

# Benchmarking Visually-Situated Translation of Text in Natural Images

Elizabeth Salesky<sup>J</sup>   Philipp Koehn<sup>J</sup>   Matt Post<sup>M, M</sup>  
<sup>J</sup>Johns Hopkins University  
<sup>M</sup>Human Language Technology Center of Excellence  
<sup>M</sup>Microsoft  
esalesky@jhu.edu

## Abstract

We introduce a benchmark, VISTRA, for visually-situated translation of English text in natural images to four target languages. We describe the dataset construction and composition. We benchmark open-source and commercial OCR and MT models on VISTRA, and present both quantitative results and a taxonomy of common OCR error classes with their effect on downstream MT. Finally, we assess direct image-to-text translation with a multimodal LLM, and show that it is able in some cases but not yet consistently to disambiguate possible translations with visual context. We show that this is an unsolved and challenging task even for strong commercial models. We hope that the creation and release of this benchmark which is the first of its kind for these language pairs will encourage further research in this direction.

## 1 Introduction

Visually-situated language concerns multimodal settings where text and vision are intermixed, and the meaning of words or phrases is directly influenced by what is observable or referenced visually. Vision-and-language research has most commonly focused on tasks where images and text can be processed as distinct channels within a joint model, such as question answering or image captioning. However, settings where text is embedded in an image are ubiquitous, ranging from text on street signs, to chyrons on news broadcasts, language embedded in figures or social media images, or non-digitized text sources.

Translating visually-situated text is a practical application of recent pixel-based translation models (Salesky et al., 2021), with new challenges due to the varied text styles, backgrounds, and complex layouts found in natural images. This task combines a series of traditionally separate steps including text detection, optical character recognition, semantic grouping, and finally machine translation.



Figure 1: Visual context can resolve translation ambiguity. Here, translating ‘EXIT’ from English to German is ambiguous without further information about the mode of travel (on foot or by car), which the visual context in the image provides.

Not only can errors propagate between steps, as generated mistakes cause mismatches in vocabulary and distribution from those observed in training and reduce downstream task performance, but processing each step in isolation separates recognized text from visual context which may be necessary to produce a correct situational translation. For example, as illustrated in Figure 1, the English word ‘Exit’ can be translated to German as either ‘Ausfahrt’ or ‘Ausgang’; without appropriate context, which may not be present in the text alone, the generated translation would be a statistical guess.

We present a publicly-released benchmark, VISTRA, for visually-situated translation (VST) of text contained in natural images. With VISTRA, we benchmark the performance of popular OCR models and conduct an error analysis of text recognition errors. We analyze which recognition errors propagate to and most significantly affect downstream translation to four target languages with varied levels of contextual dependence on the image. We also compare **direct** visually-situated translation with multimodal LLMs, and discuss whether access to visual context improves visually-situated translation with current models. Finally, given our findings, we present directions for future work and connections to recent pixel-based translation models.

## 2 Constructing the VISTRA benchmark

VISTRA comprises 772 natural images containing English text, with aligned translations to four target languages (German, Spanish, Russian, and Mandarin Chinese) with varying levels of visual contextual dependence. Each image is annotated with its height and width, a categorical label, its semantically grouped English transcript, translations to the four target languages aligned at the level of the semantic groups in the transcript, and, word-level bounding boxes specified by corner with coordinates rescaled from 0-1, matched to the aligned word in the transcript. On average, each image contains 11.2 words and 2.4 transcript groups, for a total of 1840 parallel segments in the benchmark with an average length of 4.7 words. An annotated data sample is shown in Figure 2.<sup>1</sup>

To the best of our knowledge, only one prior publicly-released data exists for in-image text translation from *natural* images (OCRMT30K: Lan et al., 2023), which contains 30k images with Chinese text manually translated to English. In the absence of datasets for this task, prior work on in-image machine translation has primarily synthetically rendered MT corpora for this task (Mansimov et al., 2020; Tian et al., 2023; Niu et al., 2024; Lan et al., 2024) or addressed PDF document translation (Ignat et al., 2022; Hsu et al., 2024), discussed further in Section 4. While these settings typically use uniform text styles and sizes and contain a single semantic unit per image, natural images are contain text with multiple sizes and styles, multiple text groups in complex layouts, and varied image backgrounds, all of which introduce additional challenges. Our task also differs from what is commonly called multimodal translation in that our setting text is embedded into the image context, as opposed to a text caption to be translated with the aid of a relevant image.

The VISTRA benchmark is released under a permissive CC BY-SA license for further scientific research and commercial use.<sup>2</sup>

### 2.1 Criteria for image selection

The dataset is primarily constructed of newly-captured photos in order that they not be under copyright or contained in LLM training data.<sup>3</sup> We

additionally include a small challenge set of public domain images from social media where text has been embedded in an image and is no longer accessible without OCR. Within this benchmark, we focus only on printed text, not handwritten. We describe the detailed criteria for image selection below.

