In-Image Machine Translation.
A Preliminary Modular Approach

Sergio Gómez González
Universitat Politècnica
de València
sgomgon@prhlt.upv.es

Miguel Domingo
Universitat Politècnica
de València,
ValgrAI - Valencian Graduate
School and Research Network
for Artificial Intelligence
midobal@prhlt.upv.es

Francisco Casacuberta
Universitat Politècnica
de València,
ValgrAI - Valencian Graduate
School and Research Network
for Artificial Intelligence
fcn@prhlt.upv.es

Abstract

In-image machine translation is a sub-task of Image-Based Machine Translation that aims to substitute text embedded in images with its translation into another language. In the current work, we define a simple task with a synthetic dataset based on rendering parallel text over a plain background. Furthermore, we experiment with different optical character recognition, machine translation and image synthesis models to include in our ensemble. Then, we present our cascade approach as a pipeline that obtains the transcript of the original image, translates it, and generates a new image (image synthesis) similar to the original one. Finally, we compare the performance of our approach with several current state of the art models, including an end-to-end approach, demonstrating its competitiveness.

1 Introduction

Machine translation (MT) is a traditional application of machine learning that allows the translation of text from one language to another, ensuring that a larger community can understand the information (Wang et al., 2022). However, traditionally, the field has been tied to text (understood as a sequence of characters contained in a vocabulary) as the primary container of language (Stahlberg, 2020), except for some research, mainly in audio (Ma et al., 2024; Radford et al., 2023; Barrault et al., 2023). Nevertheless, most of the information produced and received by humans is visual; yet it contains text embedded in visual patterns. Additionally, the textual information contained in those images is understandable only to communities that speak the language. In-image machine translation (IIMT) is a recent sub-field of MT dedicated to producing visual translations of text in images (Tian et al., 2025). Thus, more people can access the information codified in the text embedded in the images.

The task of IIMT is not intrinsically multimodal, since both its input and output are images. However, text is usually incorporated at some point in IIMT models (Lan et al., 2024; Ma et al., 2022). Actually, the IIMT task can be viewed as a single challenge, with a slight use of text, or it can be split into several sub-tasks. The second option is most evidently multimodal, since text is usually involved directly at some step of the pipeline.

In our approach[1], we perform IIMT as a combination of optical character recognition (OCR), MT and image generation in a cascade manner. First, an OCR model must detect and recognize text embedded in the image. Then a MT model would translate that text into the target language. Finally, the image generation model may create an image that is visually similar to the original but contains the translated text.

Our experimentation will focus on a simplified set as a first approach. Thus, we will translate images with a plain background and horizontal text in English and German.

In the following we present a summary of the current state of the art for IIMT. After that, we propose our experimental framework in section 3, where we detail our dataset (section 3.1), approach to the task (section 3.2) and evaluation protocol (section 3.3). Finally, we present our results and compare our approach with an end-to-end model in section 4, and draw some conclusions in section 5.

2 State of the Art

Image-based machine translation (IMT) surges under the field of multimodal machine translation (MMT) to produce translations from text embedded in visual media (Elliott and Kádár, 2017; Gao et al., 2025). Furthermore, IMT is often divided into two additional sub-fields: text-image machine translation (TIMT) and IIMT. The former produces

[1] github.com/sergiogg-ops/modular_i2imt

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 4: Student Research Workshop, pages 502–513
March 26, 2026 ©2026 Association for Computational Linguistics

textual translations from text in images (Lan et al., 2025; Ma et al., 2023). The latter, the one we address in this article, is dedicated to editing the image (Lan et al., 2024). As a result, the desired output is an image visually similar to the original, including translation of the original text instead of it (Tian et al., 2025). This task is often addressed using a cascade approach (Vaidya et al., 2025; Lan et al., 2023; Liang et al., 2024) or an end-to-end one (Lan et al., 2024; Ma et al., 2022).

In most cascade approaches, the first step involves detecting and recognizing the text present in the source image. Most OCR models focus on pictures of documents (Li et al., 2023; Cheng et al., 2017). Nevertheless, in IIMT, natural scenes are a more relevant use case. These images usually contain less text with a more complex layout. The mentioned scenes involve text with complex backgrounds, variable lighting, distortions, arbitrary orientations, and diverse fonts. Thus, OCR models that cope with these conditions have been trained on datasets that recreate them (Gupta et al., 2016; Veit et al., 2016; Wang et al., 2011). Furthermore, renowned computer vision (CV) conferences host natural scene challenges (Karatzas et al., 2015) to push the current state of the art.

The same cascade approaches, after obtaining the transcription, use a MT model to obtain its translation (Qian et al., 2024). MT is a well established field with a huge availability of models (Hameed and Al-Khateeb, 2025) and datasets (Koehn, 2005; Tiedemann and Thottingal, 2020). In addition, conferences such as *WMT* (Chatterjee et al., 2019; Farajian et al., 2020; Wang et al., 2024) promote research in several open challenges in this field. The large language model (LLM) tide has also arrived to MT, even with the support of institutions such as the European Union (Martins et al., 2025).

