [Yu et al. \(2021\)](#) propose new image tokenizers, including VQGAN and ViT-VQGAN. With VQGAN's help, MUSE ([Chang et al., 2023](#)) achieves state-of-the-art text-to-image performance with novel training paradigms of masked transformer modeling of image tokens. Additionally, [Rombach et al. \(2021\)](#) extend Diffusion Probabilistic Models (DM) ([Sohl-Dickstein et al., 2015](#)) to latent diffusion models (LDMs) and achieve new state-of-the-art scores for image inpainting and class-conditional image synthesis.

2.4. Text Rendering Models

Another significant area of image synthesis is text rendering, which is primarily based on diffusion models and aims to generate well-formed visual text. TextDiffuser ([Chen et al., 2023b](#)) synthesizes text at specific locations within images through a two-stage process: the first stage generates the layout information for the text, and the second stage synthesizes the image by combining the mask and text prompt. This process is intricate, and the resulting character accuracy, as measured by OCR accuracy, is relatively low. In its latest iteration, TextDiffuser-2 ([Chen et al., 2023a](#)) utilizes a large language model to generate layout information, including words with bounding boxes. However, these methods have primarily shown effectiveness with short text. GlyphControl ([Yang et al., 2023](#)) facilitates the generation of long-text images but requires pre-specified Glyph Instructions to determine text and position information (glyph rendering). [Liu et al. \(2023\)](#) introduce character-aware models that modify the text encoder's tokenizer from subword-level to character-level, thereby improving Word exact-match accuracy. Nevertheless, the evaluation datasets (such as WikiSpell and DrawText) and the spelling samples provided by the authors indicate that the optimized text tends to be brief, often consisting of single words. This contrasts with our work, which focuses on generating images containing sentence-level translations. Interestingly, our UMTIT reaches a similar conclusion that character-level processing yields better results in generating target text images. [Lotz et al. \(2023\)](#) investigate various rendering strategies to convert sentences into sequences of image patches with continuous characters, aiming to improve the performance of pixel language models. In summary, all the text rendering works discussed here fall short in maintaining alignment with the source image, a key aspect that sets them apart from the image-to-image translation scenario.

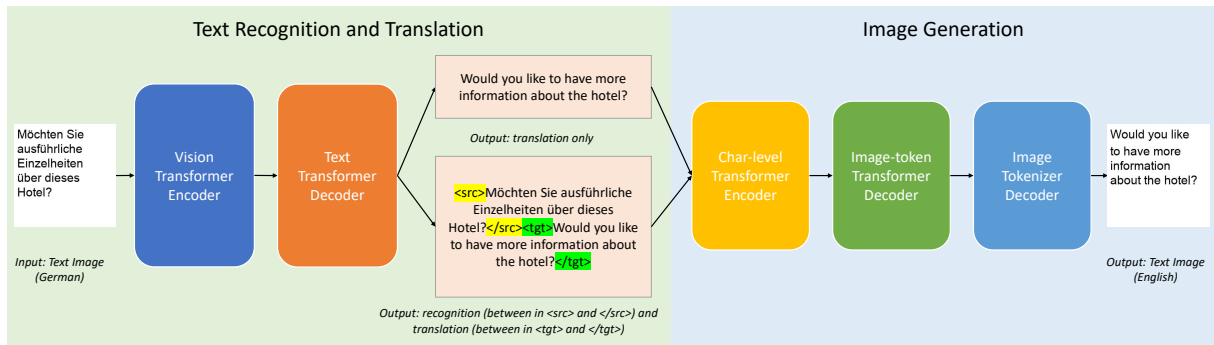


Figure 2: Overview of the UMTIT model, which integrates text recognition, translation and image generation into a single model. As an example of DE2EN multi-line text image translation, UMTIT can recognize the source text "Möchten Sie ausführliche Einzelheiten über dieses Hotel?", translate the image to the target text "Would you like to have more information about the hotel?", and most importantly, generate a new consistent English image.

3. UMTIT Model

As shown in Figure 2, we propose UMTIT, which consists of two components: the left half is for recognition and translation of the source image, and the right half is for generation of the target image.

3.1. UMTIT for Text Translation

Given a source text image X , the first goal of UMTIT is to generate a sequence of target language tokens $T = (t_1, t_2, \dots, t_n)$ for text translation. For example, as shown in Figure 2, the input is a German image containing the source text "Möchten Sie ausführliche Einzelheiten über dieses Hotel?", UMTIT needs to translate it to the target (English) sentence "Would you like to have more information about the hotel?". Without the need for OCR and NMT models, UMTIT uses an end-to-end architecture consisting of a Transformer-based visual encoder and text decoder.

3.1.1. Visual Transformer Encoder

The visual encoder converts the input RGB image $X \in R^{H \times W \times C}$ into a set of embeddings $E = (e_1, e_2, \dots, e_h)$, where $e_i \in R^d$, h is the hidden feature map size or the number of image patches and d is the dimension of hidden vectors. Transformer-based models (Carion et al., 2020; Dosovitskiy et al., 2021; Liu et al., 2021) have become the dominant approach for various computer vision tasks, compared to CNN-based models (He et al., 2016). In this work, we use the Swin Transformer¹ (Liu et al., 2021) as the visual encoder.

¹<https://github.com/microsoft/Swin-Transformer>

3.1.2. Text Transformer Decoder

Given the latent representation E of the image learned by the visual encoder, the text decoder needs to generate a sequence of tokens $T = (t_1, t_2, \dots, t_n)$, where $t_i \in R^v$ is a one-hot vector indicating the i -th target token, v is the vocabulary size, and n is the maximum length of the sequence. Unlike BERT (Devlin et al., 2019), which has only an encoder, and GPT series (Radford et al., 2018, 2019; Brown et al., 2020) with only a decoder, BART (Lewis et al., 2020) uses a Transformer-based encoder-decoder architecture. Inspired by Kim et al. (2021), we use BART as the text decoder. Another advantage of the BART decoder is that we can easily modify it to accept arbitrary inputs with a cross-attention mechanism.

3.2. UMTIT for Text Recognition

With a simple vision-encoder and text-decoder architecture, UMTIT is flexible and can be easily extended to generate arbitrary tokens. Specifically, given a source text image X , UMTIT can be modified to generate a sequence of m source language tokens $S = (s_1, s_2, \dots, s_m)$ and a sequence of n target language tokens $T = (t_1, t_2, \dots, t_n)$ simultaneously. For simplicity, we concatenate the source and target sequences into a merged sequence $M = (\langle src \rangle s_1, s_2, \dots, s_m \langle /src \rangle \langle tgt \rangle t_1, t_2, \dots, t_n \langle /tgt \rangle)$, where $\langle src \rangle, \langle /src \rangle$ and $\langle tgt \rangle, \langle /tgt \rangle$ are special tokens used to distinguish the source and target sequences, respectively.

