Generative models for assisting graphic design

Naoto Inoue (CyberAgent Inc.)

June 28, Invited talk III at SNL2023

CyberAgent AI Lab

Self Introduction

Career:

- ~ Mar. 2021: Ph. D at The University of Tokyo
- Apr. 2021 ~ : Research Scientist at CyberAgent AI Lab
  - Research related to creating advertisements

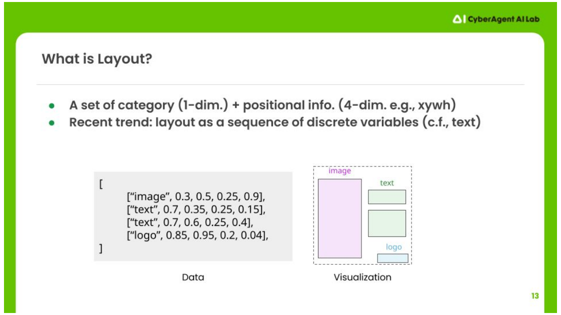
Research interest:

- generative models for graphic design

Graphic Design

- Visual + textual content
- Important to convey ideas

banner ad



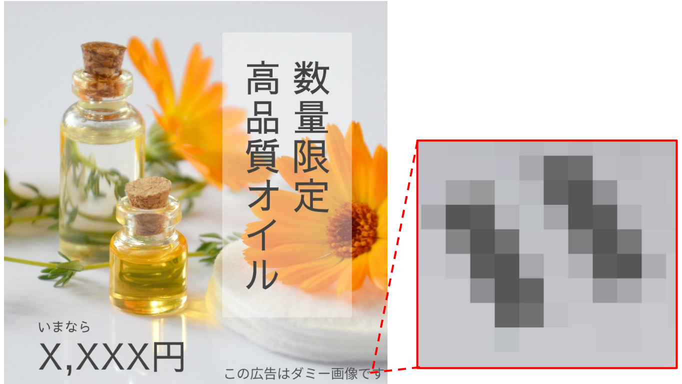
presentation

meme (credit)

Raster v.s. Vector Format

Raster

- for display
- e.g., .jpg, .png, ...



Raster v.s. Vector Format

Raster

- for display
- e.g., .jpg, .png, …

Vector

- for edit
- e.g., .pptx, .psd, .svg, …

```
<svg>
  <image xlink: href="…">
  <rect x=55 y =10 … ></rect>
  <text x=10 … >X,XXX円</text>
  …
</svg>
```

Graphic Design in Vector Format

Features

- Multi-modal attributes
- Large number of elements

Research question

- How to generate vector graphic document?

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

WED-AM-185

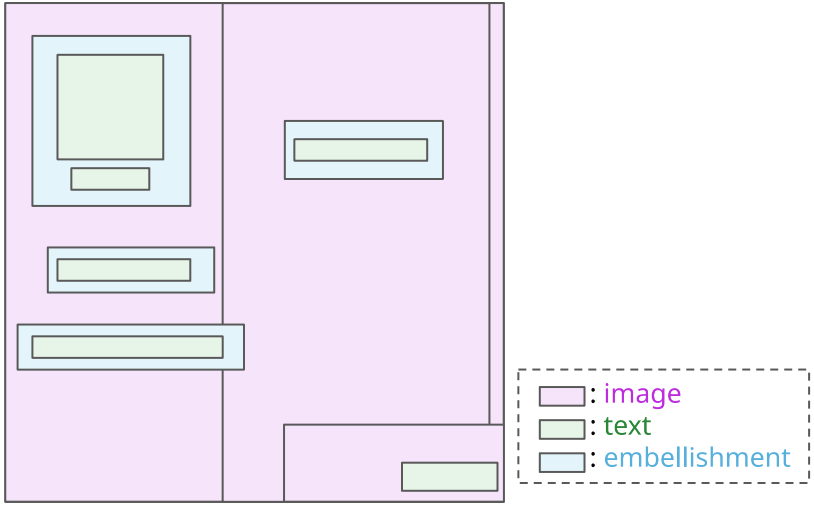
LayoutDM: Discrete Diffusion Model for Controllable Layout Generation

