JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

WED-PM-185

Towards Flexible Multi-modal Document Models

(Highlight)

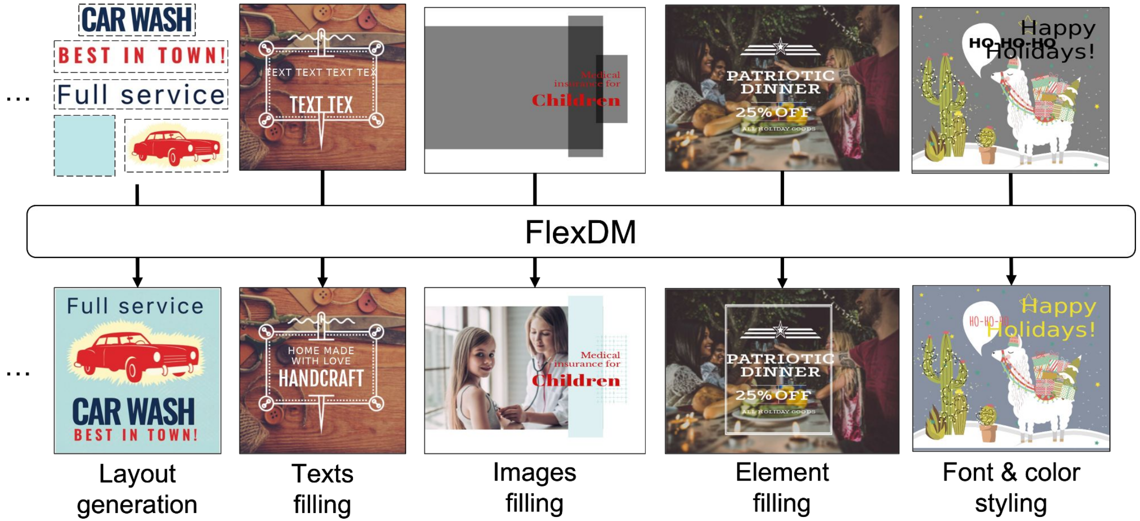
Naoto Inoue Kotaro Kikuchi Mayu Otani
Edgar Simo-Serra Kota Yamaguchi

CyberAgent AI Lab

WASEDA University
早稲田大学

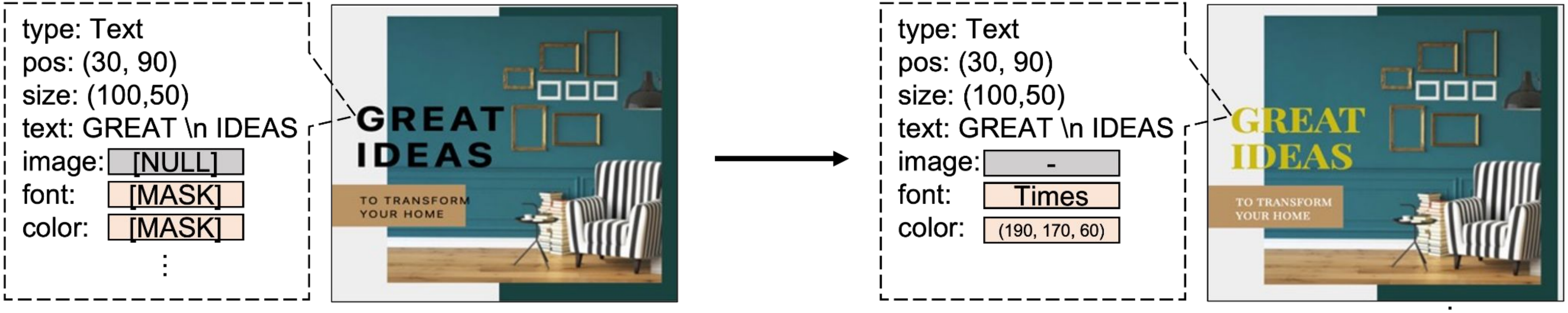
Flexible Document Model (FlexDM)