As a result, UMTIT can accomplish both text recognition and translation tasks simultaneously. As shown in Figure 2, given an input of German text image, UMTIT can generate a sequence including source German and target English tokens " $\langle src \rangle$ Möchten Sie ausführliche Einzelheiten über dieses Hotel? $\langle /src \rangle \langle tgt \rangle$ Would you like to have

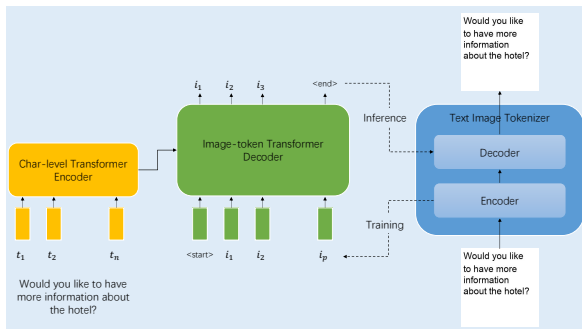


Figure 3: Text Image Generation of UMTIT with details of training and inference, as well as the interaction logic with the text image tokenizer.

more information about the hotel?". We prioritize text recognition during the decoding process because it can further enhance translation due to the self-attention mechanism in the decoder.

3.3. UMTIT for Text Image Generation

3.3.1. Text Image Tokenizer

With the help of image tokenizers, modern image generative models can directly learn in latent token space instead of raw pixels, enabling more effective training losses such as cross-entropy instead of regression. However, most existing image tokenizers (Ramesh et al., 2021; Esser et al., 2021; Yu et al., 2021) focus on natural scene images, which cannot be applied to textual images. In this work, we tokenize text images using ViT-VQGAN (Yu et al., 2021), which consists of a Vision-Transformer encoder and a Transformer decoder, with a quantization layer that maps an input image into a sequence of discrete tokens from a learned codebook. Given an image of resolution $H \times W$, ViT-VQGAN encodes it into discretized latent codes of size $H/f \times W/f$, where f is the image patch size which can also be referred to as the downsampling ratio. For example, a 256×256 RGB image can be encoded to 32×32 and 16×16 tokens with $f = 8$ and $f = 16$, respectively.

3.3.2. Text Image Generation

After obtaining the translation $T = (t_1, t_2, \dots, t_n)$, the second goal is to generate a new, consistent target image Y that aligns with the source image X . Inspired by text-to-image models like MUSE (Chang et al., 2023) and Parti (Yu et al., 2022), we use a similar sequence-to-sequence architecture as shown in Figure 3. UMTIT first encodes the target text (e.g., an English sentence) with a character transformer, and then uses an autoregressive image-token transformer to generate a sequence of image tokens $I = (i_1, i_2, \dots, i_p)$, where $i_i \in R^w$

is a one-hot vector for i -th target image token, w is the vocabulary and codebook size, and p is the maximum number of tokens. Note that the image-token transformer decoder shares the vocabulary with codebook of the image tokenizer.

4. Experiment

4.1. Datasets

Currently, there is no public dataset available for Multimodal Text Image Translation (MTIT) tasks. Therefore, we have created two datasets to address this need.

MTIT-WMT-100K: The construction process consists of two steps. First, we extract appropriate bilingual text from WMT14-en-de² under certain filtering criteria. Next, we generate English and German text images based on the aligned English and German text. Ultimately, we collect over 100,000 pairs of bilingual text images, with an image resolution of 256×256 to balance storage cost and image quality. The extraction of bilingual text is as follows: considering the limited size of the image, the extracted text must have a limited number of characters. The original WMT14-en-de contains over 400 million text pairs, we first select suitable text pairs filtered by character length with two conditions: (1) the source and target text's length must be in the range of $[1, 60]$; (2) the source text's length $\times 2$ is less than the target text's length and vice versa. To maintain the quality of text pairs, we filter out text pairs with low-frequency words. Finally, we have 106,403 text pairs for training, and 1,166 text pairs for testing.

MTIT-Multi30K: Multi30K³ is a public dataset for multilingual multimodal machine translation. We select the full *de-en* part with 29,000 training sentences, 1,014 validation sentences, and 1,000 test sentences to build a smaller dataset. Note that we do not filter out long sentences, therefore, MTIT-Multi30K text image resolution is set to 512×512 .

To generate high quality text images, we use the public Python Pillow package⁴ and the well-known *Arial*⁵ font. For each sentence, we evenly fill it in multiple rows of the image in a left-to-right, top-to-bottom order. For simplicity, all images use black text on a white background. More details of our synthetic datasets compared with related works are shown in Table 1.

²<https://github.com/facebookresearch/fairseq/tree/main/examples/translation/prepare-wmt14en2de.sh>

³<https://github.com/multi30k/dataset>

⁴<https://github.com/python-pillow/Pillow>

⁵<https://learn.microsoft.com/en-us/typography/font-list/arial>



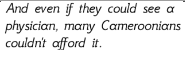
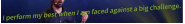
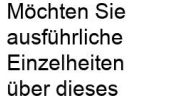
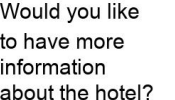
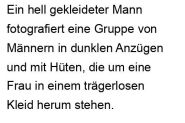
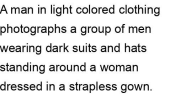
Synthetic Datasets	Single/ Multi-Line	Task	Source Image	Source Text	Target Text	Target Image
Mansimov et al. (2020)	Single-line	In-Image Translation		×	×	
Jain et al. (2021)	Multi-line	Image Translation		×	✓	×
Ma et al. (2022)	Single-line	Text Image Translation		×	✓	×
MTIT-WMT-100K (ours)	Multi-line	Multimodal Text Image Translation		✓	✓	
MTIT-Multi30K (ours)	Multi-line	Multimodal Text Image Translation		✓	✓	

Table 1: Comparison of our synthetic datasets with related works. [Mansimov et al. \(2020\)](#) dataset including $\{SourceImage, TargetImage\}$ is used for In-Image Translation, [Jain et al. \(2021\)](#) and [Ma et al. \(2022\)](#) datasets including $\{SourceImage, TargetText\}$ are used for Text Image Translation (TIT). Our MTIT-WMT-100K and MTIT-Multi30K containing $\{SourceImage, SourceText, TargetText, TargetImage\}$ are used for Multimodal Text Image Translation (MTIT).