1. **Languages:** Only images containing text in a single language (English) are included.
2. **Maximizing translatable text:** Images were chosen to maximize text which would be translated rather than transliterated or copied across languages, i.e. maximizing descriptive or instructive text and minimizing numerals and named entities. Where these are present, they may not constitute the majority of the text.
3. **Framing with sufficient context:** Sufficient context (visual or textual) must be present to reduce translation ambiguity. If, as in Figure 1, correct translation would require knowledge that the sign is by a road or a footpath, one of these should be at least partly visible.
4. **Length of text:** We aim for a balance of text lengths. While some traffic signs may have only 1-2 words, if they are sufficiently frequent that it is important for strong image translation models to get correct, they have been included; other images may include up to 100 words.
5. **Text style:** Text may contain multiple fonts, colors, and sizes within one image.
6. **Layout and number of text groups:** We include a balance of layout complexity, from single-line horizontal layouts, to complex layouts with angled text, or multiple adjacent semantic groups which prove challenging for line-level OCR.
7. **Image dimensions and resolution:** We collect high-resolution photos, non-resized and not retouched. Original dimensions may vary based on camera and conditions, but at least one dimension (length or width) must be larger than 1024.

<sup>1</sup>We omit the full list of bounding box coordinates in Figure 2 for readability.

<sup>2</sup><https://vistra-benchmark.github.io>

<sup>3</sup>Though these specific images will not have been observed in training, we cannot guarantee that the same or similar signs

in other settings have not. Though we submitted opt-out requests to exempt our data from being trained on before submitting benchmark images to commercial LLMs in our experiments, if subsequent researchers do not also do so, benchmark images may be ingested as training data.



```
{
  "image_file": "3c2b0778.png",
  "height": 1024, "width": 768,
  "category": "directional sign",
  "transcript": ["EXIT ONLY", "ONE WAY"],
  "translation":
    "de": ["NUR AUSFAHRT", "EINBAHNSTRABE"],
    "es": ["SOLO SALIDA", "UNA VÍA"],
    "ru": ["ТОЛЬКО ВЫЕЗД", "ОДНОСТОРОННЕЕ ДВИЖЕНИЕ"],
    "zh": ["仅用作出口", "单向"],
  "bounding_boxes": {'EXIT': [[0.4701, 0.2565], ...]},
  "requires_image_context":
    "de":true, "es":true, "ru":false, "zh":true
}
```

Figure 2: VISTRA data sample showing metadata, transcripts, and translations.



8. **Clean conditions:** The dataset reflects clean conditions. We require that it is not challenging for a human reader to recognize contained text. We exclude images where the text is challenging to read due to environmental conditions (such as weather: rain, fog); blur; lighting conditions; occlusions (such as graffiti, foliage).
9. **Permissive use:** All images are either photos taken for the purposes of this benchmark or in the public domain.

## 2.2 Text annotation and transcription

Text bounding boxes and transcripts were manually post-edited from Google Cloud Vision OCR with a custom interface. The annotation interface is shown in Figure 6 in Appendix A.

**Bounding boxes.** Text bounding boxes are specified at the word level. In contrast to line-level annotations, using word-level bounding boxes more flexibly allows for complex layouts where unrelated text may appear side-by-side in an image (for example, adjacent signs), but should not be grouped and translated together. Bounding boxes are rectangular (90° corners) with all four vertices specified, which allows angled rotation to match text directionality. Bounding boxes were post-edited to ensure all text was detected, no text was cropped, and hallucinated text boxes were removed.

**Transcript.** All (and only) text which was clearly human-readable with images resized to a maximum height and width of 1024px was transcribed. In the final transcripts, case and punctuation are matched as closely as possible to what is present in the original image. Non-textual symbols which may be present on some directional signs

(for example,  or ) were not transcribed or annotated.

**Semantic grouping.** Finally, we semantically group word-level text boxes. This creates text units with necessary context for translation, and separates for example different street signs which appear in the same image into distinct units for downstream translation. Not all images contain full sentences; therefore, our criteria were forming clause or phrase-level groups which appear together in the image and should be translated together. This step may be ambiguous, and so was annotated by one person to ensure consistency across the dataset.

## 2.3 Translation

We contracted Centific<sup>4</sup> to professionally translate the text in each image from English to four target languages: German, Spanish, Russian, and Mandarin Chinese. This set of languages covers multiple language families and scripts, and varied dependence on visual context. Annotators were paid a competitive market rate. Each image was translated by an individual linguist and a random sample of 10% of the image translations were checked by a second linguist.

All translations were performed from scratch in OneForma, with access to both the original image and transcript. The translation instructions and annotation interface are shown in Figure 7 in Appendix A. Translations are aligned one-to-one with the semantic groups in the transcript. We do not ask annotators to match case and punctuation in the source language, which may be unnatural for the target language, but rather localize these for the target language.

<sup>4</sup><https://www.centific.com>

Model	OCR	MT	VST	Release	OCR level	Returns bboxes?	Multilingual
PaddleOCR	✓			OPEN-SOURCE	line/word	yes	
TesseractOCR	✓			OPEN-SOURCE	word	yes	
Google Cloud Vision	✓			COMMERCIAL	word	yes	
mBART		✓		OPEN-SOURCE	—	—	✓
Google Translate		✓		COMMERCIAL	—	—	
GPT-4o	✓	✓	✓	COMMERCIAL	unknown	no	✓

Table 1: Models benchmarked for visually-situated translated (cascaded and direct).

Annotators were additionally asked whether the visual context in the image affected the resulting translation, as a binary question. Whether a translation would be ambiguous without the image can vary by target language, as exemplified by ‘Exit’ in Figure 1 which would be ambiguous in German but not in Spanish. 99.7% of images were marked as requiring image context for translation for at least one translation direction, with the following breakdown by target language: German 99%, Chinese 96%, Spanish 54%, Russian 6%.

