WED-PM-273

Towards Unified Scene Text Spotting based on Sequence Generation

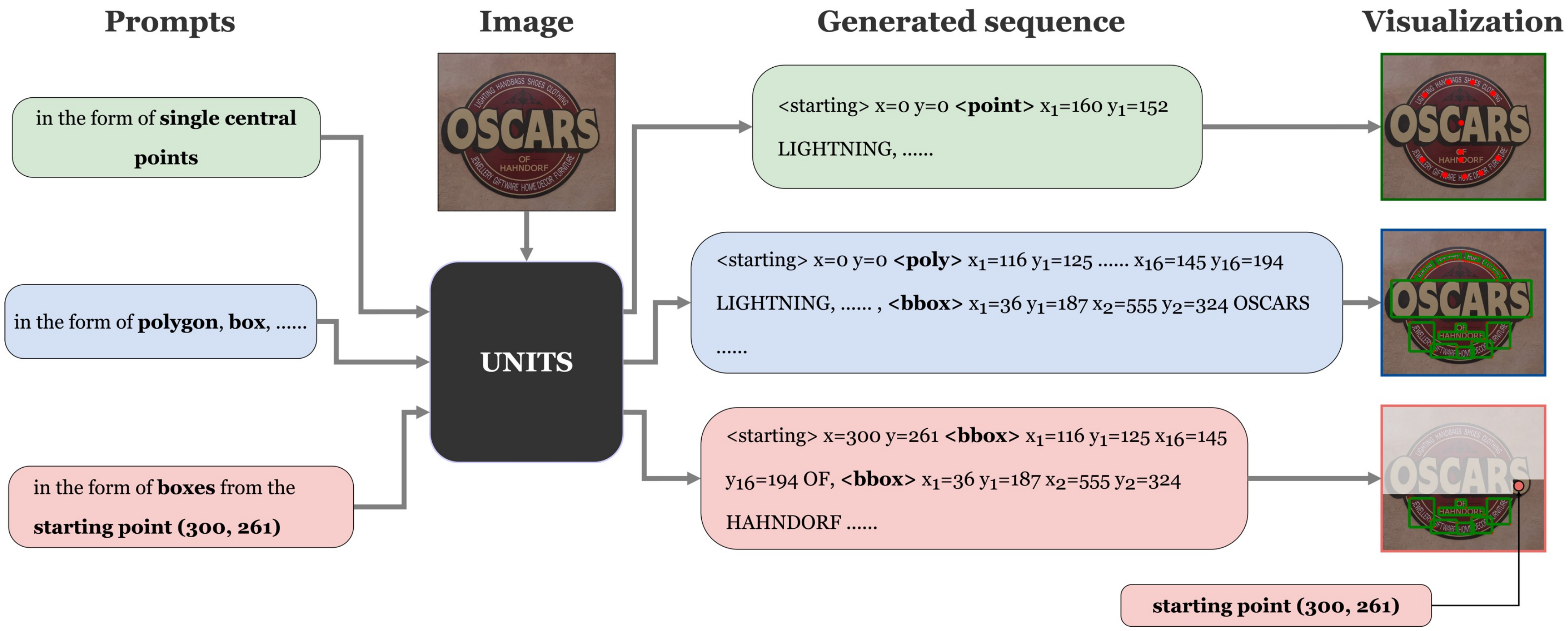
Taeho Kil[1*] Seonghyeon Kim[2†] Sukmin Seo[1] Yoonsik Kim[1] Daehee Kim[1]

[1]Naver Cloud [2]Kakao Brain

{taeho.kil, sukmin.seo, yoonsik.kim90, daehee.k}@navercorp.com, matt.mldev@kakaobrain.com

Overview

- Tackle the text spotting task using a sequence generation method



Prompts

in the form of **single central points**

in the form of **polygon**, **box**, ...

in the form of **boxes** from the **starting point (300, 261)**

Image

UNITS

Generated sequence

<starting> x=0 y=0 **<point>** x_1=160 y_1=152 LIGHTNING, ...

<starting> x=0 y=0 **<poly>** x_1=116 y_1=125 ...... x_{16}=145 y_{16}=194 LIGHTNING, ... , **<bbox>** x_1=36 y_1=187 x_2=555 y_2=324 OSCARS ......