Models	MTIT-WMT-100K		MTIT-Multi30K		#Params
	CER (De)	CER (En)	CER (De)	CER (En)	
Fine-tuned TrOCR-Base (Li et al., 2022)	0.0086	0.0332	0.0148	0.0252	384M
UMTIT (ours)	0.0012	0.0011	0.0007	0.0013	293M

Table 2: Comparison of CER and number of model parameters.

4.2. Training

With data pairs of $\{S, X, T, Y\}$, where S is the *source text*, X is the *source text image*, T is the *target text*, and Y is the *target text image*, the training process of our UMTIT consists of two parts:

Text Image Tokenizer: We first train several ViT-VQGAN models for both source and target text images $\{X, Y\}$.

UMTIT: Given a source text image X , UMTIT needs to output $\{S, T, Y\}$, so the training objective is to maximize $p(S, T, Y|X) = p_\theta(Y|T) \times p_\phi(S, T|X)$, where ϕ represents the parameters for text recognition and translation, while θ denotes the parameters for image generation.

4.3. Results and Analysis

As described in Section 3, our flexible UMTIT can recognize the source text in images, translate the image to text in target language, and generate a new target text image simultaneously. In this section, we evaluate UMTIT’s performance on multiple tasks with a comprehensive analysis.

4.3.1. Text Recognition

We first evaluate UMTIT’s performance on text recognition using the Character Error Rate (CER)⁶. For a fair comparison, we fine-tuned the pretrained base TrOCR⁷ on the same datasets and tested it with the best checkpoint. As shown in Table 2, UMTIT achieves a lower CER than the state-of-the-art TrOCR ([Li et al., 2022](#)) model with fewer model parameters. Additional comparative examples are available in Table 7 in Appendix A.1.

4.3.2. Text Translation

Next, we evaluate the text translation performance using the SacreBLEU⁸ metric and compare UMTIT models with Text-to-Text NMT models and Image-to-Text cascaded systems. We use fairseq to implement the NMT models with different architectures (IWLST as Small, Base, and Big). For cascaded systems, we use the fine-tuned TrOCR-Base model

⁶<https://github.com/huggingface/datasets/tree/main/metrics/cer>

⁷<https://github.com/microsoft/unilm/tree/master/trocr>

⁸<https://github.com/mjpost/sacrebleu>

Models	Input Modality	Output Modality	MTIT-WMT-100K		MTIT-Multi30K	#Params
			DE2EN	EN2DE	DE2EN	
<i>NMT Models (Transformer Encoder - Decoder)</i>						
Transformer-Small	Text	Text	40.46	36.27	39.81	48M
Transformer-Base	Text	Text	39.46	35.72	38.11	65M
Transformer-Big	Text	Text	31.67	32.13	38.91	213M
<i>Cascaded Systems (Fine-tuned TrOCR-Base + NMT)</i>						
TrOCR+Transformer-Small	Image	Text	38.95	33.69	38.99	432M
TrOCR+Transformer-Base	Image	Text	38.72	33.75	38.11	449M
TrOCR+Transformer-Big	Image	Text	31.49	30.92	37.68	597M
<i>End-to-End Models (Vision Encoder - Text Decoder)</i>						
UMTIT-Trans-only (ours)	Image	Text	38.89	33.11	39.44	293M
UMTIT-Recog+Trans (ours)	Image	Text	39.03	34.04	40.90	293M
<i>+ No Early Stopping</i>						
UMTIT-Trans-only (ours)	Image	Text	38.91	33.71	39.77	293M
UMTIT-Recog+Trans (ours)	Image	Text	39.12	34.60	41.00	293M

Table 3: SacreBLEU scores for Text Image Translation.

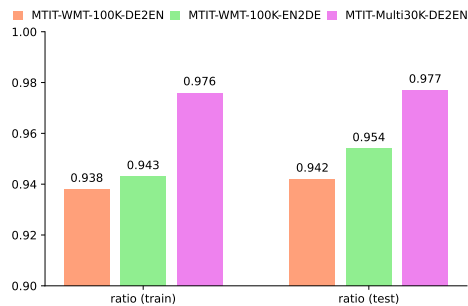


Figure 4: Comparison of the ratio of the average length of the translation and the reference.

to convert the test image into source text, which was then translated into the target language with NMT models. UMTIT is an end-to-end Image-to-Text model that supports two variants, one that outputs only the translation (UMTIT-Trans-only) and another that outputs the recognition followed by the translation (UMTIT-Recog+Trans). Note that NMT models with ground-truth source text can be viewed as the upper bound of Image-to-Text models. As shown in Table 3, we draw the following conclusions:

- All cascaded systems exhibit lower SacreBLEU scores compared to the upper-bound NMT models, primarily because OCR errors in the source image can lead to inaccuracies in text translation. Additionally, cascaded systems comprise both OCR and NMT models, resulting in a higher total number of parameters.
- Our end-to-end UMTIT models achieve comparable results with best cascaded systems when using UMTIT-Trans-only. However, with UMTIT-Recog+Trans, UMTIT performs better than the best cascaded systems on all

datasets. As mentioned in Section 3.2, the text recognition task can further improve translation performance.

- Our UMTIT performs worse than the upper bound NMT models on the MTIT-WMT-100K but still has a chance to perform better on the MTIT-Multi30K test set.
- We further analyze the difference between UMTIT on MTIT-WMT-100K and MTIT-Multi30K and find that the length of UMTIT’s translation on MTIT-WMT-100K is significantly shorter than that of MTIT-Multi30K. We calculate the ratio of the average length of the translation and the reference, as shown in Figure 4. The ratio of MTIT-Multi30K is nearly 0.98, while the ratio of MTIT-WMT-100K is less than 0.95. This behavior may be caused by the distribution of training data and the decoding strategy. Therefore, we further explore the decoding strategy and disable *early stopping*. As shown at the bottom of Table 3, without *early stopping*, UMTIT further improves performance, with 0.02-0.6 and 0.09-0.56 BLEU improvement for UMTIT-Trans-only and UMTIT-Recog+Trans, respectively.

Additional translation comparison examples are available in Table 8 and 9 in Appendix A.2.