Our work: solve many design tasks in a single model



Key Idea of FlexDM

Multi-modal masked field prediction as a unified interface

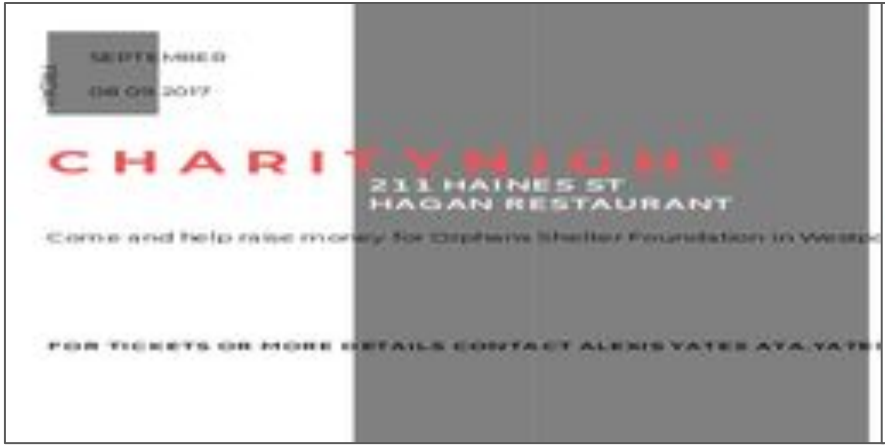


type: Text
pos: (30, 90)
size: (100,50)
text: GREAT \n IDEAS
image: [NULL]
font: [MASK]
color: [MASK]
⋮

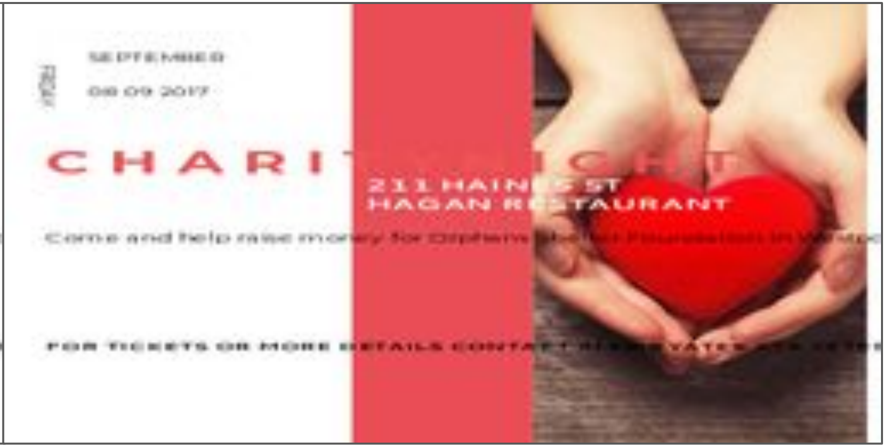
type: Text
pos: (30, 90)
size: (100,50)
text: GREAT \n IDEAS
image: -
font: Times
color: (190, 170, 60)

FlexDM Results

Input


FRIDAY
SEPTEMBER
08 09 2017
CHARITY NIGHT
211 HAINES ST
HAGAN RESTAURANT
Come and help raise money for Orphans Shelter Foundation in Westpo
FOR TICKETS OR MORE DETAILS CONTACT ALEXIS YATES ATA.YATE

Output

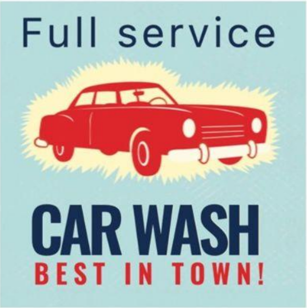

FRIDAY
SEPTEMBER
08 09 2017
CHARITY NIGHT
211 HAINES ST
HAGAN RESTAURANT
Come and help raise money for Orphans Shelter Foundation in Westpo
FOR TICKETS OR MORE DETAILS CONTACT ALEXIS YATES ATA.YATE

Vector Graphic Document

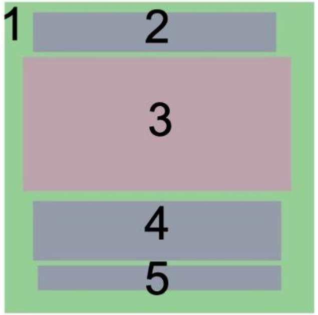
- A data format for making visual design (e.g., banner by Photoshop)
- Consists of a set of visual elements (+ global info) [Yamaguchi+, ICCV'21]
- Scalable, editable, human-interpretable