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

CyberAgent AI Lab
WASEDA University
早稲田大学

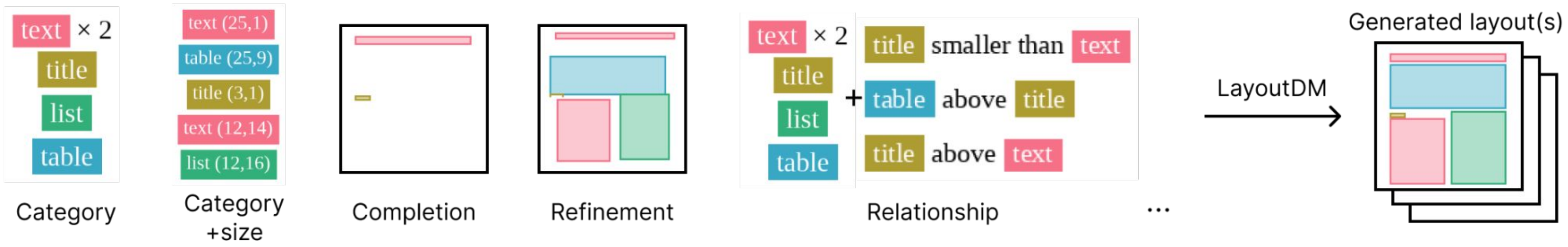
Layout

= Simple yet essential interface to understand & control visual design



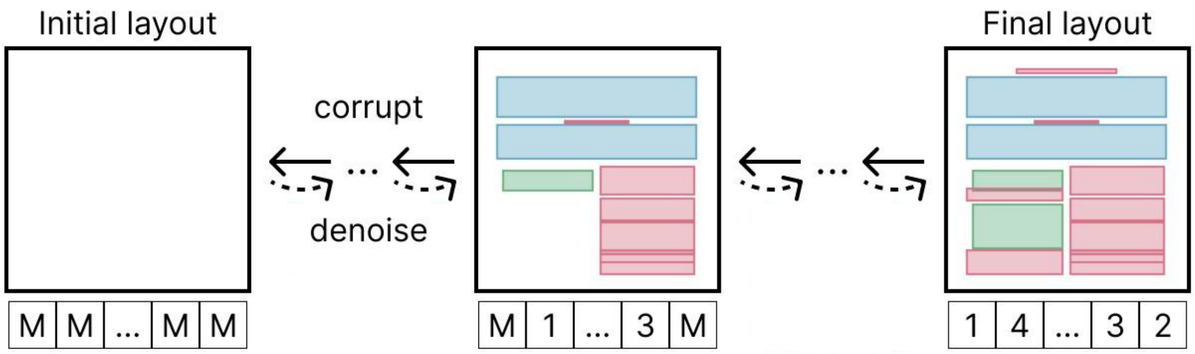
Controllable Layout Generation

Our work: solve a broad range of tasks in a single model



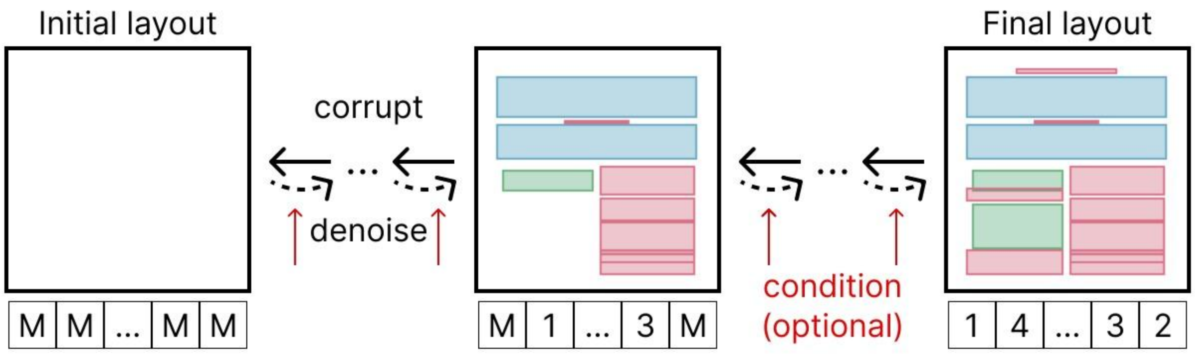
LayoutDM