In most cascade-based systems, the pipeline ends with an image synthesis stage. This area is currently dominated by diffusion transformer-based generative models (Ahsan et al., 2025), with a strong emphasis on multimodality and interactive workflows (Zhang et al., 2023a; Quan et al., 2024). Open-source models such as *Stable Diffusion* (Podell et al., 2023), *Flux*[2] and *DALLE 2* (Ramesh et al., 2022) now rival closed-source systems, including *DALLE 3*[3], *Sora*[4] and the *Banana*[5] family in terms of visual quality and controllability.

One of the main challenges of IMT is the availability of data. Although there are some limited datasets for TIMT, such as *OCRMT30k* (Lan et al., 2023), *DoTA* (Liang et al., 2024) or *ECOIT* (Zhu et al., 2023); the scarcity of data exacerbates in the IIMT field. To the best of our knowledge, there is no non-synthetic, curated dataset for the entire IIMT task. All of the articles we have read so far have used their own synthetic datasets to train and evaluate their models. Some of these sets are available (Lan et al., 2024) for general use.

In this work, we present a detailed experiment on the construction of a pipeline of models to solve an IIMT task. We provide a measure on how the current top models susceptible to being involved in this type of pipeline (Li et al., 2023; Wei et al., 2025; Costa-jussà et al., 2022; Team et al., 2025; Cheng et al., 2025; Tuo et al., 2023) perform in their respective tasks. Finally, our results are compared with *Translatotron-V* (Lan et al., 2024), which will serve as a baseline.

3 Experimental Framework

In this section, we will describe the framework surrounding our experiments and the functioning of our system.

3.1 Dataset

In order to evaluate our pipeline, we needed a dataset that fit our requirements. A dataset that allowed us to evaluate every module of our cascade approach. Since the field of IMT suffers from data scarcity (Li et al., 2025), we needed to create our own synthetic dataset.

In order to generate a synthetic dataset composed of images containing text, a source of text is needed. Therefore, we have selected the *Open Subtitles*[6] (Tiedemann, 2016) English-German parallel dataset as the source. It is an MT dataset gathered from a collection of movie subtitles and their translations. For the background, we used plain images of a single color picked at random.

To create a pair of images, two parallel sentences are extracted from the corpus and a background color is chosen. Furthermore, we select a font style and size for the text display. As suitable options for the display, we selected a collection of *Open Sans* fonts downloaded from *Google Fonts*[7]. In

[2] bfl.ai/blog/24-08-01-bfl

[3] openai.com/es-ES/index/dall-e-3/

[4] openai.com/es-ES/sora/

[5] gemini.google/es/overview/image-generation

[6] www.opensubtitles.org/

[7] github.com/googlefonts/opensans

addition, the pair of bounding boxes and the pair of text for each line are also stored. Thus, we have created a parallel IIMT dataset of 3000 pairs suitable for evaluating both end-to-end and module-wise methods. All the code used in this step will be made available in our repository[8].

3.2 Cascade approach

We aim to solve the IIMT problem by splitting it into several sub-tasks. First, we intend to obtain a transcript from the original image by using an OCR model. Then, this text will be fed into a MT model to obtain its translation into the target language. Finally, we propose using a diffusion model (Nguyen-Tri et al., 2025) to replicate the original image with the translated text in it.

3.2.1 Optical character recognition

In order to obtain a transcript from an image, we have selected three candidate open source approaches from the current state of the art:

EasyOCR[9] is available as a *Python* package with an easy-to-use, straight-forward interface. It uses a different set of weights depending on the selected language, meaning that it is not inherently multilingual. The results may vary depending on the inference language. Thus, if multilingual input is expected, this option may be problematic.

TrOCR (Li et al., 2023) is also available as a *Python* package. It is a more sophisticated approach focused on OCR for documents with a complex layout. Thus, its outputs are harder to parse than *EasyOCR*. Nevertheless, it is inherently multilingual and, thus, appropriate for a real world application in which languages can be mixed. It was developed by *Microsoft*.

DeepSeek-OCR[10] (Wei et al., 2025) is an OCR encoder-decoder model with 3 billion parameters that uses a special encoder to manage long, two-dimensional contexts. This makes it suitable for recognizing long documents. Furthermore, its decoder is the *DeepSeek* model (Liu et al., 2024a,b), which is multimodal and multilingual.

3.2.2 Machine translation

Once we have obtained the text in the original language, it must be translated into the target language. For that purpose, we selected four additional MT models from the current state of the art:

NLLB-200[11] (Costa-jussà et al., 2022) is a multilingual MT model obtained from the *No Language Left Behind* project by *Meta AI*. It can manage up to 210 languages from around the world with different alphabets. We have used the version with 3.3 billion parameters.

Seed[12] (Cheng et al., 2025) constitutes an open-source MT model that is competitive with closed-source models developed by the *ByteDance* lab. However, its language span is restricted to "just" 28 languages, with a size of 7 billion parameters.

