SEINE-DeTR: Scene-Text Aware Referring Expression Comprehension Based on Trimodal Detection Transformer

Félix DOUBLET[1,a)] Takumi KOMATSU[1] Komei SUGIURA[1,b)]

Abstract

In this study, we address the Scene-Text Oriented Referring Expression Comprehension (ST-REC) task, which requires identifying and locating a target object in an image corresponding to a scene-text oriented referring expression. Conventional REC methods do not explicitly utilize scene-text, leading to limited performance in ST-REC. To address this limitation, we propose the SEINE-DeTR model that can identify target objects based on complex textual expressions that include both scene-text and visual attributes. We validated our method on the RefText dataset. Experimental results demonstrate that our method outperforms baseline methods in terms of Precision@0.5.

1. Introduction

In the field of Vision & Language, there exist many tasks with promising societal applications, such as autonomous driving [6, 5] and real-world search [12, 16]. In these tasks, scene-text —such as road signs and product names— can serve as crucial cues for identifying or referring to objects. Particularly in the Referring Expression Comprehension (REC) task, using scene-text, alongside object and spatial information, could be beneficial across a range of applications, including assistive technology for the visually impaired, landmark-based navigation, and product search systems.

In this study, we address the Scene-Text Oriented Referring Expression Comprehension (ST-REC) task [1], which requires identifying and locating a target object in an image based on a scene-text oriented referring expression.

Figure 1 shows a typical use case for the ST-REC task.

Reference: “The bottle with Niacin”

Fig. 1 A typical use case for the ST-REC task. Given an image and a referring expression, the model aims to predict the bounding box of the target object (shown in green).

In this example, several bottles are arranged in a row. Given the input expression “The bottle with Niacin”, the goal is to identify the bottle labeled “Niacin” as shown by the green bounding box in the figure.

The ST-REC task is more challenging than the typical REC task. In fact, the human performance for this task is reported to have a Precision@0.5 score of 93.71 points. However, even the best-performing existing ST-REC model, STAN, achieved only 65.86 points for the ST-REC task, while scoring up to 83.15 points for the simple REC task [1], highlighting the difficulty of the ST-REC task.

Conventional REC methods [9, 4, 11, 20, 2, 22], do not explicitly utilize scene-text, leading to limited performance in ST-REC [1]. Thus, methods like STAN, which introduce mechanisms to process scene-text, have been proposed. Yet, their performance still remains insufficient as they struggle to accurately identify the target object in samples where similar objects are present [1].

In this study, we propose SEINE-DeTR [*1] which can identify target objects based on complex textual expres-

[1] Keio University
[a)] felixdoublet@keio.jp
[b)] komei.sugiura@keio.jp

[*1] **S**cene-t**E**xt aware referr**I**ng Expressio**N** compr**E**hension **De**tection **TR**ansformer

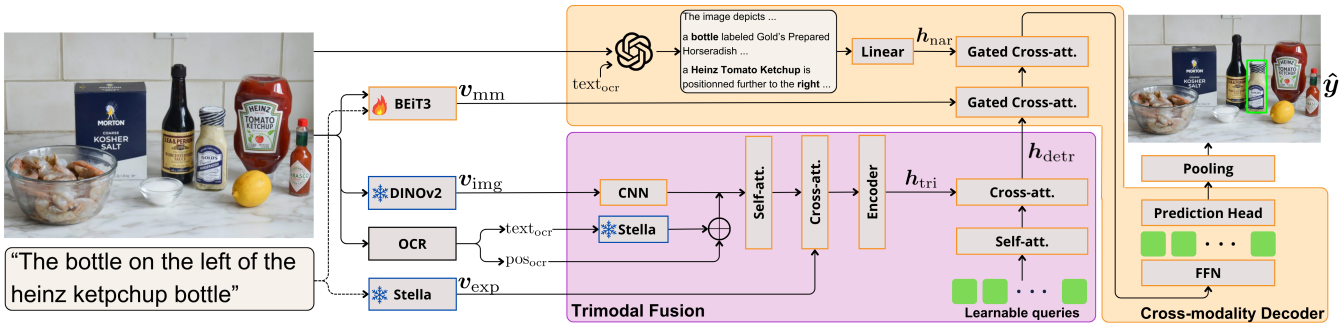


Fig. 2 Overview of SEINE-DeTR.

sions that include both scene-text and visual attributes.

The main contributions of this study are as follows:

- We introduce Trimodal Fusion Module (TFM), which enables the model to learn both the content and spatial localization of scene-text by directly integrating these cues into a query-based Transformer decoder.
- We propose Cross-Modality Decoder, which models spatial relationships between objects by leveraging detected texts, enabling the generation of hallucination-robust narrative.