- A discrete diffusion model tamed for layout generation



LayoutDM

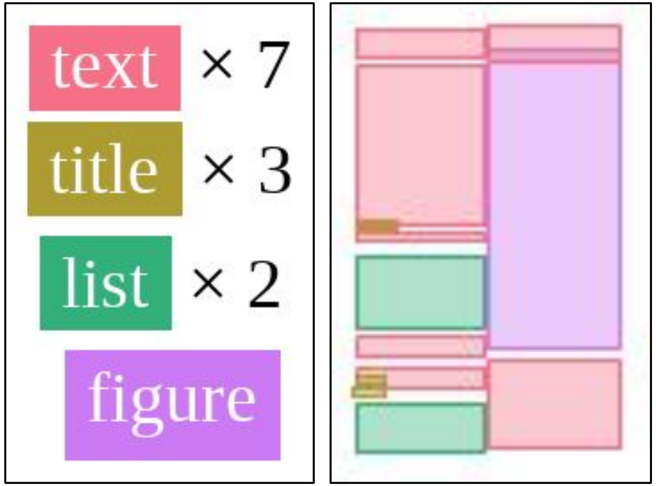
- A discrete diffusion model tamed for layout generation
- Training-free algorithm to inject various conditions during inference



LayoutDM Results

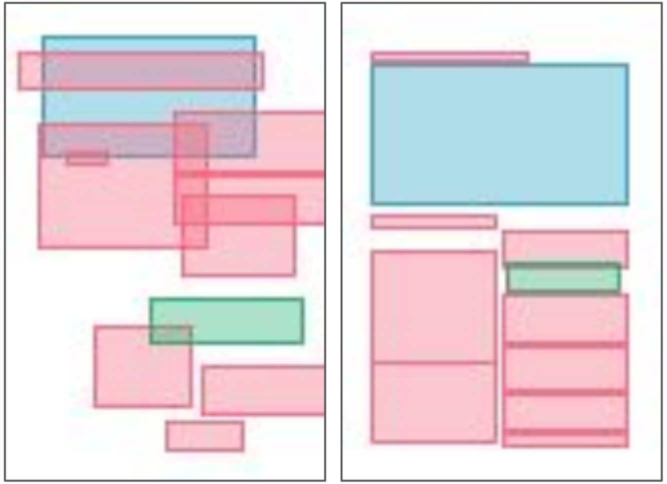
Input:
category

Output



Input:
noisy layout

Output

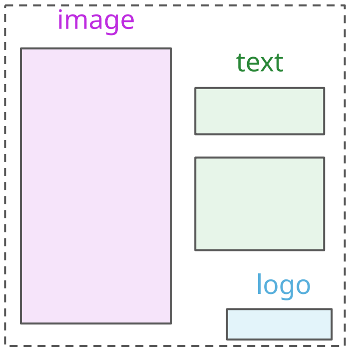


What is Layout?