Rendering

Image



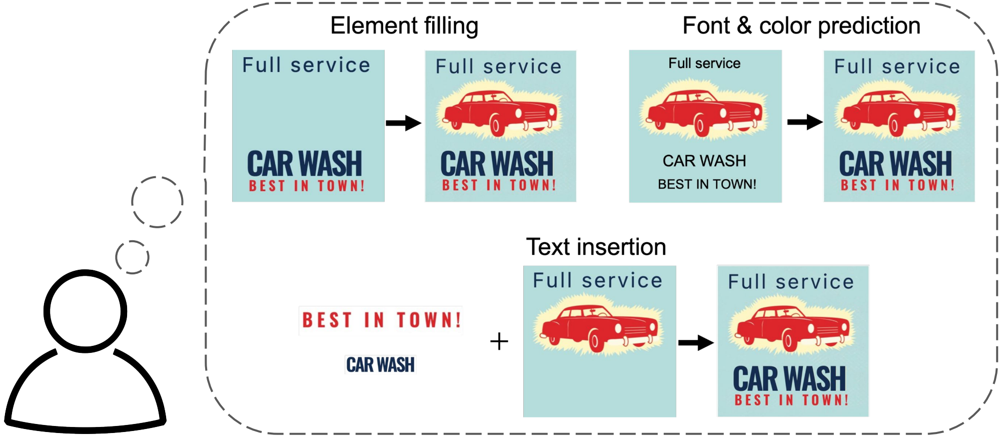
Layout



Vector graphic format

```
{
    "type": text, "position": [0.1, 0.6],
    "size": [0.8, 0.2], "text": "CAR WASH",
    "color": navy, "font_family": "Oswald", …
}, …
```

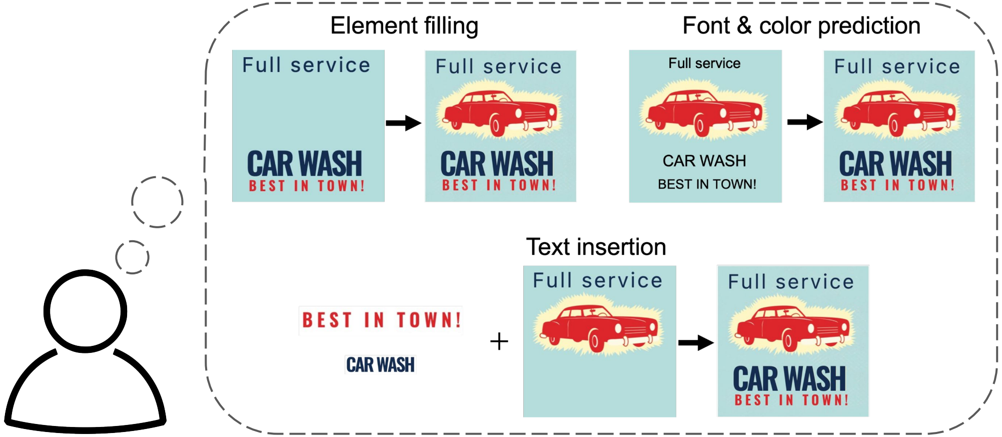
Design Tasks in Iterative Design Process



Design Tasks in Iterative Design Process

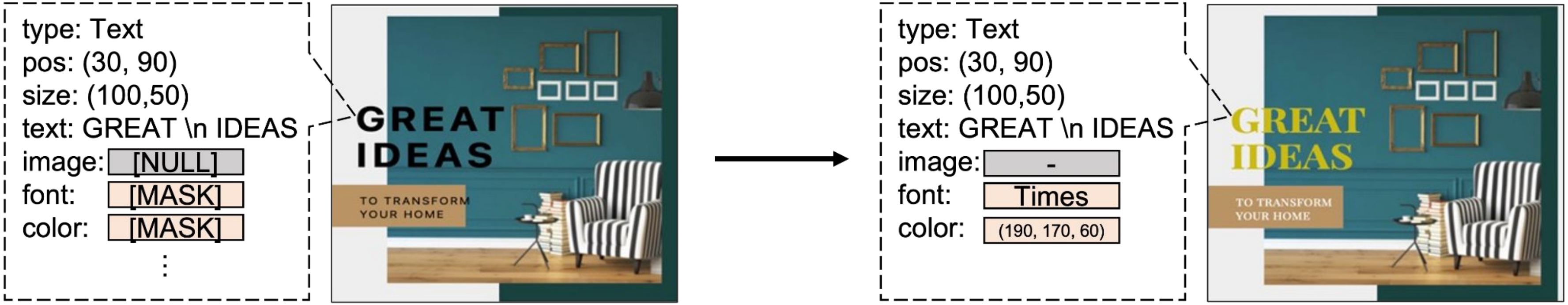
- High variety of possible actions
- Complex interaction between multi-modal elements