### 3 Benchmarking visually-situated translation

We benchmark existing models for OCR and VST using the new VISTRA dataset, and conduct an error analysis of common OCR types and their effect on downstream translation. This type of error analysis does not exist in previous work; we show that our new benchmark both illustrates these types of errors and facilitates analysis of this type.

#### 3.1 Models evaluated

We compare a variety of widely-used open-source, open-weight, and commercial models to give a representative view of the capabilities of current models for this task, and specifically, provide baseline performance on the VISTRA benchmark. We list all evaluated models with relevant characteristics in Table 1.

##### 3.1.1 OCR

**Paddle-OCR**<sup>5</sup> (PP-OCR: Du et al., 2020) is becoming one of the most commonly used open-source tools for OCR in English and Chinese (Lan et al., 2023; Yang et al., 2023, *inter alia*), due to its ease of use and free public release. PP-OCRv4 uses

<sup>5</sup><https://paddlepaddle.github.io/PaddleOCR>

Transformer models, trained per-language for English and Chinese. It produces word-level bounding boxes within detected lines. **Tesseract-OCR**<sup>6</sup> (Smith, 2007) is the longest-standing community-developed open-source toolkit for OCR. Tesseract-4 uses LSTM models, trained per-language. We additionally benchmark **Google Cloud Vision OCR**<sup>7</sup> (Popat et al., 2017; Ingle et al., 2019) to compare strong commercial performance.

##### 3.1.2 MT

We compare both an open-source machine translation model, **mBART-50** (Liu et al., 2020), which is trained primarily on clean, well-formed text, and a commercial machine translation model, **Google Translate**, as an upper-bound on expected performance with greater expected robustness to noise.

##### 3.1.3 Multimodal LLMs

Multimodal multilingual large language models which have been explicitly trained on both vision and language provide an opportunity to compare *direct* translation from an image containing English to text in a target language. By directly translating from an image with access to the full image context (as opposed to only the cropped region within bounding boxes from a text detection stage), multimodal models have the potential to be able to resolve ambiguity in translation. Here we benchmark the performance of **GPT-4o**, which was the top-performing multimodal model on a recent OCR-centric LLM evaluation (OCR-Bench: Liu et al., 2023), by prompting the model to directly translate text contained in images without intermediate steps. We also evaluate OCR only and machine translation only with this model in order to contextualize direct multimodal translation results.

<sup>6</sup><https://github.com/tesseract-ocr/tesseract>

<sup>7</sup><https://cloud.google.com/vision>



Class	Description
I	Undetected text: missing text and bounding boxes
II	Text hallucination: text detected where no text present
III	Bounding box misplaced: text clipped, cropping would affect recognition
IV	Grouping error: text from different groups intermixed in output text
V	Punctuation error
VI	Spacing error
VII	Character-level substitution
VIII	Word-level substitution

Table 2: OCR error taxonomy covering text detection (I-III) and recognition (IV-VIII) errors.

### 3.2 OCR performance and error taxonomy

We measure OCR performance with two automatic metrics: character error rate (CER) and translation error rate (TER) (Snover et al., 2006).<sup>8</sup> CER reflects the minimum number of single-character edits (insertions, deletions, or substitutions) required to change a string into the reference. TER is also an edit-distance metric which aims to capture the post-editing effort required to change a string into the reference. While OCR is typically evaluated case-insensitive with punctuation removed, with downstream MT in mind we calculate both metrics case-sensitive with punctuation.

While in e.g., speech recognition there may be one correct ordering of the output, in a 2D image, there is not necessarily only one correct order for recognized text. To facilitate scoring different text groupings across models, which would otherwise require re-alignment, we concatenate groups before applying automatic metrics. Where CER would recognize reorderings between the hypothesis and reference as several character edits, TER allows shifts of contiguous spans as a single operation, and therefore penalizes reordering less. As shown in Figure 8d, different orderings due to line vs. word level text recognition may still be errors and significantly impact downstream translation, which is why we use these two metrics together.

In addition to aggregate quantitative metrics, we create an error taxonomy of the different classes of OCR errors we observe across different models.

<sup>8</sup>We calculate TER with SacreBleu (Post, 2018), *case\_sensitive=True, no\_punct=False, normalized=False*.

Model	CER↓	TER↓	Sub.	Del.	Ins.
Paddle-OCR	13.0	21.5	963	2824	2851
Google OCR	18.0	32.0	186	381	8496
GPT-4o	23.8	36.0	1132	1277	9728
Tesseract-OCR	124.0	134.3	9597	37081	16477

Table 3: OCR results on the VISTRA benchmark.

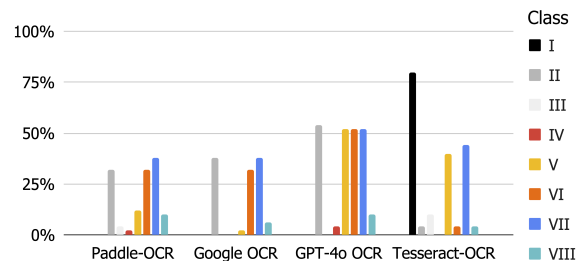


Figure 3: Proportion of OCR error classes by model.