Gemma 3[13] (Team et al., 2025) is a family of LLM developed by *Google DeepMind*, renowned for their strong performance in several languages. It is not an MT model but a generalistic LLM; however, we can use it for translation with the appropriate prompt.

3.2.3 Image generation

To generate the final image, we have made use of the *AnyText* model (Tuo et al., 2023). It is a diffusion transformer (Nguyen-Tri et al., 2025) with a *ControlNet* (Zhang et al., 2023b) specially designed for image generation and editing. Based on an original image, a mask and a prompt, it can change the text in the masked parts of the original image to that contained in the prompt. It allegedly maintains the style of the text in the original image, which is quite convenient for our task.

3.2.4 Assembly

Since our aim is to perform IIMT, each of the systems described previously must fit into a pipeline to solve the desired task. Thus, we need to make the outputs of some models compatible with the inputs of others. In our work, we have identified two concepts that are essential and shared across subtasks: bounding boxes and the text they contain. Together with the image, they completely define the task we are working on. The former are defined by the four points that form a rectangle around a piece of text

[8] github.com/sergiogg-ops/modular_i2imt

[9] github.com/JaidedAI/EasyOCR

[10] huggingface.co/deepseek-ai/DeepSeek-OCR

[11] huggingface.co/facebook/nllb-200-3.3B

[12] huggingface.co/ByteDance-Seed/Seed-X-PPO-7B

[13] huggingface.co/google/gemma-3-4b-it

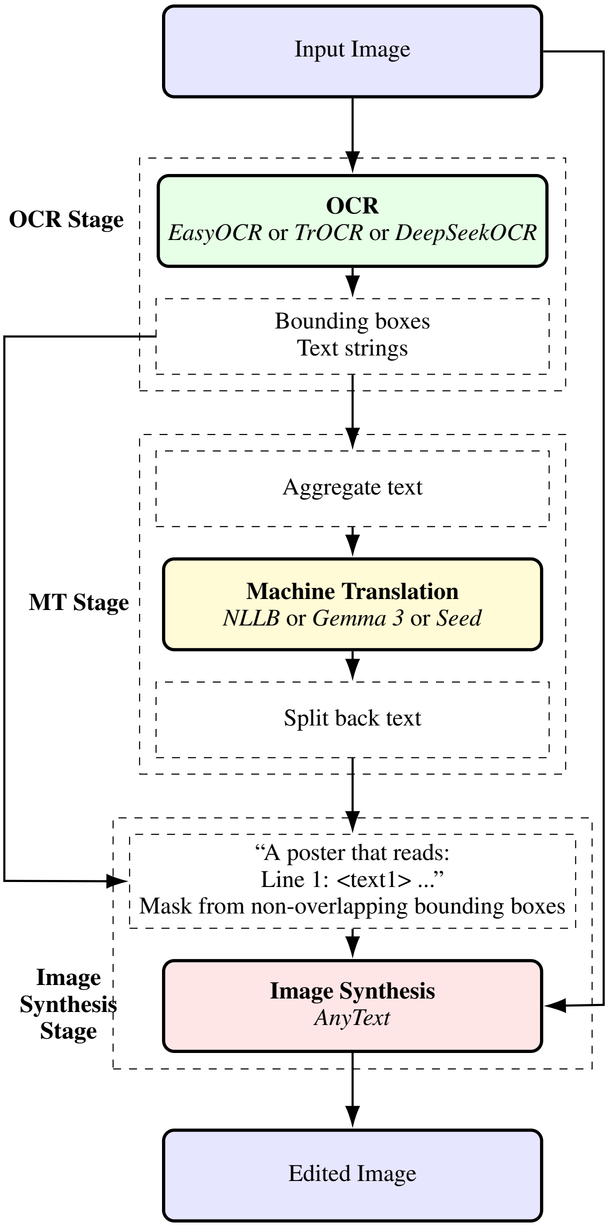


Figure 1: Proposed cascade approach.

in an image. Since our task is limited to horizontal text, they are axis-aligned rectangles. The text is stored as a common string. The bounding boxes and text are always attached to the image in a metadata file. In each step, the content is modified but the metadata structure itself remains unchanged.

As stated before and illustrated at fig. 1, our pipeline starts with the extraction of the transcription (OCR), its posterior translation MT and ends with the generation of the new image. In the OCR step, bounding boxes are extracted along with the text they contain. Depending on the model, the output is modified to be adapted to our four-point format. We maintain the reading order provided by the respective OCR model.

Thus, in the MT stage, the text of the different boxes is aggregated and fed into the MT model. After that, it is split again to fill each of the original bounding boxes without explicitly aligning with the original content.

Once the MT step is finished, the image synthesis stage can start. The *Anytext* model needs a mask to edit the original image. For that, overlaps between bounding boxes are removed and the remaining of the bounding boxes is set to black over a white background. Furthermore, the text is codified in the prompt format of appendix A, and is provided along with the mask and the original image to the *Anytext* model. As a result, we obtain the edited image that now contains the translation of the original text.

The different prompting strategies that we used can be found at appendix A.