- A set of category (1-dim.) + positional info. (4-dim. e.g., xywh)
- Recent trend: layout as a sequence of discrete variables (c.f., text)

```
[
    ["image", 0.3, 0.5, 0.25, 0.9],
    ["text", 0.7, 0.35, 0.25, 0.15],
    ["text", 0.7, 0.6, 0.25, 0.4],
    ["logo", 0.85, 0.95, 0.2, 0.04],
]
```

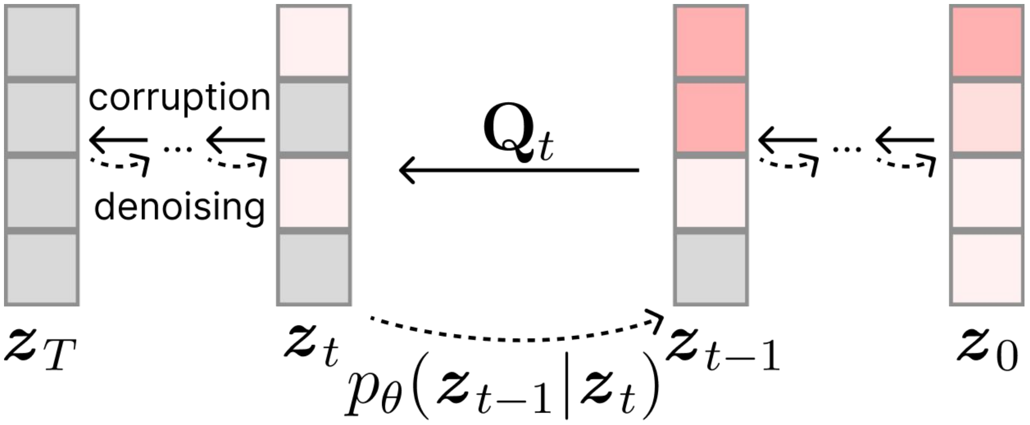
Data



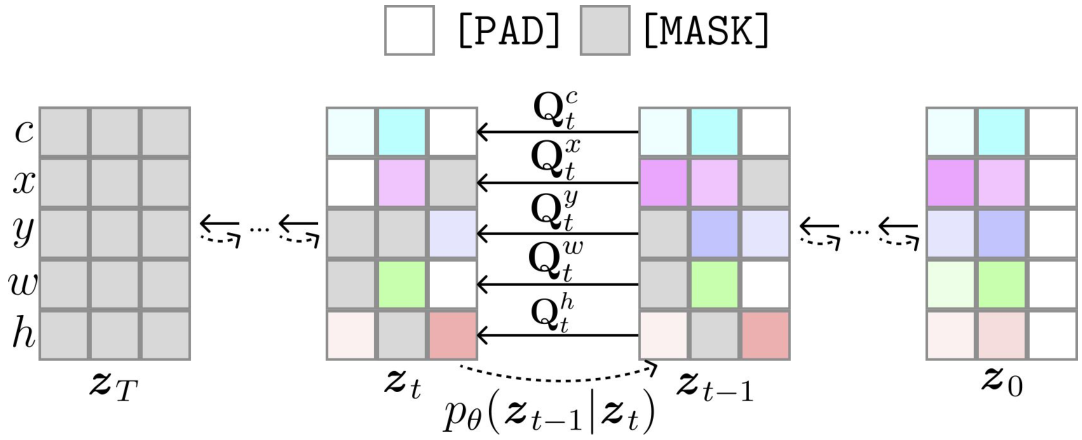
Visualization

Discrete Diffusion Models [Austin+, NeurIPS'21]

- = diffusion models for modeling categorical variables (e.g., text)
- Corruption: a token is stochastically replaced with another in vocabulary