The eight OCR error classes are listed in Table 2 and describe errors in each step of the pipeline, from text detection (text recall, text hallucinations where non-text objects are recognized as text, or bounding box placement errors which may affect downstream processing using only these regions) to recognition and generation (over and under generation of punctuation, spaces, and character- and word-level substitutions).

We provide an illustrative example for each OCR error class observed with our evaluated models on the VISTRA benchmark in Figure 8 in Appendix B. We hypothesize that differences in model design affect the proportion of each type of error, and that different error categories are likely to affect downstream translation in different ways, as we investigate in Section 3.3.

OCR performance for our 4 compared models are shown in Table 3. We observe that somewhat surprisingly, the open-source Paddle-OCR model is the highest performing on both the CER and TER metrics. While Google OCR has significantly fewer substitutions and deletions, it has a much higher insertion rate; here, this covers both more ‘benign’ insertions like whitespace, and text hallucinations as illustrated in Figure 8b, where background patterns are recognized as text characters.<sup>9</sup> GPT-4o performs slightly worse than both models on all metrics. Tesseract, on the other hand, significantly underperforms expectations set by past work (e.g.,

<sup>9</sup>It may be worth noting that where such hallucinations frequently occur as consecutive spans, and so can be significantly easier to post edit than the quantitative metrics reflect.



		CASCADED						DIRECT		
		mBART			Google Translate			GPT-4o		
OCR Model		chrF	BLEU	COMET	chrF	BLEU	COMET	chrF	BLEU	COMET
German	Tesseract-OCR	2.3	0.1	28.8	3.5	0.1	30.4	—	—	—
	Paddle-OCR	26.8	9.0	46.1	36.0	16.7	57.0	—	—	—
	GPT-4o OCR	28.1	6.9	48.0	36.4	13.2	58.2	—	—	—
	Google-OCR	31.1	9.1	47.3	37.4	14.9	55.3	—	—	—
	None	—	—	—	—	—	—	36.9	9.1	60.1
Spanish	Tesseract-OCR	2.4	0.1	30.1	3.6	0.3	31.7	—	—	—
	Paddle-OCR	17.5	3.1	44.4	60.8	33.8	75.1	—	—	—
	GPT-4o OCR	23.3	4.2	50.4	60.8	24.6	75.0	—	—	—
	Google-OCR	22.0	4.0	45.5	62.2	29.9	71.3	—	—	—
	None	—	—	—	—	—	—	54.0	21.4	73.4
Russian	Tesseract-OCR	1.7	0.1	25.3	2.6	0.1	27.3	—	—	—
	Paddle-OCR	13.0	5.8	42.4	46.5	20.0	73.0	—	—	—
	Google-OCR	16.0	7.5	42.4	48.1	18.4	71.0	—	—	—
	GPT-4o OCR	14.8	5.1	43.1	47.1	15.1	74.4	—	—	—
	None	—	—	—	—	—	—	35.6	10.7	70.2
Chinese	Tesseract-OCR	0.3	—	32.6	0.4	—	34.4	—	—	—
	Paddle-OCR	18.2	—	62.0	40.2	—	82.0	—	—	—
	GPT-4o OCR	19.7	—	63.1	40.1	—	82.5	—	—	—
	Google-OCR	18.7	—	59.2	41.6	—	77.7	—	—	—
	None	—	—	—	—	—	—	33.6	—	85.5

Table 4: Visually-situated translation results on the VISTRA benchmark. We compare both cascaded OCR and MT as well as direct translation from images with a multimodal LLM. We note results with commercial OCR and/or MT in gray, and direct translation of text in images with multimodal LLMs in blue.

Direct translation with a multimodal LLM performs quite strongly, with consistently comparable COMET scores to the strongest cascades for all target languages, though weaker comparatively on the lexical metrics chrF and BLEU; we look at this more closely in Section 3.4.

The results in Table 4 show that CER and TER alone are not sufficient indicators of performance. Translation with mBART performs more highly for Google and GPT-4o OCR than Paddle-OCR despite their higher CER and TER, suggesting the type of errors may have more significance than edit distance alone. Undetected text (Class I) has the most catastrophic effect on downstream MT. For Tesseract-OCR, recall is simply too low for non-trivial translation performance. Punctuation and whitespace are insertional errors which are detrimental to tokenization with mBART, increasing fertility by approximately 3× and resulting in input sequences which approach character level.

We hypothesize that these classes of errors may be normalized in preprocessing by the commercial MT system as they have less effect; of the sample set annotated as having these errors, segment-level chrF is 2× higher with the commercial model than mBART, which is a larger margin than observed overall (1.6×). Text hallucinations (Class II) are more difficult to remove with post-processing, though here are typically character-level rather than insertions of valid words. Character- and word-level substitutions (Classes VII and VIII) were stated to have more detrimental effect on translation for OCR’ed documents in Ignat et al. (2022) than insertions or deletions, but that is not the trend we observe here. On our type of data, natural images with complex backgrounds, we observe significantly more insertions per example than substitutions; while for example punctuation insertions (Class V) occur for a similar number of examples as character-level substitutions (Class VII) for