→ We handle design tasks in a principled manner



Masked Field prediction (MFP)

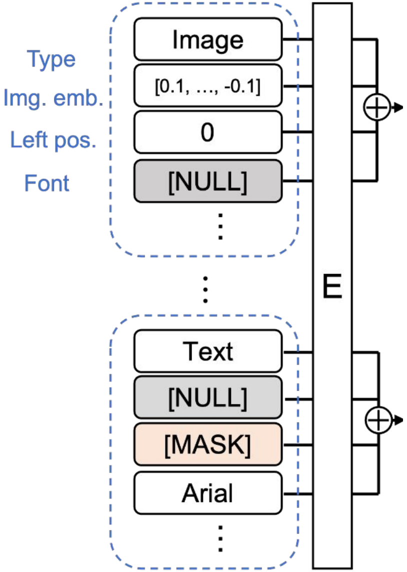
- **Predicting arbitrary number of fields hidden by [MASK]**
- **Challenges**
  - How to encode/decode various type of fields?
  - How to handle larger number of fields?



type: Text
pos: (30, 90)
size: (100,50)
text: GREAT \n IDEAS
image: [NULL]
font: [MASK]
color: [MASK]
⋮

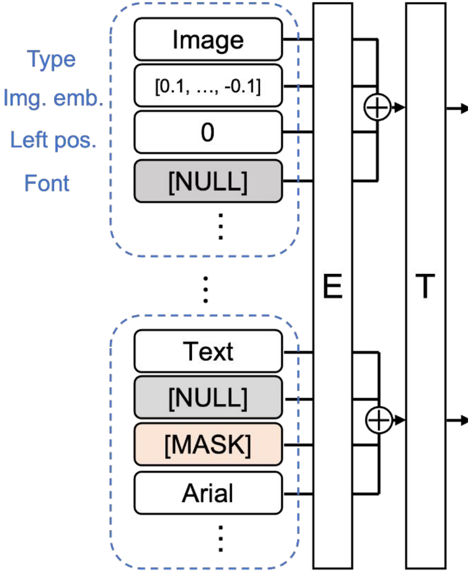
type: Text
pos: (30, 90)
size: (100,50)
text: GREAT \n IDEAS
image: -
font: Times
color: (190, 170, 60)

Network for Masked Field Prediction (MFP)



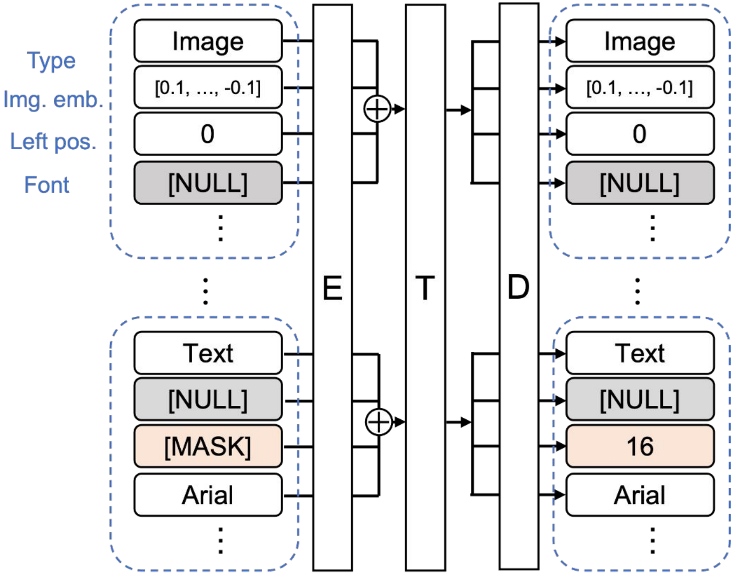
E: encoder

Network for Masked Field Prediction (MFP)