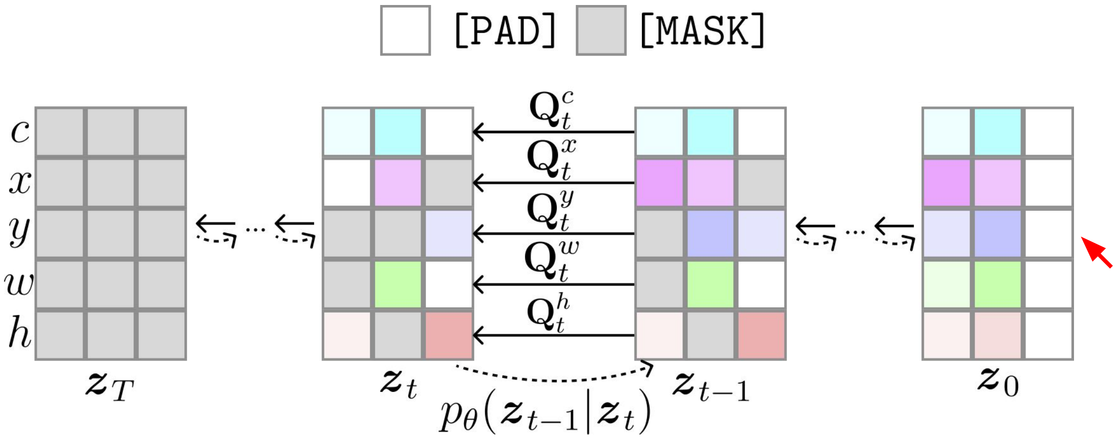


Adapting Discrete Diffusion Models for Layout



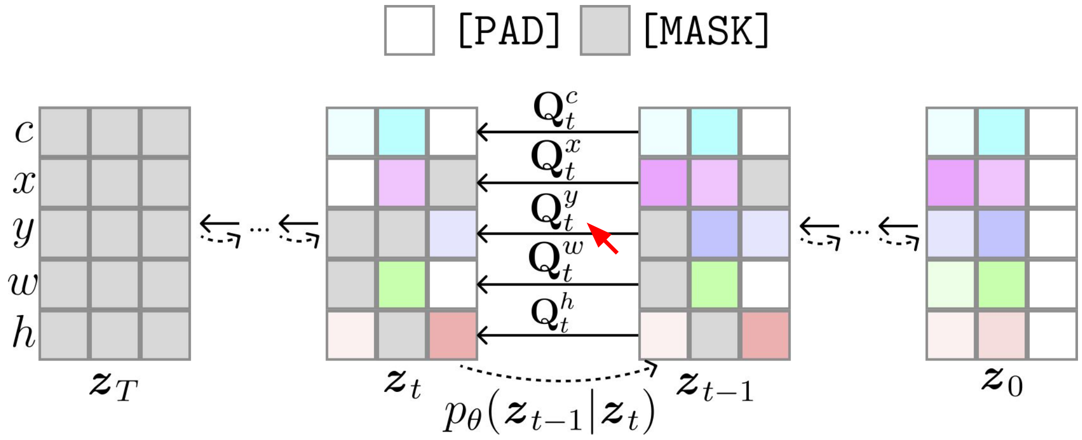
Adapting Discrete Diffusion Models for Layout

- [PAD] token to enable variable length generation

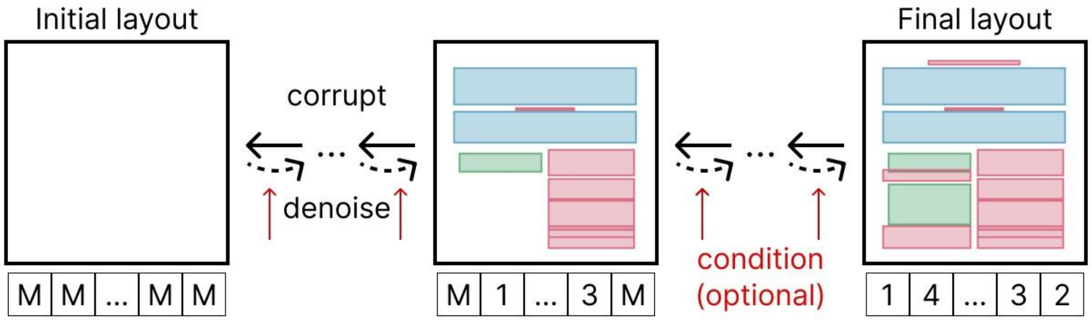


Adapting Discrete Diffusion Models for Layout

- [PAD] token to enable variable length generation
- Modality-wise corruption process