4.3.3. Text Image Tokenizers

In this section, we adapt ViT-VQGAN models from natural images to text images and compare their performance with existing trained image tokenizers, such as DALL-E (Ramesh et al., 2021), VQGAN (Esser et al., 2021), ViT-VQGAN-Base (Yu et al., 2021). The major differences of these models are

Models	Architecture	Downsampling Factor (f)	Codebook Size	#Tokens (256×256 images)	#Tokens (512×512 images)
<i>Existing Open Source Models</i>					
DALL-E (Ramesh et al., 2021)	CNN	8	8192	1024	4096
VQGAN (Esser et al., 2021)	CNN	8	8192	1024	4096
VQGAN (Esser et al., 2021)	CNN	16	16384	256	1024
VQGAN (Esser et al., 2021)	CNN	16	1024	256	1024
ViT-VQGAN (Yu et al., 2021)	ViT	8	8192	1024	4096
<i>Trained from scratch on the MTIT-WMT-100K dataset</i>					
ViT-VQGAN (ours)	ViT	16	1024	256	1024

Table 4: Comparison of our ViT-VQGAN with existing image tokenizers.

Image Size	256×256		512×512	
Metrics	FID↓	SSIM↑	FID↓	SSIM↑
DALL-E(f8,8192)	13.32	0.7041	15.51	0.7724
VQGAN(f8,8192)	9.27	0.6962	4.40	0.7650
VQGAN(f16,16384)	17.38	0.6931	14.75	0.7692
VQGAN(f16,1024)	31.69	0.6878	12.49	0.7655
ViT-VQGAN(f8,8192)	26.86	0.6992	29.47	0.7683
ViT-VQGAN (ours)	2.50	0.7091	4.04	0.7733

Table 5: Comparison of FID and SSIM for Text Image Reconstruction.

shown in Table 4 and only our ViT-VQGAN is trained from scratch using the MTIT-WMT-100K dataset.

We use Fréchet Inception Distance (FID) and Structural Similarity Index Measure (SSIM) metrics to evaluate the reconstruction quality. As shown in Table 5, without training on synthetic datasets, DALL-E(f8,8192) and VQGAN(f8,8192) achieve the best SSIM and FID. However, their down-sampling factor is small ($f=8$), leading to long sequences of image tokens (e.g., 4096 for 512×512 images), which make auto-regressive generation of images slow and difficult. Therefore, we train our ViT-VQGAN from scratch and achieve the best FID and SSIM for both 256×256 and 512×512 images even with a larger down-sampling factor ($f=16$). Further examples of image reconstruction comparisons can be found in Table 10 in Appendix A.3.

4.3.4. Text Image Generation

Finally, we evaluate the quality of text images generated by UMTIT. To balance the tradeoff between quality and cost, we resize all text images to 384×384 which can be represented by 576 tokens with a downsampling factor $f=16$. We also trained several ViT-VQGANs with different codebook size (1024, 2048, 4096) as candidate image tokenizers. We have some conclusions:

- As shown in Figure 5, character-level transformer achieves a lower training loss than word-level, and performs better in generation of text images.

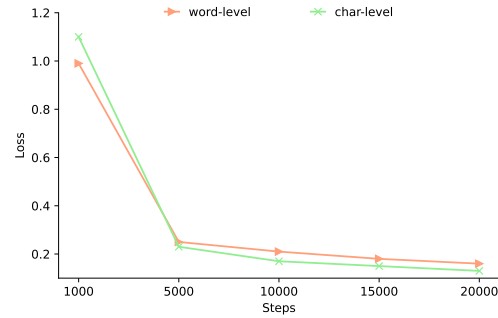


Figure 5: Comparison of training loss for Char-level and Word-level text transformer.

- As shown in Table 6, with larger training steps and codebook size, the generated images become increasingly closer to the ground-truth images. More examples can be found in Table 13 in Appendix A.4.

Target Text (EN)	Reference		Translation	
Metrics	FID↓	SSIM↑	FID↓	SSIM↑
<i>Training Steps=20000</i>				
ViT-VQGAN (1024)	20.84	0.9236	21.49	0.8082
ViT-VQGAN (2048)	22.02	0.9329	21.59	0.8179
ViT-VQGAN (4096)	18.12	0.9330	18.30	0.8191
<i>Training Steps=30000</i>				
ViT-VQGAN (1024)	18.06	0.9275	18.84	0.8139
ViT-VQGAN (2048)	20.77	0.9332	26.71	0.8089
ViT-VQGAN (4096)	17.26	0.9351	18.57	0.8191

Table 6: Comparison of FID and SSIM between generated images and ground-truth images on MTIT-WMT-100K-DE2EN task. Here, "Reference" refers to images generated using ground-truth target text, while "Translation" denotes images generated from translated text.

Observing the text images generated by UMTIT, our method has achieved consistent results between the source and target images, especially in terms of line style. Our approach encompasses several key components:

- Firstly, we have developed two bilingual image datasets that ensure strict alignment in font size, line style, and layout between the source and target images. We meticulously curated two high-quality datasets from diverse sources: MTIT-WMT-100K and MTIT-Multi30K. Additionally, we are pioneers in proposing a dataset for complex multi-line image-to-image translation, which includes four elements: the original image, the source language text, the target language translation, and the translated image. This represents a significant advancement over previous research.
- Secondly, we trained a unified ViT-VQGAN tokenizer using the high-quality dataset, which guarantees consistent representation for both source and target images, even though they are in different languages (English and German).
- Lastly, for the generation of translated images, we utilized a fine-grained, character-level encoder. Our experiments have indicated that this method ensures greater consistency in the generated images compared to using a word-level encoder.

In conclusion, our UMTIT model is capable of effectively performing multiple tasks in multimodal text image translation. Additional end-to-end examples are available in Table 14 in Appendix A.5.

5. Conclusion

Research in Image Machine Translation (IMT) can be divided into two principal categories. The first focuses on translating images into text, where traditional cascaded systems are now being outperformed by cutting-edge end-to-end models like It-Net, TIT, and DIT. These models deliver enhanced translation quality but are unable to produce corresponding images in the target language. The second category is concerned with image-to-image translation, which presents a more complex learning challenge. As a result, the images generated by these models often contain incomplete sentences and lack clarity in detail. Moreover, they are limited to handling single-line text images and do not produce translated text. To overcome these limitations, we introduce UMTIT, the pioneering model that combines multiple multimodal tasks, including text recognition, translation, and image generation. During our experimental phase, we carefully developed two multimodal translation datasets featuring multi-line text images. Our findings demonstrate that UMTIT surpasses existing models in various tasks and, crucially, it can generate high-quality target images that maintain a consistent style.