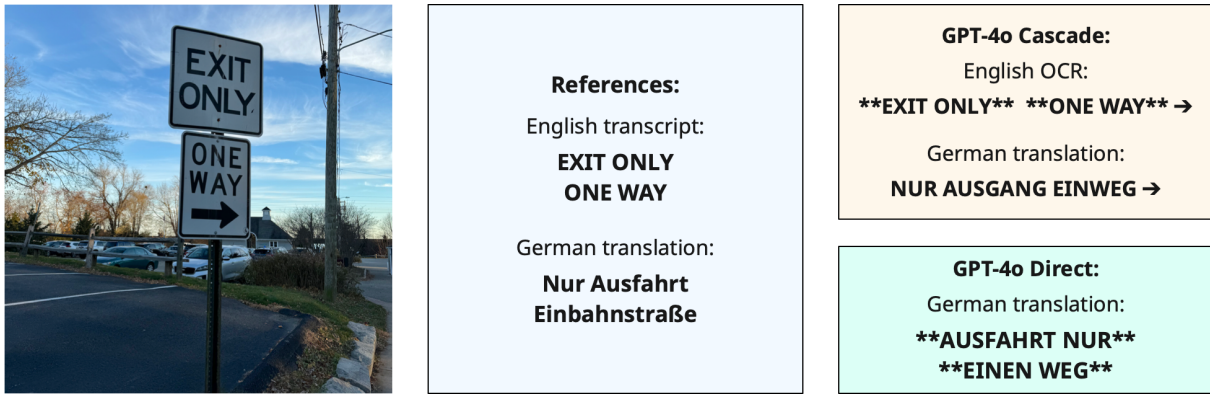


Figure 5: On this example from VISTRA, in a model cascade when translating with access to OCR output only, GPT-4o translates ‘Exit’ as ‘Ausgang,’ while when translating directly from the image with access to the visual context, GPT-4o correctly translates ‘Exit’ as ‘Ausfahrt.’

the GPT-4o OCR model in our annotated set, there are nearly 10× more insertions than substitutions in each example. When word-level substitutions occurred, they occurred at most twice per image, which both MT models were more easily able to recover from using context.

We do not observe downstream MT errors due to bounding box placement (Class III) in our sample. We note such errors may be more significant for models which process only the cropped region within bounding boxes, as in the example in Figure 8c an overly tight bounding box would cause the *g* to look like an *a* when cropped. This would be an important consideration if adapting recent visual text-based translation approaches for text in natural images (Salesky et al., 2021), as these models currently only process the region directly surrounding text.

### 3.4 Can multimodal models resolve contextual ambiguity?

Multimodal LLMs have access to both textual information contained in an image, as well as the visual context it is situated in. Cascaded OCR and MT, however, discards the visual information at translation time. Are LLMs able to use the broader visual context to resolve otherwise ambiguous translations?

It can be challenging to assess the degree to which multimodal models rely on different modalities for their predictions (Hessel and Lee, 2020), particularly for closed models without access to relative weights or the training data distribution for statistical priors. Here though direct translation with a multimodal LLM performs non-trivially, we still observe a performance gap to the same LLM

performing text translation from the reference transcripts without access to visual information: for the English→German language pair for example, 41.0 chrF and 18.8 BLEU vs. 36.9 chrF and 9.1 BLEU. Directly comparing quantitative results is not a perfect reflection of the task, because each model may get ambiguous examples wrong for different reasons. However, within the VISTRA test set we do observe examples where ambiguous source nouns are generated as only one possible translation with text input, but multiple senses with visual input. In our running example, 14 images in the benchmark contain the English word ‘Exit’; in a model cascade when translating with access to text only, GPT-4o translates all 14 instances as ‘Ausgang,’ while with visual input only 5 instances are translated this way and 4 use a variant of ‘Ausfahrt,’ as illustrated in Figure 5. Particularly when used in conjunction with models trained from scratch, this benchmark may enable further analysis of attribution.

#### Cautionary note on evaluation metrics.

Learned metrics such as COMET score paraphrases and synonyms highly, which typically leads to higher correlations with human judgments. However, for this task precisely that property may make them less reliable indicators of success. For example, returning to the motivational example in Figure 1, when translating the English sentence ‘The exit is over there,’ both possible German translations ‘Die Ausfahrt ist dort drüben’ and ‘Der Ausgang ist dort drüben’ are given identical COMET scores (97.6) with either translation as the reference. Lexical metrics such as chrF and BLEU do reflect a mismatch to the reference here, and may be more reliable in this setting specifically for measuring correct visually-situated translation. For this



reason, and given the high proportion of examples marked as contextually dependent in our benchmark (Section 2.3), the COMET scores in Table 4 should likely be viewed more cautiously than for other tasks. To properly evaluate contextually-dependent translations with multimodal input using a learned metric likely requires a new metric.

## 4 Related work

Translation of text in images has been strongly motivated by printed historical documents which require digitization (Affi and Way, 2016; Ignat et al., 2022) and PDF document translation (Zhang et al., 2023b; Hsu et al., 2024) with two column or more complex text layouts. In the absence of publicly available aligned and translated data sources, the majority of work in this space has created synthetic data for this task by rendering common machine translation corpora from sources from WMT14 (Mansimov et al., 2020; Tian et al., 2023; Niu et al., 2024; Lan et al., 2024). Ma et al. (2022) compared cascaded and direct models for in-image text translation with synthetic, cropped subtitles, and street-view images, but did not release their datasets. Lan et al. (2023) extended this work, studying auxiliary objectives for this task, and released a benchmark extending 5 Chinese OCR datasets with natural images with translations for Chinese→English. Ignat et al. (2022) perform similar analysis on the impact of OCR CER on downstream MT performance with the aim to see whether OCR’ed documents can be utilized for data augmentation for MT training with low-resource languages.