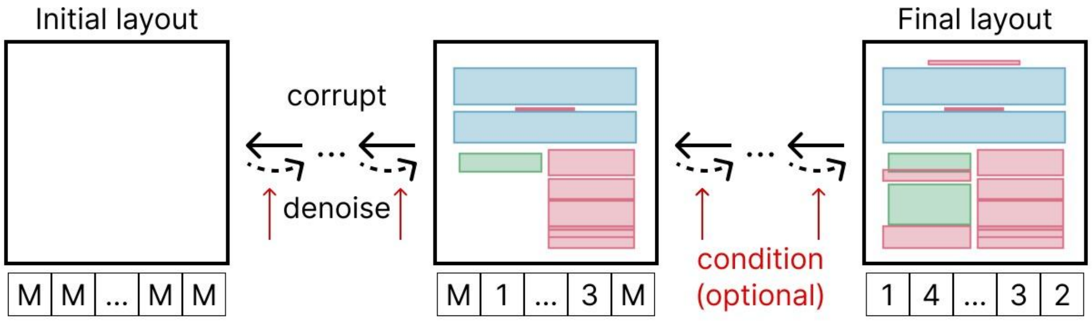


How to Feed Conditions during Inference?



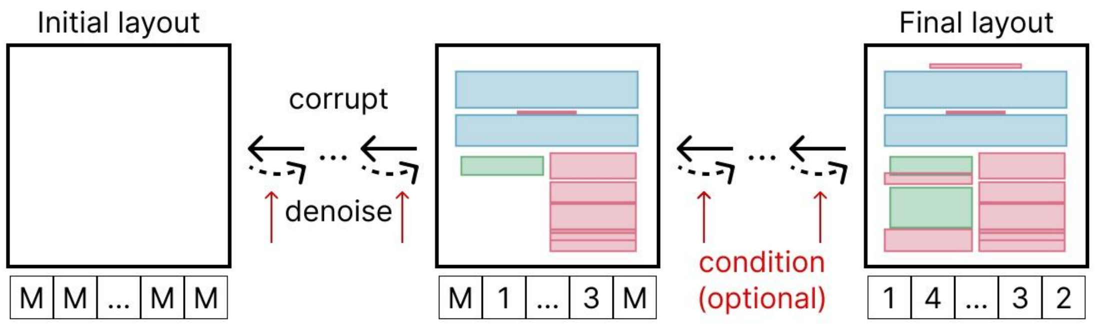
How to Feed Conditions during Inference?

- **Hard condition: masking**
  - e.g., “i-th element’s category is C”



How to Feed Conditions during Inference?

- **Hard condition: masking**
  - e.g., “i-th element’s category is C”
- **Soft condition: logit adjustment**
  - e.g., “an element at the top”, “an element bigger than another”

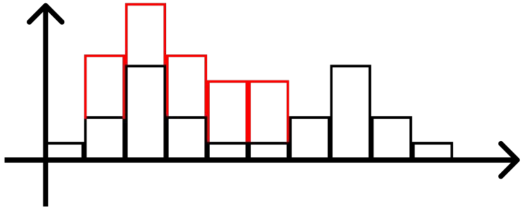


Logit Adjustment

Inject soft condition as a prior term

$$\log \hat{p}_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \log p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) + \underbrace{\lambda_\pi \boldsymbol{\pi}}_{\text{Prior term}}$$

$$\boldsymbol{z}_{t-1} \sim \hat{p}_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$$