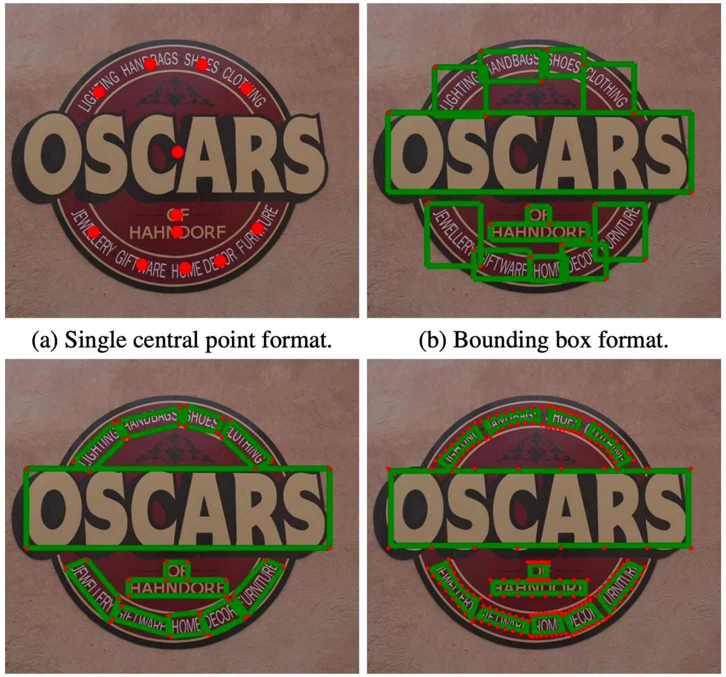
<starting> x=300 y=261 **<bbox>** x_1=116 y_1=125 x_{16}=145 y_{16}=194 OF, **<bbox>** x_1=36 y_1=187 x_2=555 y_2=324 HAHNDORF ......

Visualization

starting point (300, 261)

Problem statement

- Needs to cover all detection formats instead of relying on only one


LIGHTING HANDBAGS SHOES CLOTHING
OSCARS
OF
HAHNDORF
JEWELLERY GIFTWARE HOME DECOR FURNITURE

(a) Single central point format.

(b) Bounding box format.

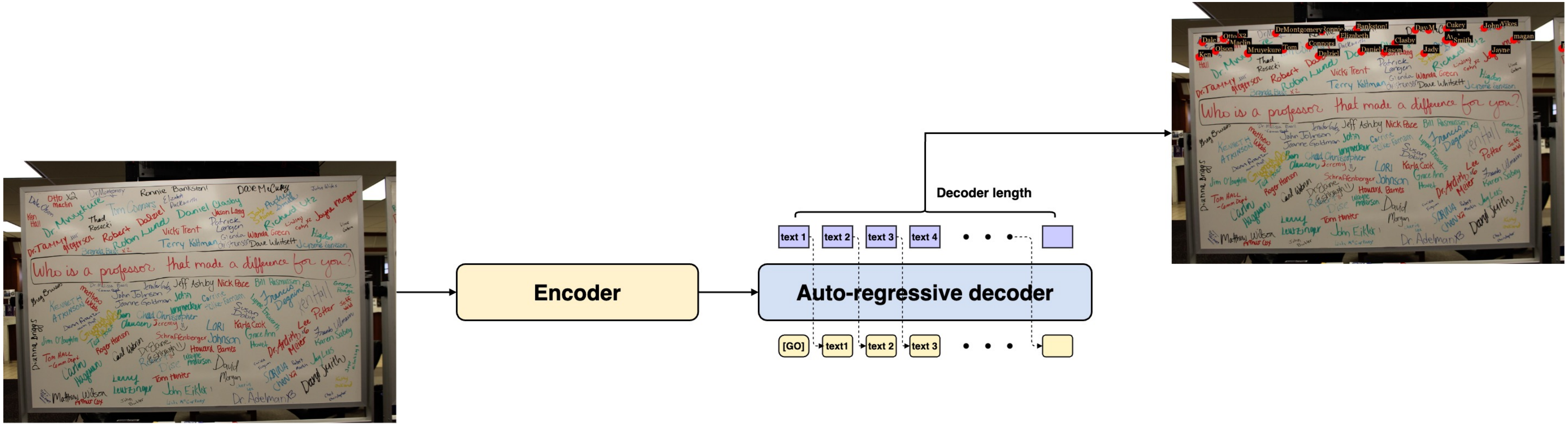
(c) Quadrilateral format.