3.3 Evaluation

Since our approach to this problem is to split it into several subtasks, we first need to evaluate the different modules. Thus, we will be able to select the best strategies for each specific task to obtain the optimal overall model. For that evaluation, we will use some metrics to determine the models' affinity for the task. Furthermore, we will use approximate randomization testing (ART) (Riezler and Maxwell, 2005) to determine whether the resulting differences in the metric scores are statistically significant.

Therefore, in this section, we will explain the evaluation strategies followed to analyze the performance of the different OCR, MT and image generation approaches. Finally, our proposal to evaluate the aggregated model will be described.

3.3.1 Optical character recognition

The first step in our approach is to obtain the transcript of the input image while preserving part of the layout. Precisely, we evaluate the similarity of the transcript with respect to the reference and the location of the bounding boxes. For that, we use Word Error Rate (WER), bag-of-words WER (bWER) and intersection over union (IOU):

WER (Morris et al., 2004): it is obtained from the *Levenshtein* or edit distance from the transcript to the reference at the word level.

bWER (Vidal et al., 2023): computes WER in a bag-of-words manner instead of an ordered list of words.

Model	English			German		
	WER (↓)	bWER (↓)	IOU (↑)	WER (↓)	bWER (↓)	IOU (↑)
EasyOCR	15.8	**12.0**†	61.7	**8.5**	**7.0**	65.5
TrOCR	**13.8**	**13.3**†	59.1	16.7	16.5	56.9
DeepSeek-OCR	40.1	40.1	**67.9**	53.1	53.1	**66.0**

Table 1: Evaluation scores for the OCR modules of the pipeline. Best results are reported in **bold**; all of the differences are statistically significant except the ones between scores marked with †.

Model	German to English			English to German		
	BLEU (↑)	TER (↓)	ChrF (↑)	BLEU (↑)	TER (↓)	ChrF (↑)
NLLB-200	**35.3**	**60.5**	**52.9**	**27.8**	**69.5**	**50.7**
Gemma 3	25.9	66.5	46.4	21.3	74.6	45.8
Seed	10.0	174.7	38.7	1.6	719.9	19.4

Table 2: Evaluation scores for the MT modules of the pipeline. Best results are reported in **bold**. All of the differences are statistically significant.

IOU (Rezatofighi et al., 2019): relation of the intersection with the union of the bounding box hypothesized and its reference. Used to evaluate text detection.

As the first step in the pipeline, it is vital that these models perform as well as possible. Otherwise, the errors will spread and it will be difficult to correct them in subsequent steps.

3.3.2 Machine translation

In the second step, the transcript should be translated into the target language. Otherwise, the final image would contain text in the original language. Furthermore, the translation model can correct any recognition errors from the previous step. To measure the performance of MT models, we use bilingual evaluation understudy (BLEU), translation error rate (TER) and F-score based on character n-grams (chrF):

BLEU (Papineni et al., 2002): based on the computation of the average of the modified n-gram precision. It is also normalized by a brevity factor that penalizes short sentence results.

TER (Snover et al., 2006): score computed by summing the number of word edit operations (insertion, substitution, deletion and swapping). It is normalized by the number of words in the reference.

ChrF (Popović, 2015): metric that applies statistics after the common F-score to assess the similarity between sentences using n-grams.

We have used the implementation of *Sacrebleu*[14] (Post, 2018) to compute these metrics as it is a renowned standard.

3.3.3 Image generation

The last step is the generation of an image with the new translated text. Ideally, it will look exactly like the original image, with the only difference being the text it contains. Even the text should use the same format and occupy a similar space in the image. To measure the similarity of the generated image with the reference, we will use two widely used metrics in computer vision: Fréchet inception distance (FID) and structural similarity index (SSIM).

FID (Heusel et al., 2017): it compares the distribution of features extracted from generated images to those from real images by calculating the Fréchet distance between the means and covariances of these feature sets.

SSIM (Wang et al., 2004): it compares luminance, contrast, and structural information between the generated image and the reference. It measures how well the structural information of the original image is preserved in the generated sample. Its values range from -1 (inverse correlation) to 1 (perfect similarity).

In addition, the evaluation of this last step will require more human supervision for qualitative analysis. In the end, the results of the cascade should

[14] `github.com/MorinoseiMorizo/sacreBLEU`

	German to English					
Model	OCR		MT		Image generation	
	bWER (↓)	IOU (↑)	BLEU (↑)	TER (↓)	SSIM (↑)	FID (↓)
EasyOCR	$\mathbf{7.0}^{\dagger}$	$\mathbf{65.5}^{\dagger}$	–	–	–	–
NLLB-200	–	–	**35.3**	**60.5**	–	–
AnyText	–	–	–	–	**46.1**	**120.4**
Cascade	$\mathbf{7.0}^{\dagger}$	$\mathbf{65.5}^{\dagger}$	32.8	63.2	38.3	147.8

Table 3: Analysis of the error propagation in the system with respect to the predictions of each of the modules separately for the German to English task. Best results are reported in **bold**; all of the differences are statistically significant except the ones between scores marked with †.