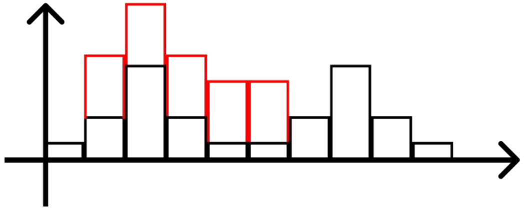


Logit Adjustment

Inject soft condition as a prior term

$$\log \hat{p}_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) = \log p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t) + \lambda_\pi \boldsymbol{\pi}$$
$$\boldsymbol{z}_{t-1} \sim \hat{p}_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$$

Prior term



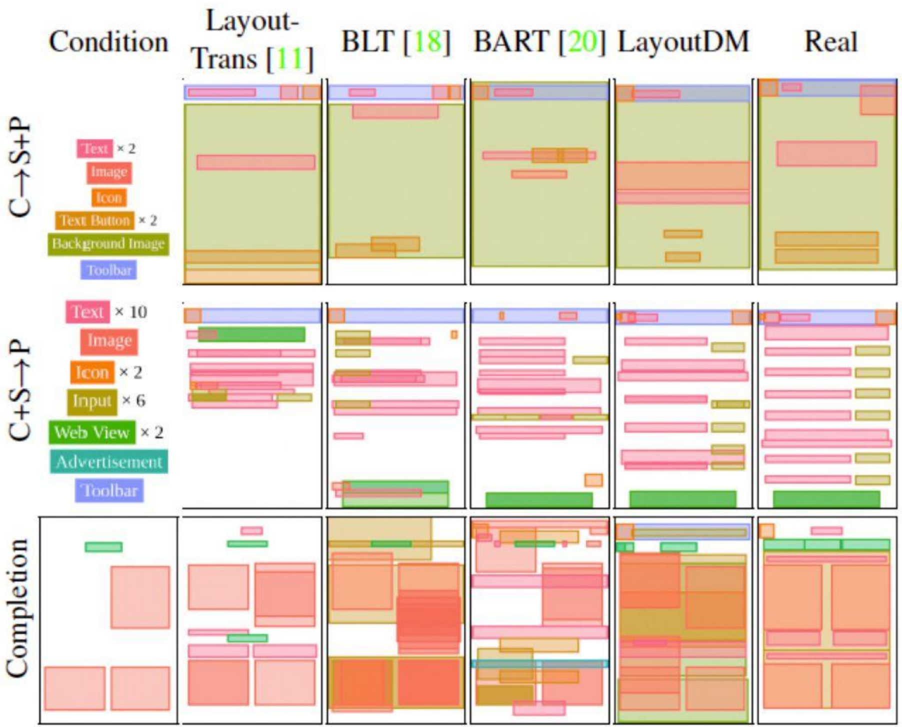
How to implement a prior?

- Hard coding (e.g., refinement task)
- Gradients from loss functions w.r.t. the prediction (e.g., relationship task)

Advantages over Existing Methods

- **No fixed generation order unlike auto-regressive models**
  - c.f., LayoutTransformer [Gupta+, ICCV'21]
- **Flexibly changing the number of elements to be generated**
  - c.f., BLT [Kong+, ECCV'22]
- **Incorporating both hard and soft conditions**
  - c.f., NDN [Lee+, ECCV'20]