(d) Polygonal format.

Figure 1. Various types of detection formats. The green line represents the boundary shape of the detection format, and the red dot represents the points used for the corresponding format.

Problem statement

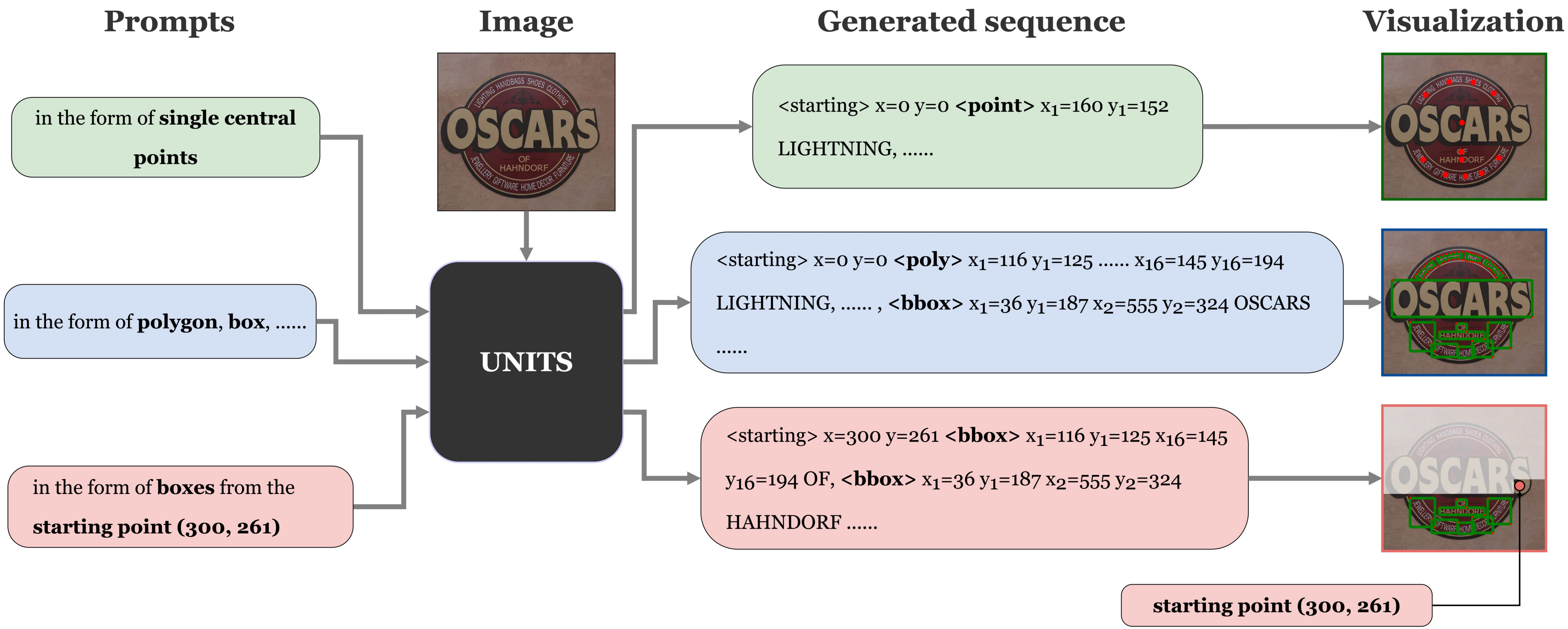
- The maximum length of the decoder limits the number of texts that can be extracted



Contribution

- Propose a novel sequence generation-based scene text spotting method
- Extract arbitrary-shaped text areas by unifying various detection formats
- Extract more texts than the decoder length allows using the starting-point prompt

Proposed method



Prompts

in the form of **single central points**

in the form of **polygon**, **box**, ...

in the form of **boxes** from the **starting point (300, 261)**

Image

UNITS

Generated sequence

<starting> x=0 y=0 **<point>** x_1=160 y_1=152 LIGHTNING, ...

<starting> x=0 y=0 **<poly>** x_1=116 y_1=125 ...... x_{16}=145 y_{16}=194 LIGHTNING, ... , **<bbox>** x_1=36 y_1=187 x_2=555 y_2=324 OSCARS ......