Limitations

Although we propose UMTIT as the first model that can accomplish text recognition and text translation, and most importantly, generate target text images for multimodal text image translation task, there are still many aspects that can be improved in the future. We list some major limitations and optimization direction as follows:

- Typically, to obtain the final translated target image, UMTIT requires three steps: translating the source text image to the target language, generating the target image tokens, and finally decoding to the target text image with image tokenizers. Even though these modules are included in a single model, the architecture of UMTIT can be optimized. For example, for some scenarios that do not require target text, UMTIT can be designed to skip text translation and directly generate the consistent target image tokens.
- Due to the absence of public datasets for multimodal text image translation task, we constructed two synthetic datasets, including MTIT-WMT-100K and MTIT-Multi30K and verified multiple objectives of our UMTIT model. Although these text images already has multi-line layouts, the font style and image backgrounds are relatively monotonous. In the future, we need to build more complex and realistic text images with different layouts, font styles, and backgrounds for large-scale MTIT tasks.

6. Bibliographical References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision—*

- ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer.
- Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.
- Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023a. [Textdiffuser-2: Unleashing the power of language models for text rendering](#).
- Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2023b. [Textdiffuser: Diffusion models as text painters](#).
- Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. [Generative pretraining from pixels](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Laurent Dinh, David Krueger, and Yoshua Bengio. 2014. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. [Multi30K: Multilingual English-German image descriptions](#). In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.
- Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. [Generative adversarial nets](#). In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Puneet Jain, Orhan Firat, Qi Ge, and Sihang Liang. 2021. [Image translation network](#). In *Image Translation Model*.
- Ye Jia, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Z. Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *ArXiv*, abs/1904.06037.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2021. Ocr-free document understanding transformer. *arXiv preprint arXiv:2111.15664*.
- Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022. [Trocr: Transformer-based optical character recognition with pre-trained models](#).
- Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, Rj Mical, Mohammad Norouzi, and Noah Constant. 2023. [Character-aware models improve visual text rendering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

- Papers), pages 16270–16297, Toronto, Canada. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual denoising pre-training for neural machine translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022.
- Jonas Lotz, Elizabeth Salesky, Phillip Rust, and Desmond Elliott. 2023. Text rendering strategies for pixel language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10155–10172.
- 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](#). In *26th International Conference on Pattern Recognition, ICPR 2022, Montreal, QC, Canada, August 21-25, 2022*, pages 1664–1670. IEEE.
- Elman Mansimov, Mitchell Stern, Mia Xu 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*, pages 70–74.
- Ru Peng, Yawen Zeng, and Jake Zhao. 2022. [Distill the image to nowhere: Inversion knowledge distillation for multimodal machine translation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2379–2390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. [High-resolution image synthesis with latent diffusion models](#).
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. [Deep unsupervised learning using nonequilibrium thermodynamics](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France. PMLR.
- ZhenHao Tang, XiaoBing Zhang, Zi Long, and XiangHua Fu. 2022. [Multimodal neural machine translation with search engine based image retrieval](#). In *Proceedings of the 9th Workshop on Asian Translation*, pages 89–98, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.
- Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo, and Bin Wang. 2023. [In-image neural machine translation with segmented pixel sequence model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15046–15057, Singapore. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *CoRR*, abs/1706.03762.
- Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. 2023. [Glyphcontrol: Glyph conditional control for visual text generation](#).
- Shaowei Yao and Xiaojun Wan. 2020. [Multimodal transformer for multimodal machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.
- Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.
- Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*.
- Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong.

2023. *LayoutDIT: Layout-aware end-to-end document image translation with multi-step conductive decoder*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10043–10053, Singapore. Association for Computational Linguistics.

A. Appendix

A.1. Text Recognition Examples

Table 7 demonstrates that our UMTIT model achieves more accurate recognition performance compared to the fine-tuned TrOCR model.

A.2. Text Translation Examples

Comparative examples of text-to-text NMT models, cascaded systems, and our UMTIT are presented in Tables 8 and 9.

A.3. Text Image Tokenizers

Examples of image reconstructions by various image tokenizers are compared in Table 10. Table 11 presents a 2D visualization of the codebook embeddings from our trained ViT-VQGAN image tokenizers using t-SNE.

A.4. Text Image Generation Examples

We compare images generated by character-level and word-level text transformer encoders, as shown in Table 12. Additionally, we compare images generated with reference to different ViT-VQGAN image tokenizers, as illustrated in Table 13.

A.5. Multimodal Text Image Translation Examples

In this section, we present case studies of our UMTIT model for multimodal text image translation, as shown in Table 14.

MTIT-WMT-100K Images (DE/EN)		
	Möchten Sie ausführliche Einzelheiten über dieses Hotel?	Would you like to have more information about the hotel?
Ground-truth	<i>Möchten Sie ausführliche Einzelheiten über dieses Hotel?</i>	<i>Would you like to have more information about the hotel?</i>
Fine-tuned TrOCR-Base (Li et al., 2022)	<i>Möchten Sie ausführliche Einzelheiten über dieses Hoteli?</i>	<i>Would you like to have more more information about the hotel?</i>
UMTIT (ours)	<i>Möchten Sie ausführliche Einzelheiten über dieses Hotel?</i>	<i>Would you like to have more information about the hotel?</i>
	Der SBB Bahnhof ist nur ein paar Gehminuten vom Hotel.	Close to transportation and shopping areas.
Ground-truth	<i>Der SBB Bahnhof ist nur ein paar Gehminuten vom Hotel.</i>	<i>Close to transportation and shopping areas.</i>
Fine-tuned TrOCR-Base	<i>Der SIB Bahnhof ist nur ein paar Gehminuten vom Hotel.</i>	<i>Close to transportation and shopping areas.</i>
UMTIT (ours)	<i>Der SBG Bahnhof ist nur ein paar Gehminuten vom Hotel.</i>	<i>Close to translation and shopping areas.</i>
	Gaddafi muss gehen.	Gaddafi must go.
Ground-truth	<i>Gaddafi muss gehen.</i>	<i>Gaddafi must go.</i>
Fine-tuned TrOCR-Base	<i>Gaddarf muss gehen.</i>	<i>Gaza must go.</i>
UMTIT (ours)	<i>Saddam muss gehen.</i>	<i>Caddafi must go.</i>
MTIT-Multi30K Images (DE/EN)		
	Ein hell gekleideter Mann fotografiert eine Gruppe von Männern in dunklen Anzügen und mit Hüten, die um eine Frau in einem trägerlosen Kleid herum stehen.	A man in light colored clothing photographs a group of men wearing dark suits and hats standing around a woman dressed in a strapless gown.
Ground-truth	<i>Ein hell gekleideter Mann fotografiert eine Gruppe von Männern in dunklen Anzügen und mit Hüten, die um eine Frau in einem trägerlosen Kleid herum stehen.</i>	<i>A man in light colored clothing photographs a group of men wearing dark suits and hats standing around a woman dressed in a strapless gown.</i>
Fine-tuned TrOCR-Base (Li et al., 2022)	<i>Ein hell gekleideter Mann fotografiert eine Gruppe von Männern in dunklen Anzügen und mit Hüten, die um eine Frau in einem trägelosen Kleid herum stehen.</i>	<i>A man in light colored clothing photographs a group of men wearing dark suits and hats standing around a woman dressed in a small dress.</i>
UMTIT (ours)	<i>Ein hell gekleideter Mann fotografiert eine Gruppe von Männern in dunklen Anzügen und mit Hüten, die um eine Frau in einem trägerlosen Kleid herum stehen.</i>	<i>A man in light colored clothing photographs a group of men wearing dark suits and hats standing around a woman dressed in a strapless gown.</i>

Table 7: Examples of Text Recognition for sampled text images, with errors highlighted in red.