2. Method

The proposed method consists of two main modules: Trimodal Fusion and Cross-modality Decoder. Figure 2 shows the structure of our proposed method.

The input to SEINE-DeTR is $\{\boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{exp}}\}$. Here, $\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{x}_{\text{exp}} \in \{0,1\}^{v\times l}$ represent the image and the referring expression, respectively. C, H and W denote the number of channels, width and height of the input image, while v and l correspond to the vocabulary size of the referring expression and the maximum number of tokens, respectively.

The language features $\boldsymbol{v}_{\text{exp}} \in \mathbb{R}^{D_{\text{txt}}}$ are extracted from $\boldsymbol{x}_{\text{exp}}$ using Stella [23]. Here, D_{txt} denotes the dimensionality of the features. The image features $\boldsymbol{v}_{\text{img}} \in \mathbb{R}^{D_{\text{img}}\times N_{\text{img}}}$ are extracted from $\boldsymbol{x}_{\text{img}}$ using DINOv2 [15]. Here, D_{img} and N_{img} denote the dimensionality of the feature vectors, and the sequence length of the extracted image tokens, respectively. The multimodal features $\boldsymbol{v}_{\text{mm}} \in \mathbb{R}^{D_{\text{mm}}\times N_{\text{mm}}}$ are extracted from $\boldsymbol{x}_{\text{img}}$ and $\boldsymbol{x}_{\text{exp}}$ using BEiT-3 [19]. Here, D_{mm} and N_{mm} denote the dimensionality of the feature vectors, and the sequence length of the extracted image tokens, respectively.

2.1 Trimodal Fusion

In TFM, we integrate scene-text cues into the final representation. Indeed, scene-text serves as crucial information to uniquely identify the object, as is the case with the expression "The bottle with Niacin" shown in Figure 1 in an environment containing several visually similar bottles. However, conventional REC methods either fail to explicitly leverage scene-text information or rely solely on implicit text representations encoded within image features. To address this limitation, we introduce TFM that explicitly aligns textual embeddings from detected scene texts with their spatial coordinates. These features are combined with $\boldsymbol{v}_{\text{exp}}$ and $\boldsymbol{v}_{\text{img}}$ to serve as input of a query based transformer decoder.

First, we apply Azure AI Vision [13] to $\boldsymbol{x}_{\text{img}}$ for OCR, extracting detected text alongside their coordinates, denoted as $\{\text{text}_{\text{ocr}}, \text{pos}_{\text{ocr}}\}$. Each detected text instance is then encoded using Stella [23], yielding the scene-text language feature representation $\boldsymbol{v}_{\text{st}} \in \mathbb{R}^{N_{\text{words}}\times D_{\text{txt}}}$. Here, N_{words} and D_{txt} denote the number of considered words and the dimensionality of the text features, respectively. At this stage, Sinusoidal Positional Encoding is applied based on the corresponding coordinates pos_{ocr}, incorporating positional information to obtain $\boldsymbol{h}_{\text{st}}$. In parallel, we downsample $\boldsymbol{v}_{\text{img}}$ using a Convolutionnal Neural Network (CNN) to obtain $\boldsymbol{h}_{\text{img}} \in \mathbb{R}^{N_{\text{down}}\times D_{\text{img}}}$. Here, N_{down} and D_{img} correspond to the downsampled sequence length and the dimensionality of the image features, respectively.

Subsequently, we derive the trimodal representation $\boldsymbol{h}_{\text{tri}} \in \mathbb{R}^{(N_{\text{words}}+N_{\text{down}})\times D_{\text{tri}}}$ where D_{tri} denotes the dimensionality to which all feature types are projected. This representation is computed as follows:

$$\boldsymbol{h}_{\text{tri}} = \text{Enc}(\text{CrossAtt}(\text{SelfAtt}([\boldsymbol{h}_{\text{img}}, \boldsymbol{h}_{\text{st}}]), \boldsymbol{v}_{\text{exp}})), \quad (1)$$

where Enc., CrossAtt. and SelfAtt. represent a transformer encoder, a cross-attention and a self-attention architecture, respectively.

Following the MDETR architecture [9], we initialise learnable queries $\boldsymbol{q} \in \mathbb{R}^{N_{\text{queries}}\times D_{\text{dec}}}$ where N_{queries} and D_{dec} are the number of queries and the dimensionality of the decoder, respectively. The output of this module is computed as follows:

$$\boldsymbol{q}_{\text{detr}} = \text{CrossAttn}(\text{SelfAttn}(\boldsymbol{q}), \boldsymbol{h}_{\text{tri}}). \quad (2)$$

Table. 1 Quantitative comparison between the proposed method and baseline methods.