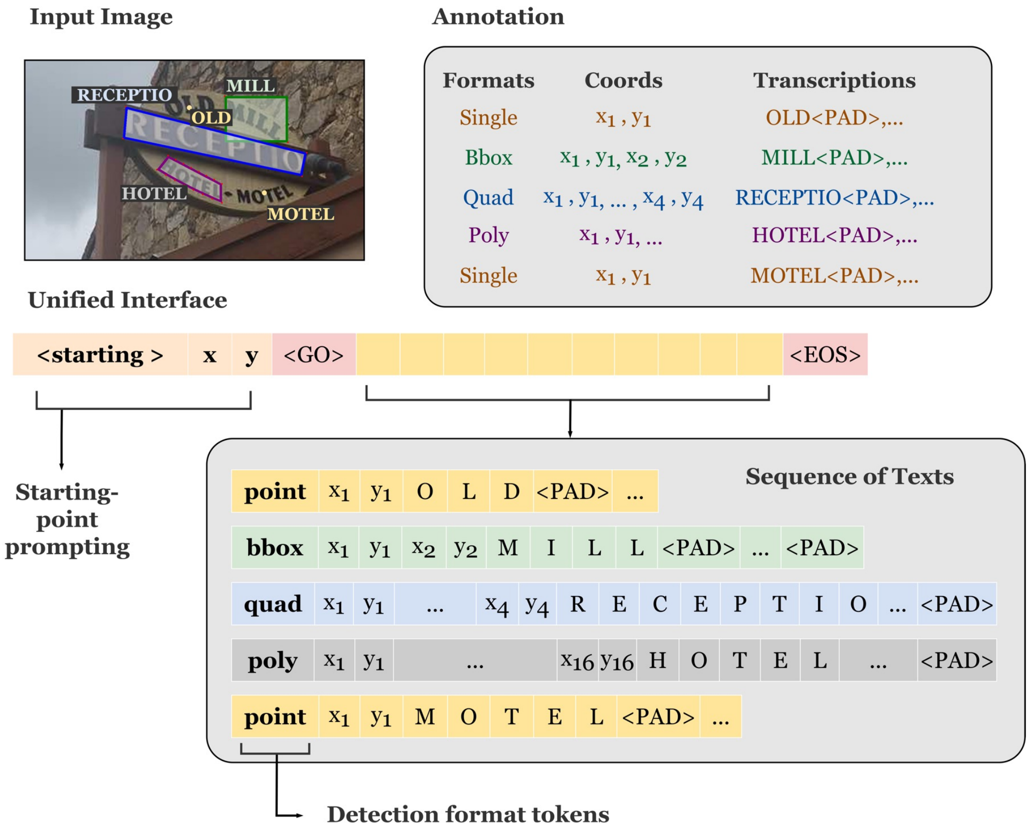
<starting> x=300 y=261 **<bbox>** x_1=116 y_1=125 x_{16}=145 y_{16}=194 OF, **<bbox>** x_1=36 y_1=187 x_2=555 y_2=324 HAHNDORF ......

Visualization

starting point (300, 261)

LIGHTING HANDBAGS SHOES CLOTHING
OSCARS
OF
HAHNDORF
JEWELLERY GIFTWARE HOME DECOR FURNITURE

Unified interface for text spotting



Input Image

Annotation

Formats	Coords	Transcriptions
Single	x_1 , y_1	OLD<PAD>,...
Bbox	x_1 , y_1, x_2 , y_2	MILL<PAD>,...
Quad	$x_1 , y_1, ... , x_4 , y_4$	RECEPTIO<PAD>,...
Poly	$x_1 , y_1, ...$	HOTEL<PAD>,...
Single	x_1 , y_1	MOTEL<PAD>,...