Similar to our task, multimodal translation uses auxiliary visual context to improve text translation, typically of image captions (Elliott et al., 2016; Specia et al., 2016; Elliott and Kádár, 2017; Elliott et al., 2017; Barrault et al., 2018; Li et al., 2022). Recent work has adapted pretrained model components into a single ViT model for this task (Gupta et al., 2023). As in our setting, it is challenging to assess the degree to which multimodal models make use of visual context in addition to text representations (Hessel and Lee, 2020); some studies investigating the usage of visual input in multimodal MT have found that do so primarily in the case of ambiguity or limited text input (Caglayan et al., 2019; Raunak et al., 2019) or provide regularization only (Wu et al., 2021).

Beyond machine translation, significant work has studied problems in text-centric visual pro-

cessing such as document and table layout understanding through visual means (Long et al., 2022; Alonso et al., 2024; Zheng et al., 2024), OCR-free language understanding (Tanaka et al., 2021; Ye et al., 2023), and modeling language in screenshots (Kim et al., 2022; Lee et al., 2023; Gao et al., 2024). As multimodal LLMs become increasingly strong, analyzing their capabilities and limitations for text-rich image understanding (Zhang et al., 2023a, 2024; Li et al., 2024) and OCR (Liu et al., 2023) is a growing area. As we saw here, though they are strong general purpose models, there can remain a gap to task-specific models for complex and specialized tasks.

## 5 Conclusions

We introduce a benchmark, VISTRA, for visually-situated translation of English text in natural images to four target languages. We describe the dataset construction and composition. We benchmark multiple commonly used OCR models on VISTRA, both open-source and commercial, and evaluate cascaded OCR and MT performance. We present both quantitative result and create a taxonomy of common error classes, and investigate their impact on downstream MT. Finally, we assess direct image-to-text translation with a multimodal LLM, and show that it is able in some cases but not yet consistently to disambiguate possible translations with visual context. We show that this is an unsolved and challenging task even for strong commercial models. We hope that the creation and release of our benchmark, which is the first of its kind for these language pairs, will encourage further research in this direction.

## Limitations

Our dataset is limited in scale and language coverage to English text, with images predominantly taken in a single country (USA). The majority of photos were taken by a single photographer, which may lead to more consistent image quality and application of inclusion criteria, but likely also limits diversity through a locale bias to their surroundings. Transcriptions were performed by 3 individuals, and all checked by the same annotator for consistency, while translations were professionally done with a subset checked by a second annotator.

## Acknowledgements

We are grateful to Carol Edwards and Ellen Thompson for help with transcription; Centific for professional translations; and Qin Gao, Zhe Gan, Yova Kementchedjhieva, George Sherif Botros Ibrahim, and Carlos Aguirre for helpful discussions. Elizabeth Salesky is supported by the Apple Scholars in AI/ML fellowship.

## References

- Haithem Afli and Andy Way. 2016. [Integrating optical character recognition and machine translation of historical documents](#). In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 109–116, Osaka, Japan. The COLING 2016 Organizing Committee.
- Iñigo Alonso, Eneko Agirre, and Mirella Lapata. 2024. [PixT3: Pixel-based table-to-text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6721–6736, Bangkok, Thailand. Association for Computational Linguistics.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. [Findings of the third shared task on multimodal machine translation](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.
- Ozan Caglayan, Pranava Swaroop Madhyastha, Lucia Specia, and Loïc Barrault. 2019. [Probing the need for visual context in multimodal machine translation](#). *ArXiv*, abs/1903.08678.
- Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. [Pp-ocr: A practical ultra lightweight ocr system](#).
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. [Findings of the second shared task on multimodal machine translation and multilingual image description](#). In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.
- Desmond Elliott, Stella Frank, K. Sima’an, and Lucia Specia. 2016. [Multi30k: Multilingual english-german image descriptions](#). *ArXiv*, abs/1605.00459.
- 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, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Tianyu Gao, Zirui Wang, Adithya Bhaskar, and Danqi Chen. 2024. [Improving language understanding from screenshots](#). *ArXiv*, abs/2402.14073.
- D. Gupta, S. Kharbanda, J. Zhou, W. Li, H. Pfister, and D. Wei. 2023. [Cliptrans: Transferring visual knowledge with pre-trained models for multimodal machine translation](#). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2863–2874, Los Alamitos, CA, USA. IEEE Computer Society.
- Jack Hessel and Lillian Lee. 2020. [Does my multimodal model learn cross-modal interactions? it’s harder to tell than you might think!](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.
- Benjamin Hsu, Xiaoyu Liu, Huayang Li, Yoshinari Fujinuma, Maria Nadejde, Xing Niu, Ron Litman, Yair Kittenplon, and Raghavendra Pappagari. 2024. [M3T: A new benchmark dataset for multi-modal document-level machine translation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 499–507, Mexico City, Mexico. Association for Computational Linguistics.
- Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. [OCR improves machine translation for low-resource languages](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.
- R. Reeve Ingle, Yasuhisa Fujii, Thomas Deselaers, Jonathan Baccash, and Ashok Popat. 2019. [A scalable handwritten text recognition system](#). *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 17–24.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoon Yun, Dongyoon Han, and Seunghyun Park. 2022. [Ocr-free document understanding transformer](#). In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, page 498–517, Berlin, Heidelberg. Springer-Verlag.
- Zhibin Lan, Liqiang Niu, Fandong Meng, Jie Zhou, Min Zhang, and Jinsong Su. 2024. [Translatotron\(vision\): An end-to-end model for in-image machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.
- 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](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3479–3491, Toronto, Canada. Association for Computational Linguistics.

- Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvasi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. [Pix2struct: Screenshot parsing as pretraining for visual language understanding](#). In *ICML*, pages 18893–18912.
- Xiujun Li, Yujie Lu, Zhe Gan, Jianfeng Gao, William Yang Wang, and Yejin Choi. 2024. [Text as images: Can multimodal large language models follow printed instructions in pixels?](#)
- Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio Feris, David Cox, and Nuno Vasconcelos. 2022. [Valhalla: Visual hallucination for machine translation](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5206–5216.
- 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.
- Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai. 2023. [On the hidden mystery of ocr in large multimodal models](#). *ArXiv*, abs/2305.07895.
- Shangbang Long, Siyang Qin, Dmitry Panteleev, A. Bisacco, Yasuhisa Fujii, and Michalis Raptis. 2022. [Towards end-to-end unified scene text detection and layout analysis](#). *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1039–1049.
- 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](#). *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1664–1670.
- 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.
- Liqiang Niu, Fandong Meng, and Jie Zhou. 2024. [UMTIT: Unifying recognition, translation, and generation for multimodal text image translation](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16953–16972, Torino, Italia. ELRA and ICCL.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Ashok C Popat, Jonathan Michael Baccash, Karel Driesen, Patrick Michael Hurst, and Yasuhisa Fujii. 2017. [Sequence-to-label script identification for multilingual ocr](#). In *Proceedings of the 14th International Conference on Document Analysis and Recognition (ICDAR)*.
- 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, Lisbon, Portugal. Association for Computational Linguistics.
- 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, Brussels, Belgium. Association for Computational Linguistics.
- Vikas Raunak, Sang Keun Choe, Quanyang Lu, Yi Xu, and Florian Metze. 2019. [On leveraging the visual modality for neural machine translation](#). In *International Conference on Natural Language Generation*.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Elizabeth Salesky, David Etter, and Matt Post. 2021. [Robust open-vocabulary translation from visual text representations](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- R. Smith. 2007. An overview of the tesseract ocr engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02, ICDAR '07*, page 629–633, USA. IEEE Computer Society.
- Matthew Snover, Bonnie Dorr, Rich 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, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. [A shared task on multimodal machine translation and crosslingual image description](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

- Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. [Visualmrc: Machine reading comprehension on document images](#). *ArXiv*, abs/2101.11272.
- Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo, and Bin Wang. 2023. [In-image neural machine translation with segmented pixel sequence-to-sequence model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15046–15057, Singapore. Association for Computational Linguistics.
- Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. [Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.
- Yukang Yang, Dongnan Gui, YUHUI YUAN, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. 2023. [Glyphcontrol: Glyph conditional control for visual text generation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 44050–44066. Curran Associates, Inc.
- Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. 2023. [UReader: Universal OCR-free visually-situated language understanding with multimodal large language model](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, Singapore. Association for Computational Linguistics.
- Ruiyi Zhang, Yufan Zhou, Jian Chen, Jiuxiang Gu, Changyou Chen, and Tongfei Sun. 2024. [Llava-read: Enhancing reading ability of multimodal language models](#). *ArXiv*.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tongfei Sun. 2023a. [Llavar: Enhanced visual instruction tuning for text-rich image understanding](#). *ArXiv*, abs/2306.17107.
- Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023b. [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.
- Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. [Multimodal table understanding](#). *ArXiv*, abs/2406.08100.