MTIT-WMT-100K Images (DE2EN and EN2DE)		
	Möchten Sie ausführliche Einzelheiten über dieses Hotel?	Would you like to have more information about the hotel?
Reference	<i>Would you like to have more information about the hotel?</i>	<i>Möchten Sie ausführliche Einzelheiten über dieses Hotel?</i>
NMT Models	<i>Do you want more detailed details about this hotel?</i>	<i>Möchten Sie weitere Informationen über das Hotel?</i>
Cascaded System	<i>Do you want more detailed details on this House?</i>	<i>Möchten Sie mehr Informationen über das Hotel haben?</i>
UMTIT-Trans-only (ours)	<i>Would you like to stay at the hotel?</i>	<i>Möchten Sie mehr Informationen zum Hotel übernachten?</i>
UMTIT-Recog+Trans (ours)	<i>Would you like detailed information about this hotel?</i>	<i>Möchten Sie mehr Informationen über das Hotel?</i>
UMTIT-Trans-only (ours) + No Early Stopping	<i>Would you like to receive detailed information about this hotel?</i>	<i>Möchten Sie mehr Informationen zum Hotel übernachten?</i>
UMTIT-Recog+Trans (ours) + No Early Stopping	<i>Would you like detailed information about this hotel?</i>	<i>Möchten Sie mehr Informationen über das Hotel?</i>
	Der SBB Bahnhof ist nur ein paar Gehminuten vom Hotel.	The SBB railway station is in proximity to the hotel.
Reference	<i>The SBB railway station is in proximity to the hotel.</i>	<i>Der SBB Bahnhof ist nur ein paar Gehminuten vom Hotel.</i>
NMT Models	<i>SBB station is a few minutes walk from the hotel.</i>	<i>Der Bahnhof von SBB ist in der Nähe des Hotels.</i>
Cascaded System	<i>The SIB station is a few minutes walk from the hotel.</i>	<i>Der Bahnhof & Bahnhof befindet sich in der Nähe des Hotels.</i>
UMTIT-Trans-only (ours)	<i>The SBS Bahnhof is just a few minutes from the hotel.</i>	<i>Der BBQ Bahnhof liegt in der Nähe des Hotels.</i>
UMTIT-Recog+Trans (ours)	<i>The SBG railway station is just a few minutes from the hotel.</i>	<i>Die SBS-Strategie befindet sich in der Nähe vom Hotel.</i>
UMTIT-Trans-only (ours) + No Early Stopping	<i>SBS railway station is only a few minutes from the hotel.</i>	<i>Der BBQ Bahnhof befindet sich in der Nähe des Hotels.</i>
UMTIT-Recog+Trans (ours) + No Early Stopping	<i>The SBG railway station is just a few minutes from the hotel.</i>	<i>Die SBS-Strategie befindet sich in der Nähe vom Hotel.</i>
	Genehmigung des Protokolls der vorangegangenen Sitzung	Approval of the Minutes of the previous sitting
Reference	<i>Approval of the Minutes of the previous sitting</i>	<i>Genehmigung des Protokolls der vorangegangenen Sitzung</i>
NMT Models	<i>Approval of the Minutes of the previous sitting</i>	<i>Genehmigung des Protokolls der vorangegangenen Sitzung</i>
Cascaded System	<i>Approval of the Minutes of the previous sitting</i>	<i>Anmeldungen zum Protokoll der vorangegangenen Sitzung</i>
UMTIT-Trans-only (ours)	<i>Approval of the Minutes of the previous sitting</i>	<i>Genehmigung des Protokolls der vorangegangenen Sitzung</i>
UMTIT-Recog+Trans (ours)	<i>Approval of the Minutes of previous sitting</i>	<i>Genehmigung des Protokolls der letzten Sitzung</i>
UMTIT-Trans-only (ours) + No Early Stopping	<i>Approval of the Minutes of the previous sitting</i>	<i>Genehmigung des Protokolls der vorangegangenen Sitzung</i>
UMTIT-Recog+Trans (ours) + No Early Stopping	<i>Approval of the Minutes of the previous sitting</i>	<i>Genehmigung des Protokolls der vorangegangenen Sitzung</i>

Table 8: Examples of Text Translation for MTIT-WMT-100K DE2EN and EN2DE. Note that NMT models use ground-truth source text as input, while Cascaded System and UMTITs use source image.

MTIT-Multi30K Images (DE2EN)		
	<p>Ein hell gekleideter Mann fotografiert eine Gruppe von Männern in dunklen Anzügen und mit Hüten, die um eine Frau in einem trägerlosen Kleid herum stehen.</p>	<p>Ein Mädchen beim Seilhüpfen auf dem Gehweg nahe einer Garage.</p>
Reference	<i>A man in light colored clothing photographs a group of men wearing dark suits and hats standing around a woman dressed in a strapless gown.</i>	<i>A girl jumping rope on a sidewalk near a parking garage.</i>
NMT Models	<i>A man dressed in bright colors a group of men in dark suits and hats stand around a woman in a heard.</i>	<i>A girl jumping rope on the sidewalk near a garage.</i>
Cascaded System	<i>A man dressed in bright colors a group of men in dark suits and hats stand around a woman in a hears' dress.</i>	<i>A girl jumping rope on the sidewalk near a garage.</i>
UMTIT-Trans-only (ours)	<i>A man dressed in light colored photographs a group of men dressed in dark suits and hats stand around a woman wearing a saddle.</i>	<i>A girl is on the sidewalk near a garage.</i>
UMTIT-Recog+Trans (ours)	<i>A man dressed in brightly colored photographs a group of men dressed in dark suits and hats standing around a woman wearing a pink dress standing around an outdoor dress.</i>	<i>A girl is sculpting on the sidewalk near a garage.</i>
UMTIT-Trans-only (ours) + No Early Stopping	<i>A man dressed in light colored photographs a group of men dressed in dark suits and hats stand around a woman wearing a saddle.</i>	<i>A girl is on the sidewalk near a garage.</i>
UMTIT-Recog+Trans (ours) + No Early Stopping	<i>A man dressed in brightly colored photographs a group of men in dark suits and hats standing around a woman in a pink dress standing around.</i>	<i>A girl doing a rope jump on the sidewalk near a garage.</i>

Table 9: Examples of Text Translation for MTIT-Multi30K DE2EN. Note that NMT models use ground-truth source text as input, while Cascaded System and UMTITs use source text image.