Unified Interface

<starting > x y <GO> <EOS>

Starting-point prompting

Sequence of Texts

point x_1 y_1 O L D <PAD> ...

bbox x_1 y_1 x_2 y_2 M I L L <PAD> ... <PAD>

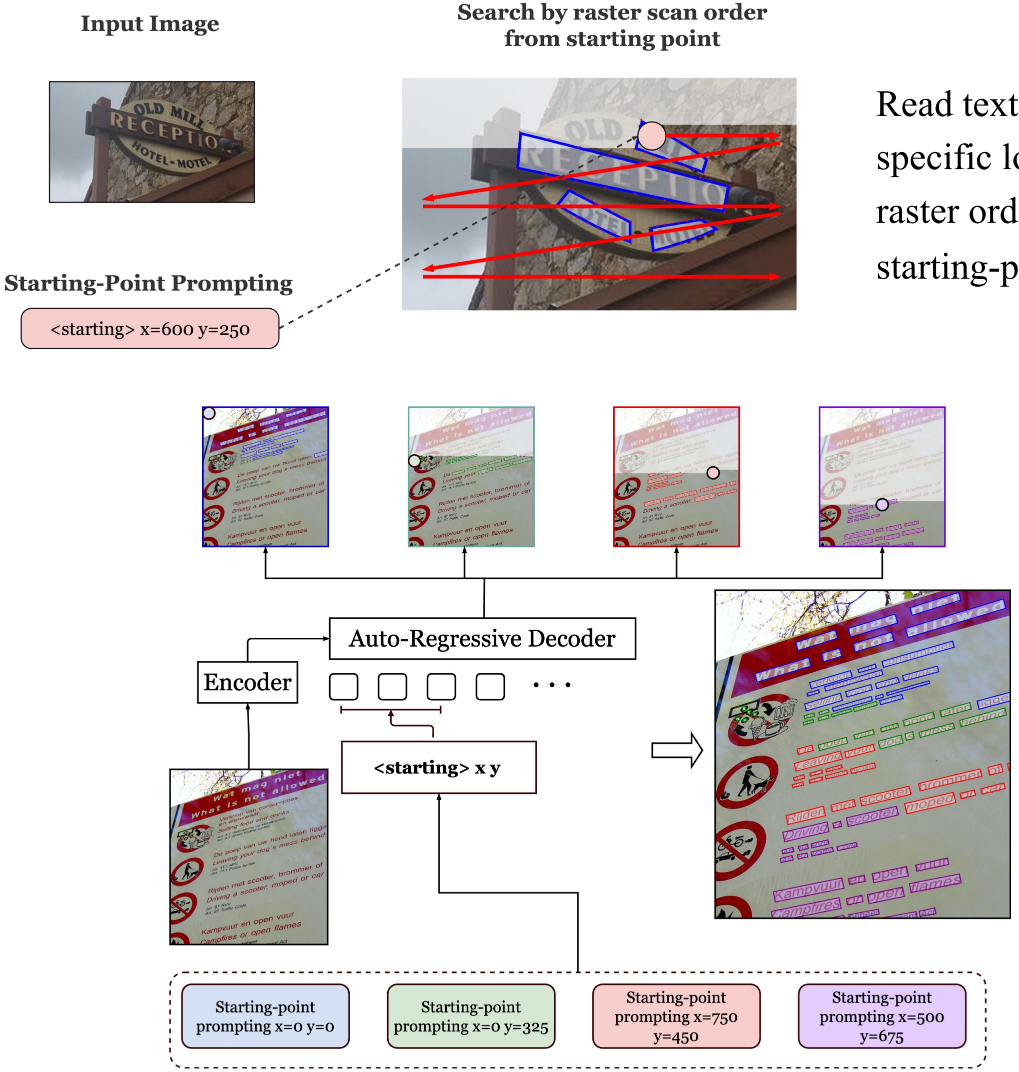
quad x_1 y_1 ... x_4 y_4 R E C E P T I O ... <PAD>

poly x_1 y_1 ... x_{16} y_{16} H O T E L ... <PAD>

point x_1 y_1 M O T E L <PAD> ...