	English to German					
Model	OCR		MT		Image generation	
	bWER (↓)	IOU (↑)	BLEU (↑)	TER (↓)	SSIM (↑)	FID (↓)
EasyOCR	$\mathbf{12.0}^{\dagger}$	$\mathbf{61.7}^{\dagger}$	–	–	–	–
NLLB-200	–	–	**27.8**	**69.5**	–	–
AnyText	–	–	–	–	$\mathbf{47.6}^{\dagger}$	$\mathbf{113.4}^{\dagger}$
Cascade	$\mathbf{12.0}^{\dagger}$	$\mathbf{61.7}^{\dagger}$	25.5	73.7	$\mathbf{47.9}^{\dagger}$	$\mathbf{113.3}^{\dagger}$

Table 4: Analysis of the error propagation in the system with respect to the predictions of each of the modules separately for the English to German task. Best results are reported in **bold**; all of the differences are statistically significant except the ones between scores marked with †.

contain the required information while also being harmonious and pleasing to the eye.

4 Results

In this section, we will discuss the evaluation of the predictions made by both the modules and the cascade system over the synthetic dataset.

4.1 Optical character recognition

After the evaluation, it becomes clear that the best text recognizer is the one offered by *EasyOCR*. Scores are available at table 1. Even if its performance in English is slightly surpassed by *TrOCR*, *EasyOCR* is decisively superior for German. Actually, only the WER score is worse, while the bWER score is similar. This could imply that the differences in WER must be related to the ordering of the recognized words. However, the text recognizer of *DeepSeek-OCR* is far worse than the others due to problems with the spacing of the words.

In contrast, the best text detector is *DeepSeek-OCR*. The IOU scores from table 1 show that it is consistently and significantly superior to the other text recognizers that we have studied.

4.2 Machine translation

Among the MT modules, *NLLB-200* is the one with the best performance among those that we have used. This is shown by the scores displayed at table 2. It is superior to any other in all metrics for both translation directions. Despite the translations provided by *Gemma 3* being readable and useful to some extent, they are of a lower quality than those produced by *NLLB-200*. The model *Seed* has offered feasible translations that do not correspond to the references. Thus, the metric computation has penalized *Seed* since it is restricted to a single reference.

4.3 Image generation

For this subtask, we have performed image editing by replacing the original text with their translations. We have provided the original bounding boxes and the text of the source image to the model. As a result, we obtained images that should be similar to the references in our dataset. After that, we evaluated them using SSIM and FID, obtaining the results shown at tables 3 and 4. The scores of both metrics are better for English text editing than for German text editing.

However, as stated at section 3.3, this task re-

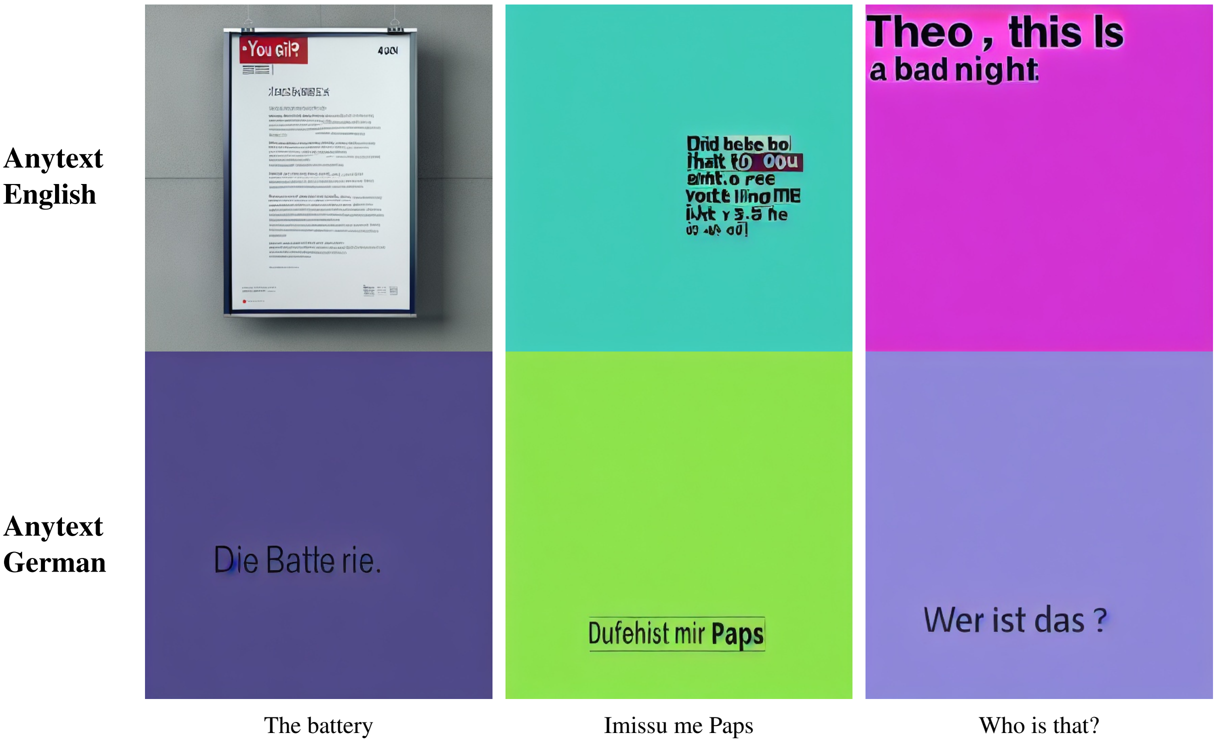


Figure 2: Example images synthesized using the *Anytext* (Tuo et al., 2023) from the intermediate references.