E: encoder, T: Transformer encoder

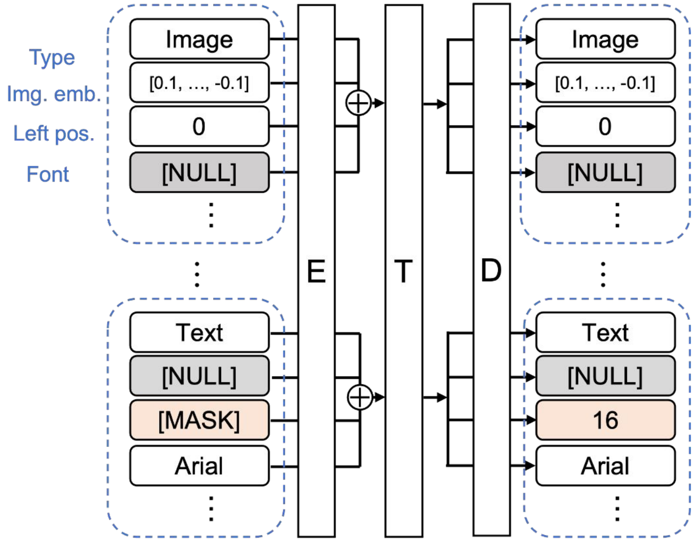
Network for Masked Field Prediction (MFP)



E: encoder, T: Transformer encoder, D: decoder

Challenges and solutions in MFP

- Various type of fields → attribute-specific enc. and dec.
- Large number of fields → consider interaction only in element-level



Training FlexDM

Training

1. In-domain pre-training (15% random masking)
2. Explicit multi-task learning for target design tasks

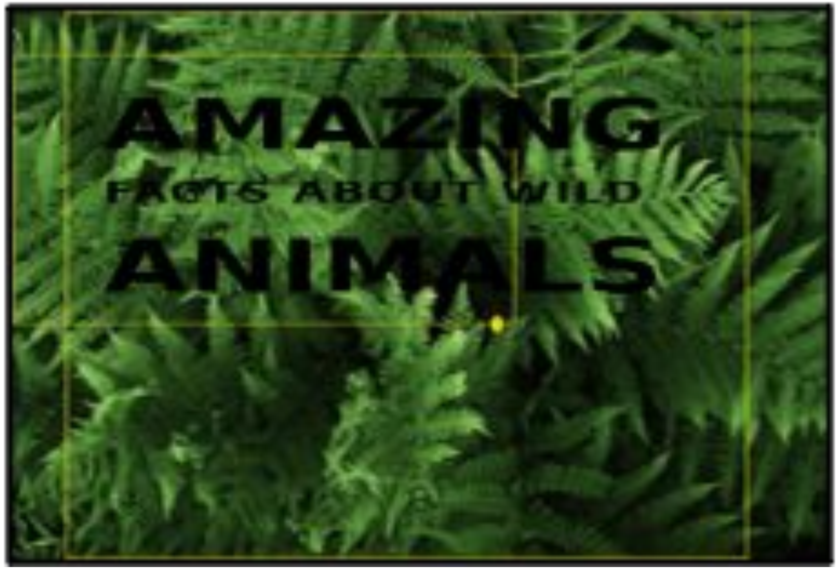
Loss: reconstruction error

Preprocess

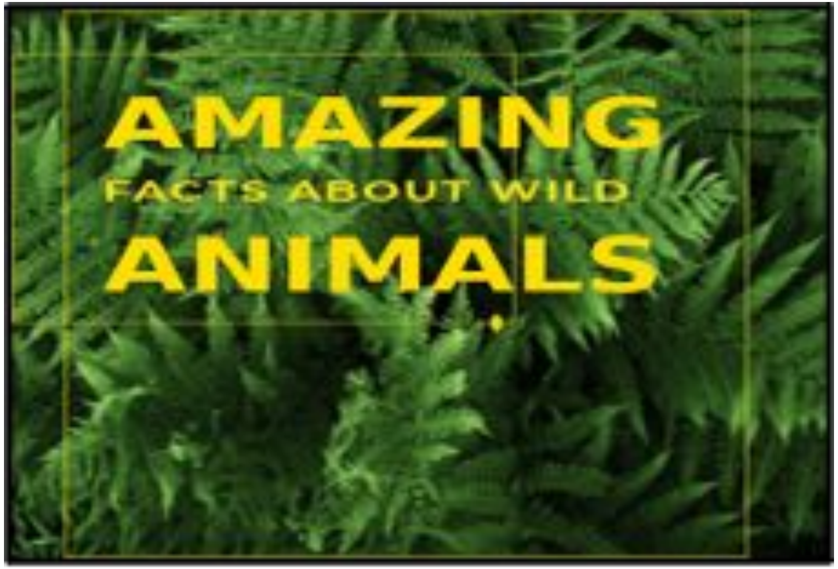
- Quantization for numerical attributes
- Feature extraction using pre-trained models for image and text