Detection format tokens

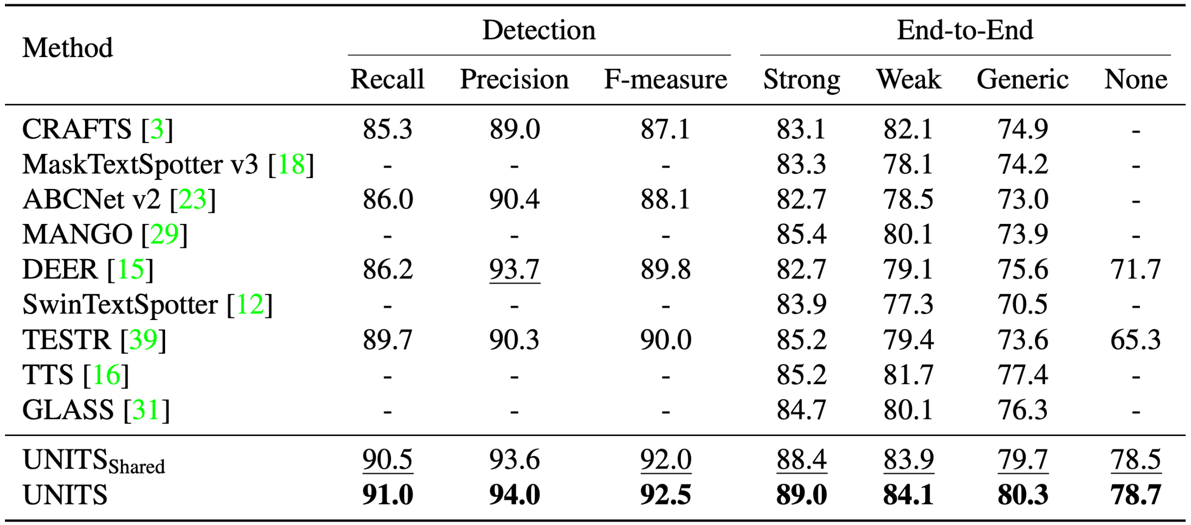
Detection format token enables a single model to handle multiple detection foramts



Input Image
Search by raster scan order from starting point
Starting-Point Prompting
<starting> x=600 y=250
Encoder
Auto-Regressive Decoder
<starting> x y
Starting-point prompting x=0 y=0
Starting-point prompting x=0 y=325
Starting-point prompting x=750 y=450
Starting-point prompting x=500 y=675

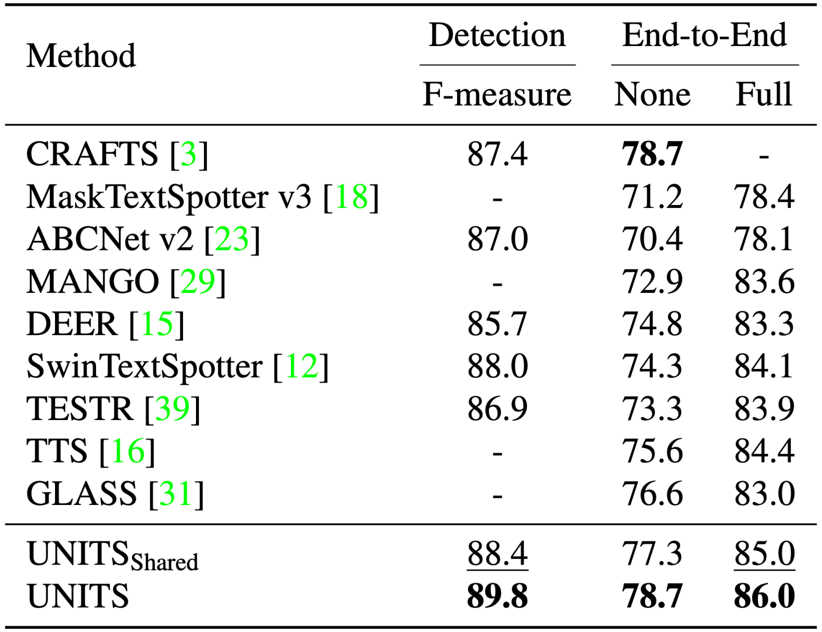
Read texts from a specific location in raster order by using starting-point prompt

Comparisons



Method	Detection			End-to-End			
	Recall	Precision	F-measure	Strong	Weak	Generic	None
CRAFTS [3]	85.3	89.0	87.1	83.1	82.1	74.9	-
MaskTextSpotter v3 [18]	-	-	-	83.3	78.1	74.2	-
ABCNet v2 [23]	86.0	90.4	88.1	82.7	78.5	73.0	-
MANGO [29]	-	-	-	85.4	80.1	73.9	-
DEER [15]	86.2	93.7	89.8	82.7	79.1	75.6	71.7
SwinTextSpotter [12]	-	-	-	83.9	77.3	70.5	-
TESTR [39]	89.7	90.3	90.0	85.2	79.4	73.6	65.3
TTS [16]	-	-	-	85.2	81.7	77.4	-
GLASS [31]	-	-	-	84.7	80.1	76.3	-
$\text{UNITS}_{\text{Shared}}$	90.5	93.6	92.0	88.4	83.9	79.7	78.5
UNITS	**91.0**	**94.0**	**92.5**	**89.0**	**84.1**	**80.3**	**78.7**

Table 1. Experiment results on ICDAR 2015. “Strong”, “Weak”, “Generic”, and “None” represent recognition with each lexicon respectively.