[%] Methods	Scenes						OOV ↑	Semantic Info ↑	MiniRefer-Text ↑
	All ↑	Street ↑	Shelf ↑	Home ↑	Sport ↑	Others ↑			
EVF-SAM [24]	–	–	–	–	–	–	–	–	29.47
Qwen2-VL-2b [18]	–	–	–	–	–	–	–	–	63.29
Qwen2-VL-7b [18]	–	–	–	–	–	–	–	–	64.73
GPT-4o [8]	5.68	7.38	6.60	6.67	4.62	3.19	5.14	3.94	1.45
STAN [1]	65.86	73.77	61.68	73.97	70.98	42.78	62.03	62.07	–
STAN (rep.) [1]	64.58±0.90	73.19±2.26	60.14±1.44	72.77±1.24	68.80±0.93	42.31±1.65	61.08±0.96	54.88±1.40	27.54±1.23
SEINE-DeTR (Ours)	**70.37±0.30**	**80.02±0.70**	**61.79±0.63**	**74.06±0.54**	**84.30±0.30**	**42.95±0.77**	**62.35±0.57**	**62.13±0.79**	**66.95±1.48**
Human performance	93.71	–	–	–	–	–	–	–	–

2.2 Cross-modal Decoder

In Cross-Modality Decoder, we enhance the representation of queries by integrating multimodal features alongside narrative representation features. In fact, some studies [7, 21] report that integrating narrative representation features can enhance the comprehension of spatial relationships, which is beneficial for tasks such as ST-REC. Most of them propose approaches using MLLMs to generate narrative representations, however, MLLMs frequently produce hallucinations due to their limited comprehension of scene-text [3]. To address this limitation, we incorporate the text extracted via third-party OCR directly into the prompt, thereby enhancing scene text comprehension of the MLLM.

In this module, we firstly generate hallucination-robust narrative representation $\boldsymbol{h}_{\text{nar}}$ from $\boldsymbol{x}_{\text{img}}$. Specifically, we generate a narrative description of $\boldsymbol{x}_{\text{img}}$ using GPT-4o [8] with a prompt p which includes text_{ocr}. $\boldsymbol{h}_{\text{nar}}$ is obtained by encoding this narrative description using Stella. The module then produces the refined output queries $\tilde{\boldsymbol{q}} \in \mathbb{R}^{N_{\text{queries}} \times D_{\text{dec}}}$ defined as:

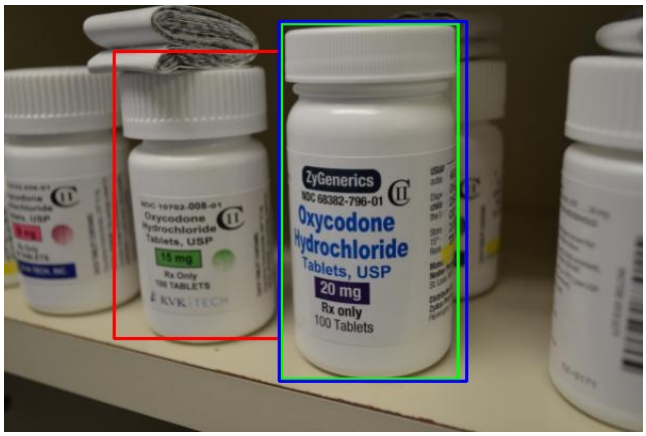
$$\tilde{\boldsymbol{q}} = \text{GtdCrossAtt}(\text{GtdCrossAtt}(\boldsymbol{q}_{\text{detr}}, \boldsymbol{v}_{\text{mm}}), \boldsymbol{h}_{\text{nar}}), \quad (3)$$

where GtdCrossAtt. represents a gated cross-attention architecture.

The final model's prediction is denoted as:

$$\hat{\boldsymbol{y}} = \{\hat{c}_x, \hat{c}_y, \hat{w}, \hat{h}\} = \frac{1}{N_q} \sum_{i=1}^{N_q} \text{MLP}(\tilde{\boldsymbol{q}}_{\boldsymbol{i}}), \quad (4)$$

where $(\hat{c}_x, \hat{c}_y)$ are the predicted center coordinates of the bounding box, and $\hat{w}$, $\hat{h}$ denote the predicted width and height of the bounding box, respectively. Additionally, N_q represents the number of queries while $\tilde{\boldsymbol{q}}_{\boldsymbol{i}}$ is the i-th output query.



Reference: "a bottle with a label that says 20 mg"

Fig. 3 A successful sample on the RefText dataset.