Attributes Prediction (ATTR)

Input

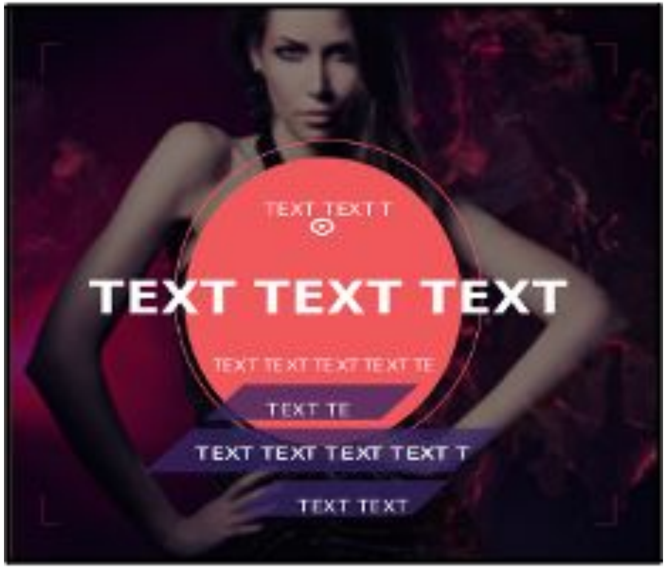


Output

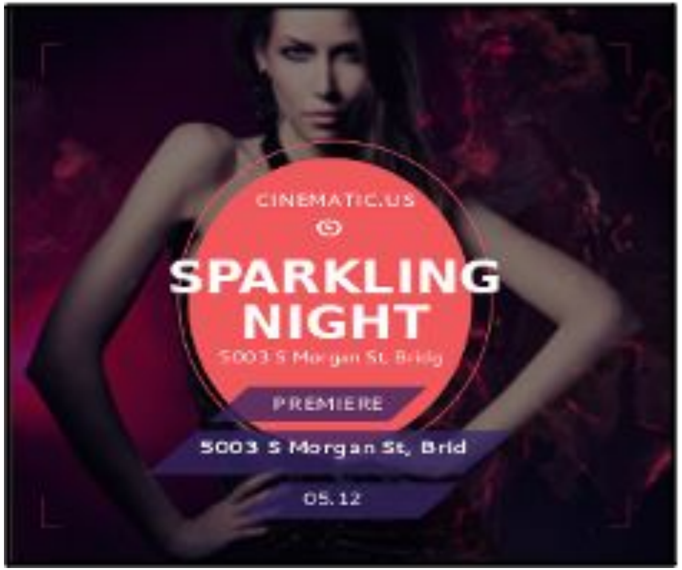


Texts Prediction (TXT)