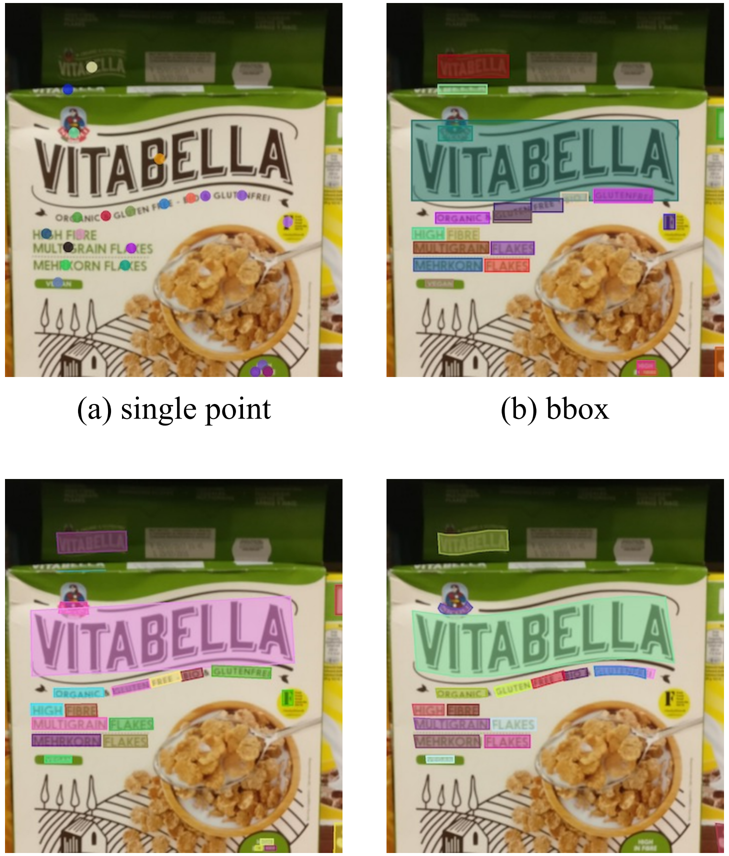
Method	Detection	End-to-End	
	F-measure	None	Full
CRAFTS [3]	87.4	**78.7**	-
MaskTextSpotter v3 [18]	-	71.2	78.4
ABCNet v2 [23]	87.0	70.4	78.1
MANGO [29]	-	72.9	83.6
DEER [15]	85.7	74.8	83.3
SwinTextSpotter [12]	88.0	74.3	84.1
TESTR [39]	86.9	73.3	83.9
TTS [16]	-	75.6	84.4
GLASS [31]	-	76.6	83.0
$\text{UNITS}_{\text{Shared}}$	88.4	77.3	85.0
UNITS	**89.8**	**78.7**	**86.0**

Table 2. Experiment results on Total-Text. “Full” and “None” represent recognition with each lexicon respectively.

Ablations

Method	End-to-End		
	Strong	Weak	Generic
SPTS [28]	77.5	70.2	65.8
UNITS$_{\text{Shared}}$ – Point	89.9	84.1	79.3
UNITS$_{\text{Shared}}$ – Box	**90.1**	**84.5**	79.3
UNITS$_{\text{Shared}}$ – Quad	89.9	**84.5**	**79.5**
UNITS$_{\text{Shared}}$ – Polygon	89.4	84.0	79.0

Table 3. The end-to-end recognition performance evaluated by the point-based metric [28] on ICDAR 2015.