quires a thorough and qualitative analysis. Thus, we observed the synthesized images and detected five major cases, represented in fig. 2:

1. The first category corresponds to satisfactory images. They contain legible text with a similar style and an adequately reconstructed background.

2. Sometimes, images contain legible text on a clean background, but the lines of text are not arranged in reading order.

3. In other images, the model has successfully extracted the original text and reconstructed their backgrounds. However, it fails to write the new text, generating a cumulus of unreadable script-like symbols. Sometimes, it writes actual characters from a random alphabet.

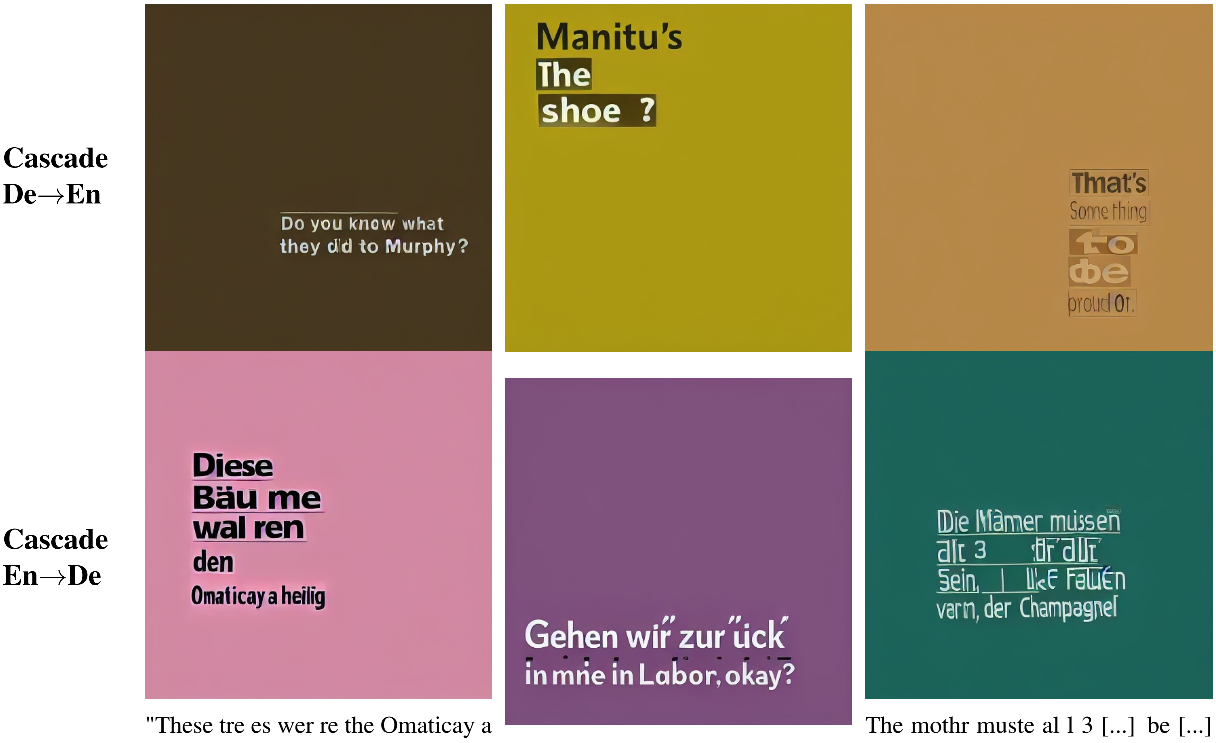
4. For some other images, the model behaves in the opposite manner: it produces legible text but cannot manage to reproduce the background without artefacts. The most common artefacts take the shape of a frame over the text.

5. Finally, the model hallucinates for some images, generating content that is irrelevant to the original image. Nevertheless, this is the rarest case, and we have detected it only a few times.

4.4 Cascade system

After applying the cascade model and evaluating the intermediate results, we can state that the cascade of models suffers from some error propagation. The scores are presented at tables 3 and 4. Intermediate translations are of lower quality than those obtained from the references for both translation directions. For the German to English direction, the error continues to propagate to the last step of the pipeline. The similarity between the images synthesized with the pipeline and the reference is significantly lower than that of the experiment of section 4.3. Nevertheless, in the opposite direction, the results of the pipeline are very similar to those of the aforementioned experiment.