Input

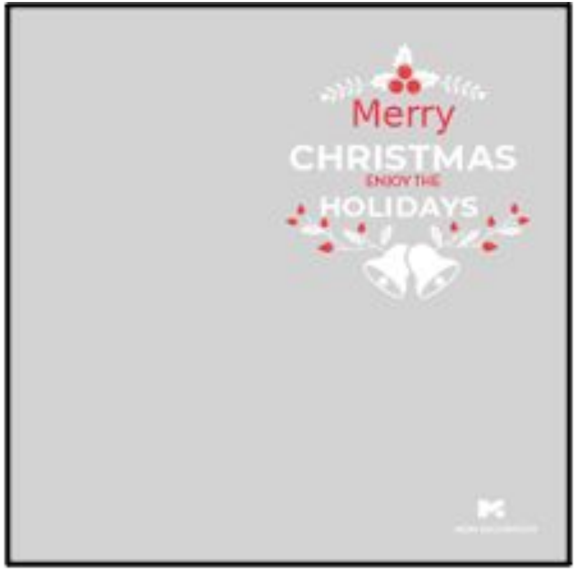


Output

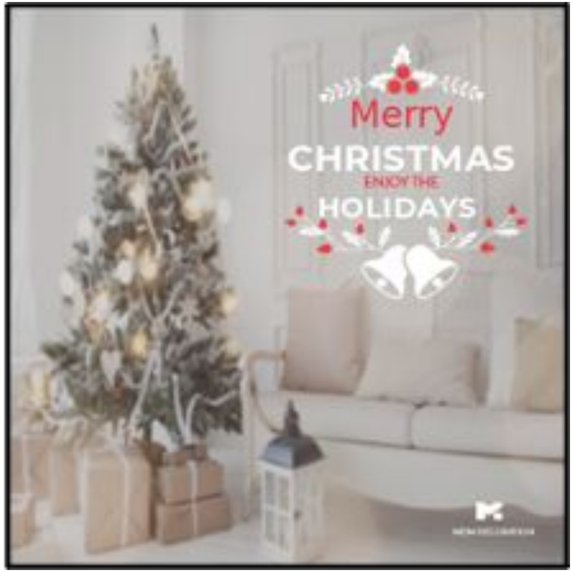


Element Filling (ELEM)