	Input Image (DE)	Input Image (EN)	Input Image (DE)	Input Image (EN)
Image Size	256×256	256×256	512×512	512×512
Original Image	Ein Europa ohne Grenzen liegt im Interesse der Verbraucher.	A Europe without borders is to the benefit of consumers.	Diesen sehr realen Problemen müssen wir uns zuwenden.	We must address those very real issues.
<i>Existing Models</i>				
DALL-E(f8,8192)	Ein Europa ohne Grenzen liegt im Interesse der Verbraucher.	A Europe without borders is to the benefit of consumers.	Diesen sehr realen Problemen müssen wir uns zuwenden.	We must address those very real issues.
VQGAN(f8,8192)	Ein Europa ohne Grenzen liegt im Interesse der Verbraucher.	A Europe without borders is to the benefit of consumers.	Diesen sehr realen Problemen müssen wir uns zuwenden.	We must address those very real issues.
VQGAN(f16,16384)	Ein Europa ohne Grenzen liegt im Interesse der Verbraucher.	A Europe without borders is to the benefit of consumers.	Diesen sehr realen Problemen müssen wir uns zuwenden.	We must address those very real issues.
VQGAN(f16,1024)	Ein Europa ohne Grenzen liegt im Interesse der Verbraucher.	A Europe without borders is to the benefit of consumers.	Diesen sehr realen Problemen müssen wir uns zuwenden.	We must address those very real issues.
ViT-VQGAN(f8,8192)	Ein Europa ohne Grenzen liegt im Interesse der Verbraucher.	A Europe without borders is to the benefit of consumers.	Diesen sehr realen Problemen müssen wir uns zuwenden.	We must address those very real issues.
<i>Our Trained Models</i>				
ViT-VQGAN(f16,1024)	Ein Europa ohne Grenzen liegt im Interesse der Verbraucher	A Europe without borders is to the benefit of consumers	Diesen sehr realen Problemen müssen wir uns zuwenden.	We must address those very real issues.

Table 10: Demonstration of image reconstruction using different Tokenizers for text images sampled from the MTIT-WMT-100K test set.

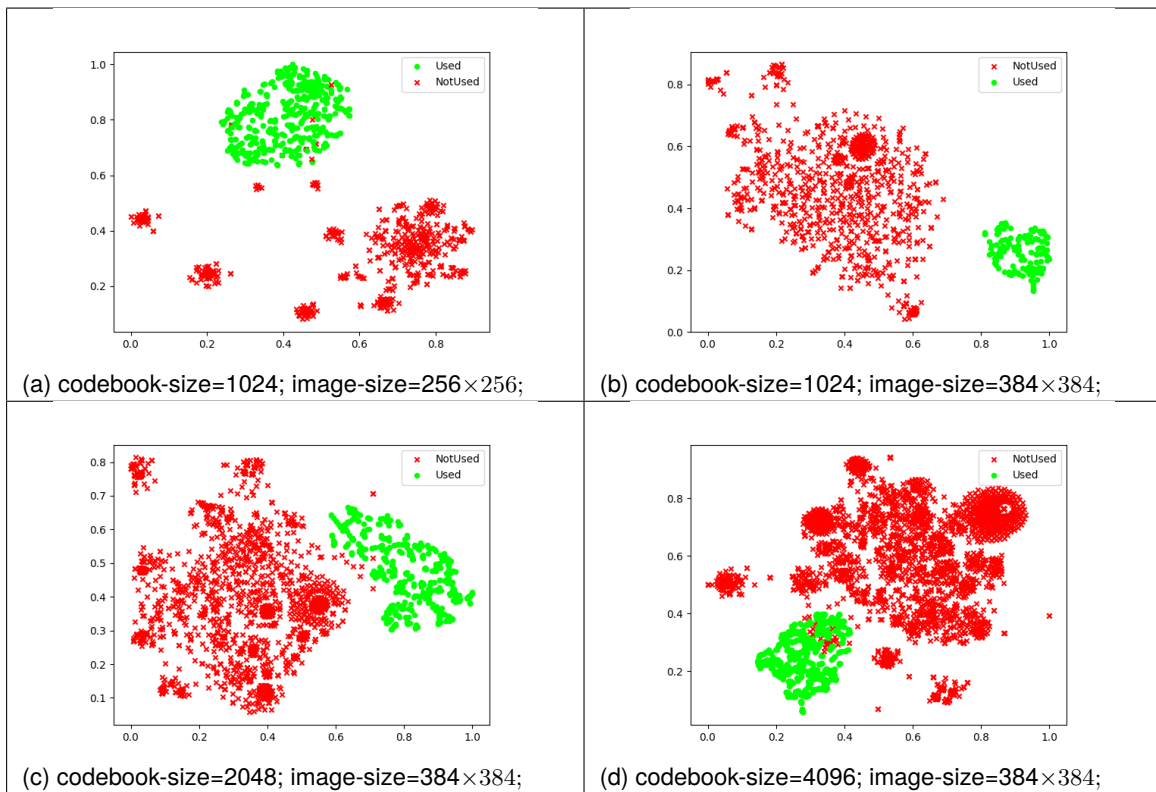


Table 11: The t-SNE visualization illustrates the codebook embeddings for various ViTVQGAN image tokenizers with a downsampling ratio of 16 ($f=16$). Green points represent embeddings utilized in image reconstruction, whereas red points indicate those that are not used.