Analyzing the synthesized images, we have observed the same phenomena as in section 4.3 as we show in fig. 3. The cases described in that section are inherited by the cascade model from the *Anytext* module. Additionally, the text in those images can be slightly different, but they are usually feasible translations when legible. However, the main difference is that, due to error propagation, there are more images with artifacts or unreadable text than in the experiment of section 4.3.


Cascade De→En
Do you know what they d'd to Murphy?
Manitu's The shoe ?
That's Some thing to be proud of.
Cascade En→De
Diese Bäu me wal ren den Omaticay a heilig
Gehen wir zur'uck in mne in Labor, okay?
Die Mämer mussen alt 3 für alt Sein, | like Falten varn, der Champagnel

"These tre es wer re the Omaticay a sacred"

Shall we go ba ck to m y in lab okay?

The mothr muste al l 3 [...] be [...] drivn, the Champagnel

Figure 3: Example images synthesized using our cascade of models approach.

Other current state-of-the-art models have been tested on our dataset, such as *Translatotron*. The model of Lan et al. (2024) is an end-to-end approach that aims to reduce the model's size and avoid error propagation. Compared with our cascade approach in table 5, both seem to perform at a similar level. Taking into account the SSIM score in section 2, our approach preserves more similarity with the original image. Furthermore, the text is more precisely placed in the produced images as the IOU scores of their text state. However, this score along with bWER have been obtained from transcripts of the images produced by the models. As a result, those scores might include some additional error from the OCR model, reporting a more pessimistic result. The bWER scores for both systems are characteristic of poor performance.

The images generated by *Translatotron-V* are often empty of text. The model successfully removes the original text, but struggles to generate the new one. This lack of text is probably what heavily penalizes its IOU score. In the images that we have studied, we have detected a majority of empty images but also some correctly translated images. This phenomenon, along with the results reported by Lan et al. (2024) in their work, inclines us to think that the model might be over-fitted to its original task.

Model	FID	IOU	bWER
Translatotron-V	133.3	1.8	**104.6**
Cascade	**113.3**	**24.4**	112.2

Table 5: Comparison with *Translatotron-V* (Lan et al., 2024) on the English to German task. All scores are computed from the images produced by both models. The bounding boxes to compute IOU were obtained with the *DeepSeek-OCR* model (Wei et al., 2025). The transcripts for bWER and BLEU computation were obtained with the *EasyOCR* library. Best results are reported in **bold**; all of the differences are statistically significant.

5 Conclusions

This work presented a preliminary modular framework for IIMT, which addresses the absence of suitable datasets through the creation of a controlled synthetic benchmark for English–German text. By decomposing IIMT into OCR, MT, and image synthesis, we evaluated state-of-the-art pretrained models in isolation and in combination, allowing us to analyze their interactions and the propagation of errors across the pipeline.

Despite error propagation, the proposed cascade demonstrated performance comparable to the end-to-end *Translatotron-V* baseline, particularly in layout preservation and structural similarity. This suggests that modular approaches remain viable, especially when leveraging high-quality pre-trained components.

Future work should focus on improving text-aware image generation and developing evaluation metrics that better capture textual fidelity, typographic consistency, and perceptual similarity. Expanding beyond simple synthetic scenes toward more realistic image conditions will be essential for assessing generalizability. Furthermore, as newer, more powerful models are developed, it will be necessary to integrate them into the pipeline and assess their performance. In general, this study establishes a transparent baseline for modular IIMT and provides the foundation for more robust hybrid or end-to-end systems.

Limitations

As the title of our article states, our research is limited to a preliminary approach to a complete IIMT system. Thus, our proposal has room for improvement and we intend to iterate through it to achieve better performance. In fact, we have used the modules of the pipeline "as they are" without any adjustments to the task, other than the prompts. We would like to experiment with some model tuning to improve performance.

Our task has been limited to images that contain horizontal text with a plain background. Seeing that the font style and size, the background color and the text location were mutable; any other variable has remained constant. In further iterations, we plan to expand this test set with a more challenging environment.

Finally, the evaluation of the generated images is still experimental. It is not clear how to automatically evaluate the similarity of text styles yet. Despite FID and SSIM metrics being standard in general image evaluation, the fact that our task is text-oriented should imply a greater importance of the similarity of text style and its location in the image. Therefore, the development of a complete evaluation strategy is one of our future steps.

Acknowledgements

This work received funding from *ValgrAI* (*Valencian Graduate School and Research Network for Artificial Intelligence*) and *Generalitat Valenciana*.

References

Md Manjurul Ahsan, Shivakumar Raman, Yingtao Liu, and Zahed Siddique. 2025. A comprehensive survey on diffusion models and their applications. *Applied Soft Computing*, page 113470.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, and 1 others. 2023. SeamlessM4T: massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, and 1 others. 2025. Seed-x: Building strong multilingual translation llm with 7b parameters. *arXiv preprint arXiv:2507.13618*.

Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. 2017. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5076–5084.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141.

M Amin Farajian, António V Lopes, André FT Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the WMT 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 65–75.

Yue Gao, Jing Zhao, Shiliang Sun, Xiaosong Qiao, Tengfei Song, and Hao Yang. 2025. Multimodal machine translation with text-image in- depth questioning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9274–9287.

Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2315–2324.

Diadeen Ali Hameed and Belal Al-Khateeb. 2025. Deep learning techniques for machine translation: A survey. *Procedia Computer Science*, pages 1022–1037.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. pages 6629—-6640.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, and 1 others. 2015. ICDAR 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, Min Zhang, and Jinsong Su. 2024. Translatotron-V(ison): An end-to-end model for in- image machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5472–5485.

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. *arXiv preprint arXiv:2305.17415*.

Zhibin Lan, Jiawei Yu, Shiyu Liu, Junfeng Yao, Degen Huang, and Jinsong Su. 2025. Towards better text image machine translation with multimodal codebook and multi-stage training. *Neural Networks*, page 107599.

Bo Li, Shaolin Zhu, and Lijie Wen. 2025. MIT-10M: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13094–13102.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024b. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Cong Ma, Xu Han, Linghui Wu, Yaping Zhang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. Modal contrastive learning based end-to-end text image machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 2153–2165.

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 *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1664–1670.

Zhengrui Ma, Qingkai Fang, Shaolei Zhang, Shoutao Guo, Yang Feng, and Min Zhang. 2024. A non-autoregressive generation framework for end-to-end simultaneous speech-to-any translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1557–1575.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, pages 53–62.

Andrew Cameron Morris, Viktoria Maier, and Phil D Green. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech*, pages 2765–2768.

Quan Nguyen-Tri, Cong Dao Tran, and Hoang Thanh-Tung. 2025. Diffusion directed acyclic transformer for non-autoregressive machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 814–828.

Kishore Papineni, Salim Roukos, Todd Ward, and Zhu Wei-Jing". 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Zhipeng Qian, Pei Zhang, Baosong Yang, Kai Fan, Yiwei Ma, Derek F. Wong, Xiaoshuai Sun, and Rongrong Ji. 2024. AnyTrans: Translate AnyText in the image with large scale models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2432–2444.

Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. 2024. Deep learning-based image and video inpainting: A survey. *Int. J. Comput. Vision*, pages 2367—2400.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, page 3.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666.

Stefan Riezler and John T Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Felix Stahlberg. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, pages 343–418.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Yanzhi Tian, Zeming Liu, Zhengyang Liu, and Yuhang Guo. 2025. Exploring in-image machine translation with real-world background. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 124–137.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT–building open translation services for the world. In *Annual Conference of the European Association for Machine Translation*, pages 479–480.

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*.

Shreyas Vaidya, Arvind Kumar Sharma, Prajwal Gatti, and Anand Mishra. 2025. Show me the world in my language: Establishing the first baseline for scene-text to scene-text translation. In *Pattern Recognition*, pages 312–328.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*.

Enrique Vidal, Alejandro H. Toselli, Antonio Ríos-Vila, and Jorge Calvo-Zaragoza. 2023. End-to-end page-level assessment of handwritten text recognition. *Pattern Recognition*, page 109695.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2022. Progress in machine translation. *Engineering*, pages 143–153.

Kai Wang, Boris Babenko, and Serge Belongie. 2011. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464.

Longyue Wang, Siyou Liu, Chenyang Lyu, Wenxiang Jiao, Xing Wang, Jiahao Xu, Zhaopeng Tu, Yan Gu, Weiyu Chen, Minghao Wu, Liting Zhou, Philipp Koehn, Andy Way, and Yulin Yuan. 2024. Findings of the WMT 2024 shared task on discourse- level literary translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 699–700.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, pages 600–612.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. 2023a. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*.

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.

Shaolin Zhu, Shangjie Li, Yikun Lei, and Deyi Xiong. 2023. PEIT: Bridging the modality gap with pre-trained models for end-to-end image translation. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13433–13447.

A Prompting

Below, the reader can find the prompting strategies used for each model that required it:

A.1 DeepSeek-OCR

As the authors from Wei et al. (2025) instruct, this model must be conditioned to generate transcripts, and different prompts work better depending on the type of images. For our work, we have used the most appropriate prompt that they advised for scene text recognition:

```
<image>
<|grounding|>OCR this image.
```

A.2 Gemma 3

This model is a multimodal generative LLM that must be conditioned by the prompt to solve each task presented to it. Thus, we have codified the following instruction for it to translate:

```
Translate the following <English/German>
text to <English/German> without further
 comments:
 <sentence>
```

Languages are inverted depending on the translation direction.

A.3 Seed

This translation model needs to be conditioned in the prompt to perform the task. It can also be instructed to explain the translation, aiming for better quality. This *chain of thought* can be easily erased from the model's output. Below, we present the prompt that we have used:

```
Translate the following <English/German>
text to <English/German> and explain it
 in detail:
 <sentence>
```

Languages are inverted depending on the translation direction.

A.4 AnyText

Finally, the image generation diffusion model we used also needs some prompting. One needs to introduce the text that will be displayed between double quotation marks. For example, in a text with N lines:

```
A poster that reads:
Line 1: "<text1>"
Line 2: "<text2>"
...
Line N: "<textN>"
```