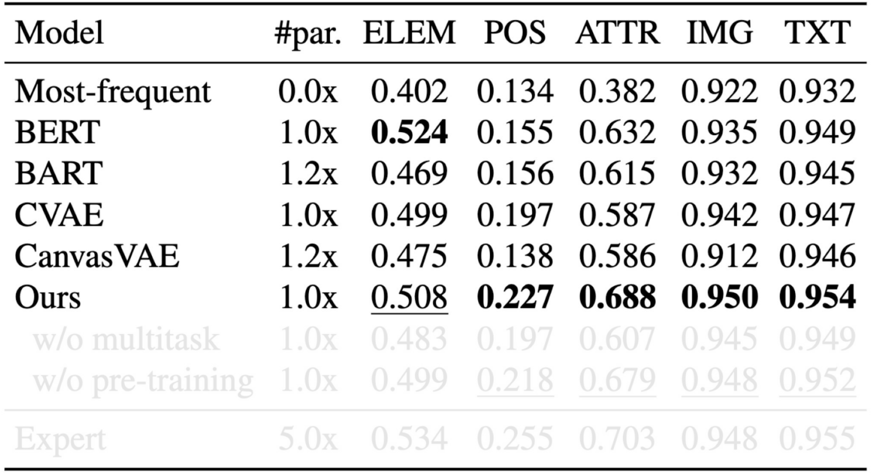
Input



Output



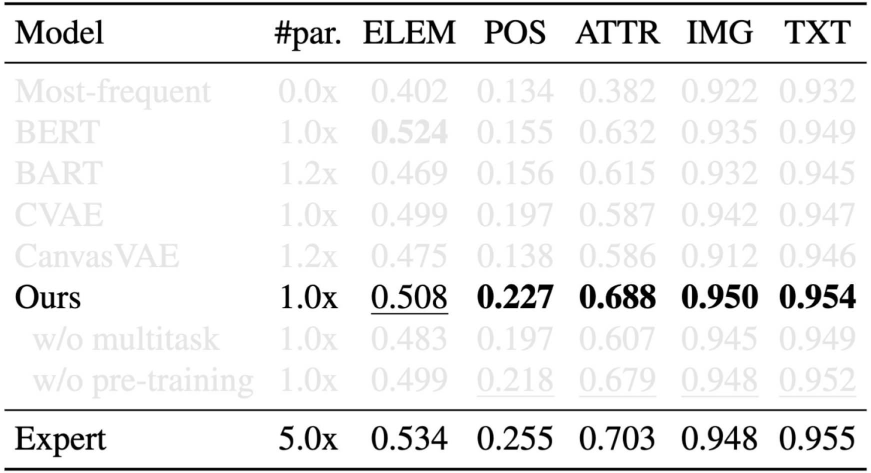
Quantitative Evaluation in Crello



Model	#par.	ELEM	POS	ATTR	IMG	TXT
Most-frequent	0.0x	0.402	0.134	0.382	0.922	0.932
BERT	1.0x	**0.524**	0.155	0.632	0.935	0.949
BART	1.2x	0.469	0.156	0.615	0.932	0.945
CVAE	1.0x	0.499	0.197	0.587	0.942	0.947
CanvasVAE	1.2x	0.475	0.138	0.586	0.912	0.946
Ours	1.0x	0.508	**0.227**	**0.688**	**0.950**	**0.954**
w/o multitask	1.0x	0.483	0.197	0.607	0.945	0.949
w/o pre-training	1.0x	0.499	0.218	0.679	0.948	0.952
Expert	5.0x	0.534	0.255	0.703	0.948	0.955

1. Much better than baselines
2. Almost close to task-specific expert
3. Both components are important

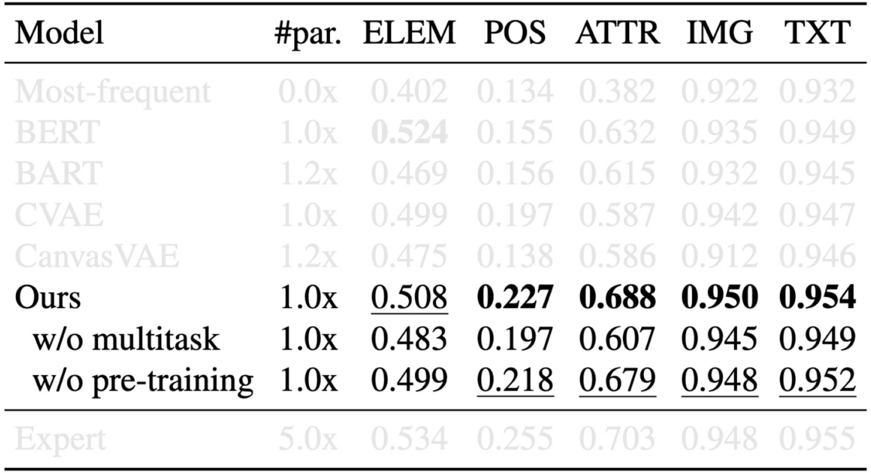
Quantitative Evaluation in Crello



Model	#par.	ELEM	POS	ATTR	IMG	TXT
Most-frequent	0.0x	0.402	0.134	0.382	0.922	0.932
BERT	1.0x	**0.524**	0.155	0.632	0.935	0.949
BART	1.2x	0.469	0.156	0.615	0.932	0.945
CVAE	1.0x	0.499	0.197	0.587	0.942	0.947
CanvasVAE	1.2x	0.475	0.138	0.586	0.912	0.946
Ours	1.0x	0.508	**0.227**	**0.688**	**0.950**	**0.954**
w/o multitask	1.0x	0.483	0.197	0.607	0.945	0.949
w/o pre-training	1.0x	0.499	0.218	0.679	0.948	0.952
Expert	5.0x	0.534	0.255	0.703	0.948	0.955

1. Much better than baselines
2. Almost close to task-specific expert
3. Both components are important

Quantitative Evaluation in Crello



Model	#par.	ELEM	POS	ATTR	IMG	TXT
Most-frequent	0.0x	0.402	0.134	0.382	0.922	0.932
BERT	1.0x	**0.524**	0.155	0.632	0.935	0.949
BART	1.2x	0.469	0.156	0.615	0.932	0.945
CVAE	1.0x	0.499	0.197	0.587	0.942	0.947
CanvasVAE	1.2x	0.475	0.138	0.586	0.912	0.946
Ours	1.0x	0.508	**0.227**	**0.688**	**0.950**	**0.954**
w/o multitask	1.0x	0.483	0.197	0.607	0.945	0.949
w/o pre-training	1.0x	0.499	0.218	0.679	0.948	0.952
Expert	5.0x	0.534	0.255	0.703	0.948	0.955

1. Much better than baselines
2. Almost close to task-specific expert
3. Both components are important

Summary

- Masked field prediction (MFP) as a unified interface
- A model handling larger number of fields and tasks efficiently
- Promising performance in various documents (e.g., banner, web, …)

Check codes and more results at

https://cyberagentailab.github.io/flex-dm/