3. Experiments

3.1 RefText dataset and MiniReferText dataset

We primarily used the RefText dataset [1] to evaluate our method. The RefText dataset contains 31,082 samples which were randomly divided into training, validation, and test sets [1]. The training, validation, and test sets contain 25,030, 2,022, and 4,030 samples, respectively, with no duplicate images across these sets. The test set is further divided into the subsets "Street", "Shelf", "Home", "Sport", and "Others" containing 854, 561, 899, 996, and 720 samples, respectively. To evaluate the performance of the model on out-of-vocabulary (OOV) and semantic understanding, the test set includes additional subsets "OOV" and "Semantic Information" which consist of 1,459 and 762 samples, respectively.

To ensure a fair and reliable evaluation of model performance, we constructed the MiniReferText dataset to address potential data leakage concerns. Indeed, the RefText dataset was randomly partitioned into training, validation, and test sets. Consequently, potential data leakage may arise when evaluating models trained on the

original datasets from which the samples or images were sourced. Notably, RefText incorporates samples from widely used datasets, such as COCO-Text [17] and Visual Genome [10], which are also utilised in training models such as Qwen2-VL [18] and EVF-SAM [24].

To enable a comparative evaluation of our proposed method against additional models under a zero-shot setting, we developed the MiniReferText dataset. We constructed the dataset by selecting images from the recently introduced Megalith-MDQA dataset [14]. Additionally, we incorporated images we collected from a shopping mall, resulting in a total of 54 images. Each image was manually annotated, resulting in a total of 207 samples. Each sample consists of an English scene-text oriented referring expression, accompanied by the corresponding bounding box of the target object.

3.2 Quantitative results

Table 1 presents the quantitative results of the comparison between the baseline methods and the proposed method. We reported the results of STAN [1] based on the values reported in [1] as well as the values obtained from our reproduction experiments (STAN (rep.)). We conducted five experiments for STAN (rep.) and our method and reported the mean and standard deviation. The best scores are in bold.

The evaluation of the predicted bounding boxes was based on Precision@0.5 (P@0.5). P@0.5 is calculated as the proportion of correctly predicted bounding boxes out of all the predicted bounding boxes where samples with an IoU of 0.5 or higher were considered correct.

$$P@0.5 = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\mathrm{IoU}_i \geq 0.5), \tag{5}$$

where N denotes the total number of predicted bounding boxes. For each sample, the Intersection over Union (IoU) was calculated as shown below.

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|}, \tag{6}$$

where, A and B denote the sets of pixels contained within the ground truth and predicted bounding boxes, respectively.

As listed in Table 1, SEINE-DeTR achieved a P@0.5 of 70.37%, while GPT-4o and STAN scored 5.68% and 65.86%, respectively. The proposed method exceeded the highest among baselines, STAN, by 4.51 points in P@0.5 and achieved the best performances on all the different subsets. In particular, on the subsets Street, Shelf, Home, Sport and Others, our method outperformed STAN, the highest among baselines by 6.25, 0.11, 0.09, 13.32 and 0.17 points, respectively. Furthermore, on the OOV and SI, the proposed method outperformed STAN by 0.32 and 0.06 points, respectively.

On the MiniReferText dataset, SEINE-DeTR achieved a P@0.5 of 66.95%, while EVF-SAM, Qwen2-VL-2b, Qwen2-VL-7b, GPT-4o and STAN scored 29.47%, 63.29%, 64.73%, 1.45% and 27.54%, respectively. The proposed method exceeded the highest among baselines, Qwen2-VL-7b, by 2.22 points in P@0.5.

3.3 Qualitative results

Figure 3 shows a successful example on the RefText dataset. In Figure 3, the green box shows the ground truth bounding box, while the red box and blue box show the bounding box predicted by STAN and SEINE-DeTR, respectively.

In Figure 3, $\boldsymbol{x}_{\mathrm{exp}}$ is "a bottle with a label that says 20 mg". The image depicts three bottles, with the two rightmost ones labeled "15mg" and "20mg", respectively. STAN incorrectly predicted a bounding box for the bottle labeled "15mg". In contrast, the proposed method correctly predicted a bounding box for the bottle labeled "20mg".

This result suggests that the proposed method effectively identifies and understands the scene-text described in the reference.

4. Conclusion

In this study, we handled the ST-REC task. Our key contributions in this study are as follows:

- We introduced Trimodal Fusion Module (TFM), which enables the model to learn both the content and spatial localization of scene-text by directly integrating these cues into a query-based Transformer decoder.
- We proposed Cross-Modality Decoder, which models spatial relationships between objects by leveraging detected texts, enabling the generation of hallucination-robust narrative.
- Our method outperformed the baseline methods in terms of the standard metrics.