## A Annotation Interfaces

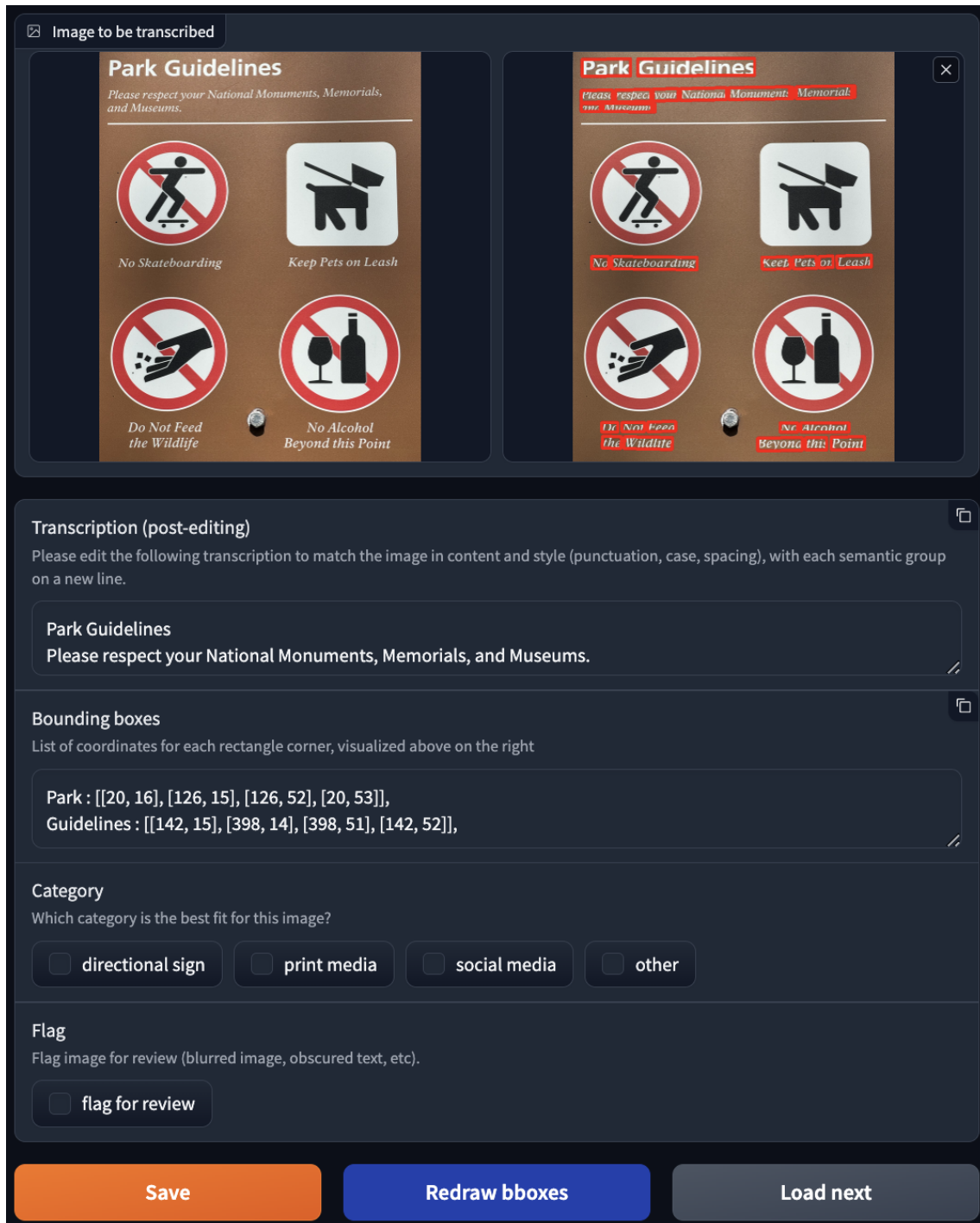



Figure 6: Text annotation interface for VISTRA benchmark.

Task **John\_Hopkinsuniversity\_20240701\_DE\_test** Hit 509595993 SKIP HIT



Please provide translations as they would appear on corresponding traffic signs locally where possible

If corresponding sign exists in your location, but with significantly different text, staying true to the text in image takes priority

If sign does not exist in your location, then provide a translation which stays true to text in image while being an appropriate translation for a sign.

If for some reason it is impossible to provide a translation which works in local language, while staying true to source please raise a query and we will confirm with client.

Transcript **directional sign** Translation **de**

CAUTION	
WET FLOOR	

Did the visual context influence the translation? \*

Yes

No

Figure 7: Translation annotation interface for VISTRA benchmark.

## B Examples of each OCR error class

Here we show an illustrative example of each OCR error class described in [Table 2](#) from the VISTRA benchmark, with the model which produced each output.



Model: Google OCR  
 Output: ESPASSING  
 Reference: NO TRESPASSING STATE  
 HIGHWAY ADMINISTRATION

(a) CLASS I: Undetected text



Model: Google OCR  
 Output: ACCESS RAMP ...  
 H H H H H H H H I  
 Reference: ACCESS RAMP



(b) CLASS II: Text hallucination



Model: Paddle-OCR  
 Output: Private Sign DONOTREAD  
 Reference: Private Sign DO NOT READ

(c) CLASS III: Bounding box error



Model: Paddle OCR  
 Output: ... I'M THINKING OF HAVE  
 YOU GOT ANY DRAWING A NEW  
 GOOD IDEAS? COMIC STRIP  
 Reference: ... I'M THINKING OF  
 DRAWING A NEW COMIC STRIP  
 HAVE YOU GOT ANY GOOD IDEAS?

(d) CLASS IV: Grouping error

Figure 8: Examples of each OCR error class from [Table 2](#).



Model: Tesseract-OCR  
 Output: | (NO OUTSIDE)!  
 ; FOOD OR !  
 || DRINKS |  
 ] ] ALLOWED |,  
 Reference: NO OUTSIDE FOOD OR  
 DRINKS ALLOWED

(e) CLASS V: Punctuation error



Model: Paddle-OCR  
 Output: PULLTOOPEN|PUSHTOCLOSE  
 Reference: PULL TO OPEN |  
 PUSH TO CLOSE

(f) CLASS VI: Spacing error



Model: Google OCR  
 Output: NO QVERNIGHT PARKING  
 Reference: NO OVERNIGHT PARKING

(g) CLASS VII: Character-level substitution



Model: Google OCR  
 Output: TOWN OF FEAST LYME ...  
 Reference: TOWN OF EAST LYME ...

(h) CLASS VIII: Word-level substitution

Figure 8: Examples of each OCR error class from Table 2 (cont.)