(a) single point

(b) bbox

(c) quad format

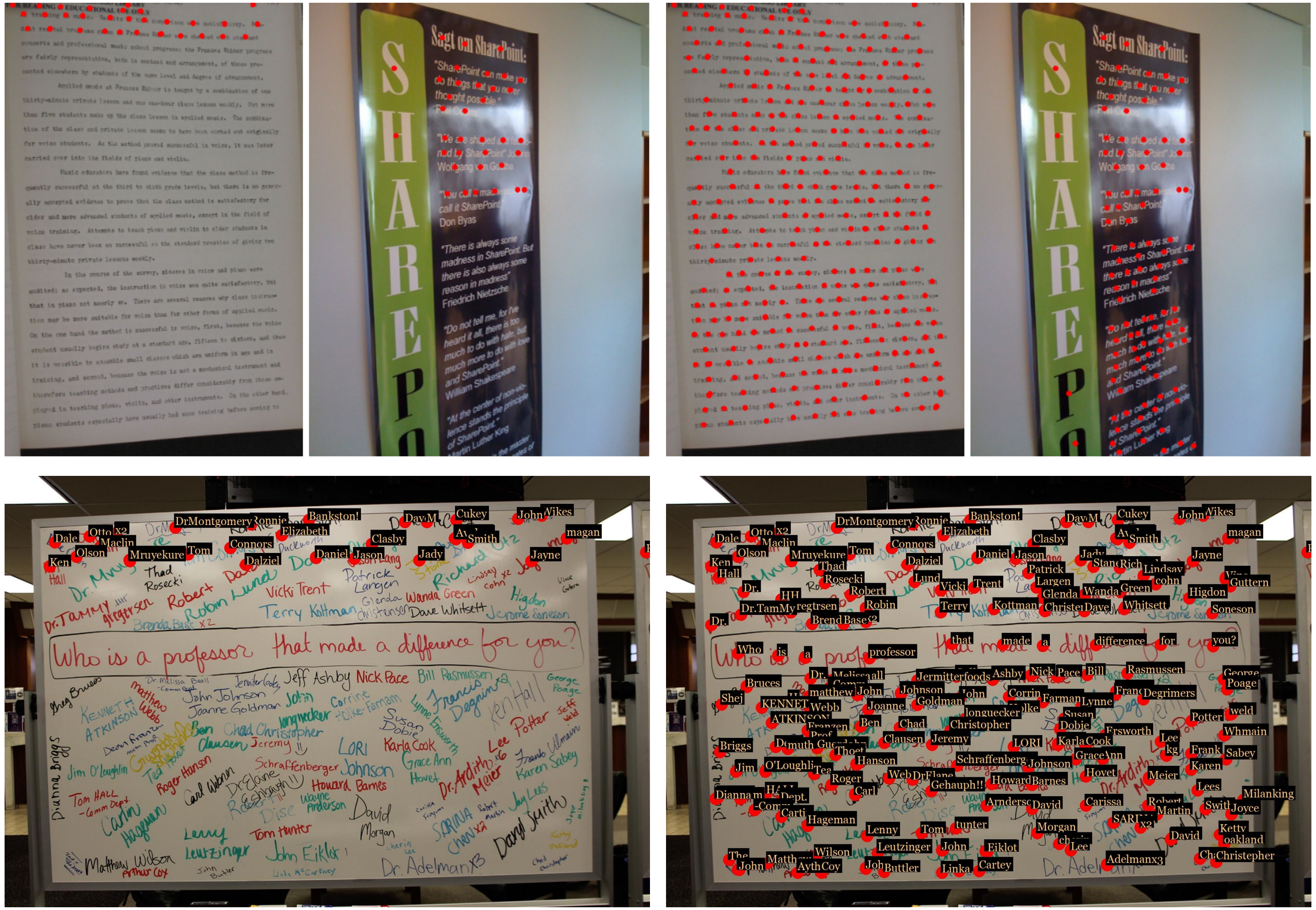
(d) polygon

The proposed method can extract text in several detection formats with a single model

Ablations

Starting-Point Prompt	End-to-End		
	Precision	Recall	F-measure
-	78.4	30.6	44.0
✓	**80.2**	**54.2**	**64.7**

Table 5. Ablation study of the starting-point prompting on TextOCR. The starting-point prompting enables UNITS to extract a large number of text instances even with a limited decoder length.



w/o starting-point prompt

w/ starting-point prompt

The proposed method overcomes limitations of existing methods by using the starting-point prompt

Qualitative results