	Word-level		Char-level	
Step-5000	have pointed out, much is going to be done and many	have poied out, much bis gging to be don and many	have pointed out, much is going to be done and many	have pointed out, much is ogoing to be onne anl m any
	1 night in a junior suite including buffet breakfast.	1 int nt in a jour-ior isite in=lyding feuf Pradjst.l.	1 night in a junior suite including buffet breakfast.	1 night in a junior suite including buffet breakfast.
	hotel has 52 rooms equipped with complete	hot l ll/as 57 rooms ged pfwlal with ccetein	hotel has 52 rooms equipped with complete	hote! has 52 rooms equipped with complete
	[Teacher Training University], over 6,000 jobs,	!Teacher- ktidliff(Fin fejiugy/, over, 6Ny13/igjll,	[Teacher Training University], over 6,000 jobs,	Teacher Training University], over 6,000 joks,
	named Planken, is also only a few minutes walk from the	named Pardiar, tic orlos ny a a a fe ymrwnrr4N U' nrethe	named Planken, is also only a few minutes walk from the	named Plan\$an, ii a lso only a few reaintee vsdlak from the
	some of the comments made at that time, even though the	some if the commentranale at ttrr t time.ine, tJurthuhtlr the	some of the comments made at that time, even though the	come of the comments made at that time, even though in the
Step-15000	A banquet hall and conference room are available on request.	A bagquet hall and conference room are avai:vh/e on request.	A banquet hall and conference room are available on request.	1 banquet hall and conference room are available on request.
	Grossetête recommendation for second reading (A5-008-0/1999)	GrosseVtfe- recn:nme- dantnn for sencnd 0adada /k5-0P6/03/15)	Grossetête recommendation for second reading (A5-008-0/1999)	Grossatete t2.L:lr me- ndation for second reading (A5.008 .if:if918)
	who have multiple accounts (used maybe in different	who have r nttiple accounts (usee rraybe in diffirerat	who have multiple accounts (used maybe in different	whc' have multiple accourts (used maybe in different
	of the Council, you have not really been very	of the Coundi, you have not really been verv	of the Council, you have not really been very	of the Council, you have not really been very
	These differences matter.	These Vttrreences matter.	These differences matter.	Thhese diffrr'ences matter.
	affect conditions of competition and affect the	affect concon itin of comonettioa e and aficlt the	affect conditions of competition and affect the	affer: ctitndions of competition and affect the

Table 12: Comparison examples of text images generated by word-level and character-level text transformers. Within each cell, the left column displays the ground truth, and the right column shows the generated images.

Reference Images	Generated Text Images using Reference Text					
	Training Steps=20000			Training Steps=30000		
Codebook	1024	2048	4096	1024	2048	4096
	Reference Text: <i>We await the outcome with great interest.</i>					
We await the outcome with great interest.	We await the outcome with great interest.	We await the outcome with great interest.	We await the outcome with great interest.	We await the outcome with great interest.	We await the outcome with great interest.	We await the outcome with great interest.
	Reference Text: <i>I am glad that this point is made in the report.</i>					
I am glad that this point is made in the report.	I am glad that this point is made in the report.	I am glad that this point is made in the report.	I am glad that this point is made in the report.	I am glad that this point is made in the report.	I am glad that this point is made in the report.	I am glad that this point is made in the report.
	Reference Text: <i>The principles are totally unchanged.</i>					
The principles are totally unchanged.	The principles are totally unchanged.	The principles are totally unchanged.	The principles are totally unchanged.	The principles are totally unchanged.	The principles are totally unchanged.	The principles are totally unchanged.
	Reference Text: <i>Other proposals will have to be taken into consideration.</i>					
Other proposals will have to be taken into consideration.	Other proposals will have to be taken into consideration.	Other proposals will have to be taken into consideration.	Other proposals will have to be taken into consideration.	Other proposals will have to be taken into consideration.	Other proposals will have to be taken into consideration.	Other proposals will have to be taken into consideration.
	Reference Text: <i>This is a healthy and sensible approach.</i>					
This is a healthy and sensible approach.	This is a healthy and sensible approach.	This is a healthy and sensible approach.	This is a healthy and sensible approach.	This is a healthy and sensible approach.	This is a healthy and sensible approach.	This is a healthy and sensible approach.
	Reference Text: <i>Die erste Phase wird derzeit durchgeführt.</i>					
Die erste Phase wird derzeit durchgeführt.	Die erste Phase wird derzeit durchgeführt.	Die erste Phase wird derzeit durchgeführt.	Die erste Phase wird derzeit durchgeführt.	Die erste Phase wird derzeit durchgeführt.	Die erste Phase wird derzeit durchgeführt.	Die erste Phase wird derzeit durchgeführt.

Table 13: Examples of generated reference text images for different ViT-VQGAN tokenizers.

Ground Truth		Results of UMTIT		
Source Image	Target Image	Source Text Recognition	Target Text Translation	Image Generation
<i>Good Cases (DE2EN)</i>				
Wir erwarten die Ergebnisse mit großem Interesse.	We await the outcome with great interest.	Wir erwarten die Ergebnisse mit großem Interesse.	We expect the results with great interest.	We expect the results with great interest.
Kernkraft ist keine Lösung.	Nuclear power is no solution.	Kernkraft ist keine Lösung.	Nuclear power is not a solution.	Nuclear power is not a solution.
Dieses Problem darf einfach nicht so hingenommen werden.	That is a situation which simply cannot be tolerated.	Dieses Problem darf einfach nicht so hingenommen werden.	This problem simply cannot be accepted.	This problem simply cannot be accepted.
Das Parlament hat eine wichtige Rolle zu spielen.	Parliament has a major role to play.	Das Parlament hat eine wichtige Rolle zu spielen.	Parliament has an important role to play.	Parliament has an important role to play.
Das ist heutzutage nicht mehr der Fall.	Now this no longer holds true.	Das ist heutzutage nicht mehr der Fall.	This is no longer the case today.	This is no longer the case today.
<i>Good Cases (EN2DE)</i>				
Thank you very much, Commissioner Fischler.	Vielen Dank, Herr Kommissar Fischler.	Thank you very much, Commissioner Fischler.	Vielen Dank, Herr Kommissar Fischler.	Vielen Dank, Herr Kommissar Fischler.
Parliament has a major role to play.	Das Parlament hat eine wichtige Rolle zu spielen.	Parliament has a major role to play.	Das Parlament hat eine große Rolle zu spielen.	Das Parlament hat eine große Rolle zu spielen.
<i>Bad Cases (DE2EN)</i>				
Wir teilen sie mit allen Menschen auf dieser Erde.	We share them with everyone on this earth.	Wir teilen sie mit allen Menschen auf dieser Erde.	We share them with all of them.	We share them with all of them. 1 f them.
An den Prinzipien hat sich nichts geändert.	The principles are totally unchanged.	An den Prinzipien hat sich nichts geändert	In principles, nothing has changed.	In principles, nothing has changed. changed.
<i>Bad Cases (EN2DE)</i>				
Thank you very much, Commissioner.	Vielen Dank, Frau Kommissarin.	Thank you very much, Commissioner.	Vielen Dank, Herr Kommissar.	Vielen Dank, Herr Kommissar. Kommissar.

Table 14: Examples for Multimodal Text Image Translation.