In future studies, we plan to segment the original images into distinct patches in order to improve both input resolution and spatial understanding.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 23K28168 and JST Moonshot.

References

[1] Bu, Y., Li, L., Xie, J., Liu, Q., Cai, Y., Huang, Q. and Li, Q.: Scene-text oriented referring expression comprehension, *Transactions on Multimedia*, Vol. 25, pp. 7208–7221 (2022).

[2] Chen, L., Ma, W., Xiao, J. et al.: Ref-nms: Breaking proposal bottlenecks in two-stage referring expression grounding, *AAAI*, Vol. 35, No. 2, pp. 1036–1044 (2021).

[3] Chen, X., Wang, C., Xue, Y. et al.: Unified hallucination detection for multimodal large language models, *ACL* (2024).

[4] Deng, J., Yang, Z., Chen, T. et al.: Transvg: End-to-end visual grounding with transformers, *ICCV*, pp. 1769–1779 (2021).

[5] Ding, K., Chen, B., Su, Y. et al.: Hint-ad: Holistically aligned interpretability in end-to-end autonomous driving, *CoRL* (2024).

[6] Dong, Z., Zhu, Y., Li, Y. et al.: Generalizing End-To-End Autonomous Driving In Real-World Environments Using Zero-Shot LLMs, *CoRL* (2024).

[7] Goko, M., Kambara, M., Saito, D. et al.: Task success prediction for open-vocabulary manipulation based on multi-level aligned representations, *CoRL* (2024).

[8] Hurst, A., Lerer, A., Goucher, A. P. et al.: Gpt-4o system card, *arXiv preprint arXiv:2410.21276* (2024).

[9] Kamath, A., Singh, M., LeCun, Y. et al.: Mdetr-modulated detection for end-to-end multi-modal understanding, *ICCV*, pp. 1780–1790 (2021).

[10] Krishna, R., Zhu, Y., Groth, O. et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations, *International journal of computer vision*, Vol. 123, pp. 32–73 (2017).

[11] Li, L., Bu, Y. and Cai, Y.: Bottom-up and bidirectional alignment for referring expression comprehension, *ACM International Conference on Multimedia*, pp. 5167–5175 (2021).

[12] Majumdar, A., Ajay, A., Zhang, X. et al.: Openeqa: Embodied question answering in the era of foundation models, *CVPR*, pp. 16488–16498 (2024).

[13] Microsoft Corporation: Azure AI Vision, `https://azure.microsoft.com/en-us/products/ai-services/ai-vision` (2025). Accessed: 2025-04-17.

[14] moondream: Megalith-MDQA, `https://huggingface.co/datasets/moondream/megalith-mdqa` (2025). Accessed: 2025-04-17.

[15] Oquab, M., Darcet, T., Moutakanni, T. et al.: Dinov2: Learning robust visual features without supervision, *Transactions on Machine Learning Research* (2024).

[16] Sigurdsson, G. A., Thomason, J., Sukhatme, G. S. and Piramuthu, R.: Rrex-bot: Remote referring expressions with a bag of tricks, *IROS*, IEEE, pp. 5203–5210 (2023).

[17] Veit, A., Matera, T., Neumann, L. et al.: Coco-text: Dataset and benchmark for text detection and recognition in natural images, *arXiv preprint arXiv:1601.07140* (2016).

[18] Wang, P., Bai, S., Tan, S. et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, *arXiv preprint arXiv:2409.12191* (2024).

[19] Wang, W., Bao, H., Dong, L., Bjorck, J. et al.: Image as a foreign language: Beit pretraining for vision and vision-language tasks, *CVPR*, pp. 19175–19186 (2023).

[20] Yang, Z., Chen, T., Wang, L. and Luo, J.: Improving one-stage visual grounding by recursive sub-query construction, *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, Springer, pp. 387–404 (2020).

[21] Yashima, D., Korekata, R. and Sugiura, K.: Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning with Dense Labeling, *Robotics and Automation Letters* (2024).

[22] Yu, L., Lin, Z., Shen, X. et al.: Mattnet: Modular attention network for referring expression comprehension, *CVPR*, pp. 1307–1315 (2018).

[23] Zhang, D., Li, J., Zeng, Z. and Wang, F.: Jasper and Stella: distillation of SOTA embedding models, *arXiv preprint arXiv:2412.19048* (2024).

[24] Zhang, Y., Cheng, T., Zhu, L. et al.: Evf-sam: Early vision-language fusion for text-prompted segment anything model, *arXiv preprint arXiv:2406.20076* (2024).