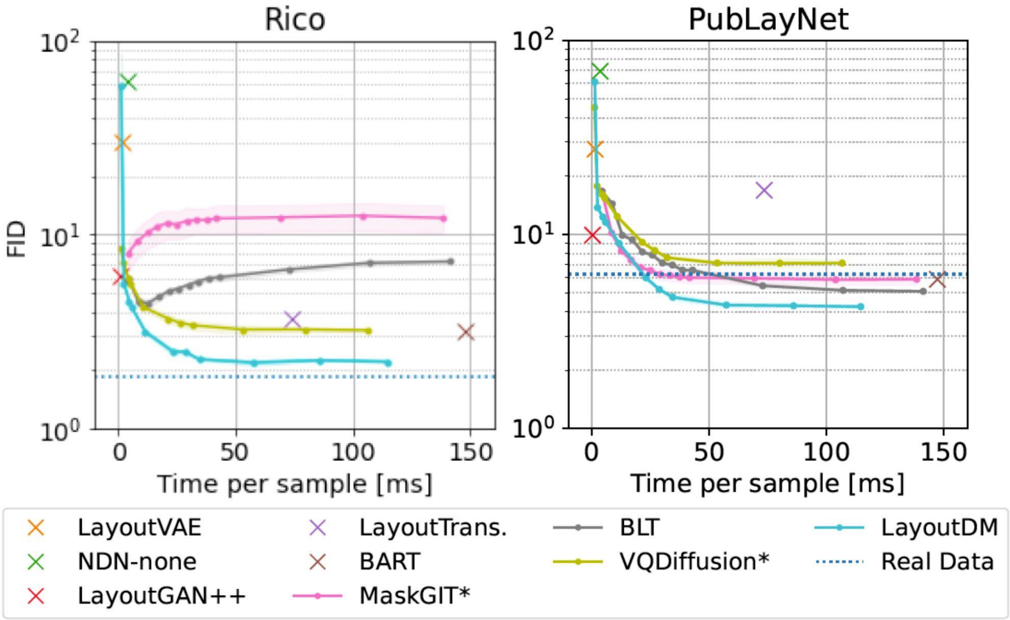
Results in Rico [Deka+, UIST'17]



Results in PubLayNet [Zhong+, ICDAR'19]

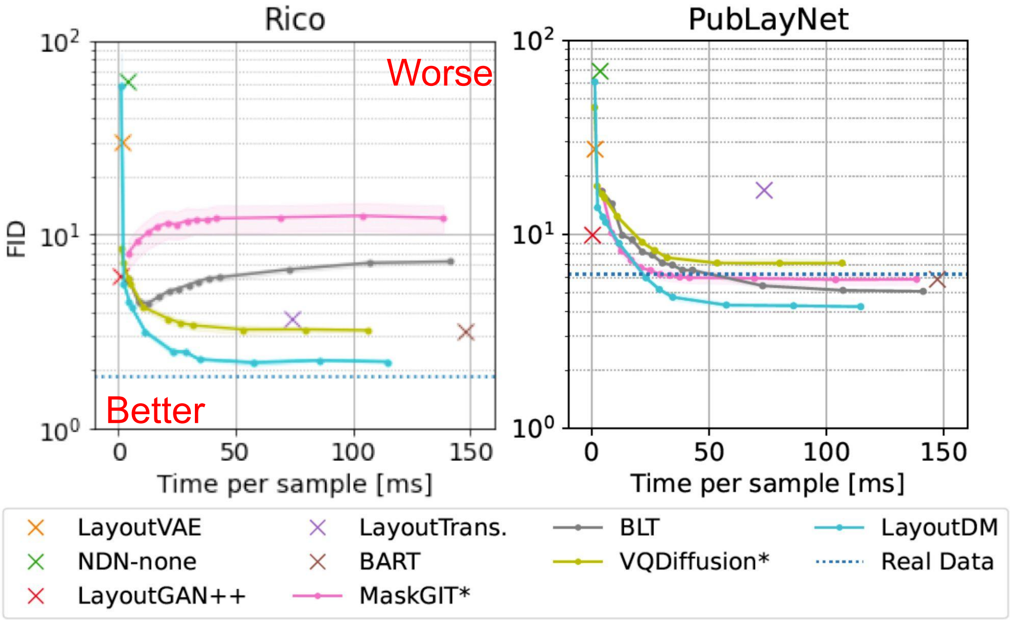
Quantitative Evaluation (in category + size → position)

LayoutDM achieves the best speed-quality tradeoff



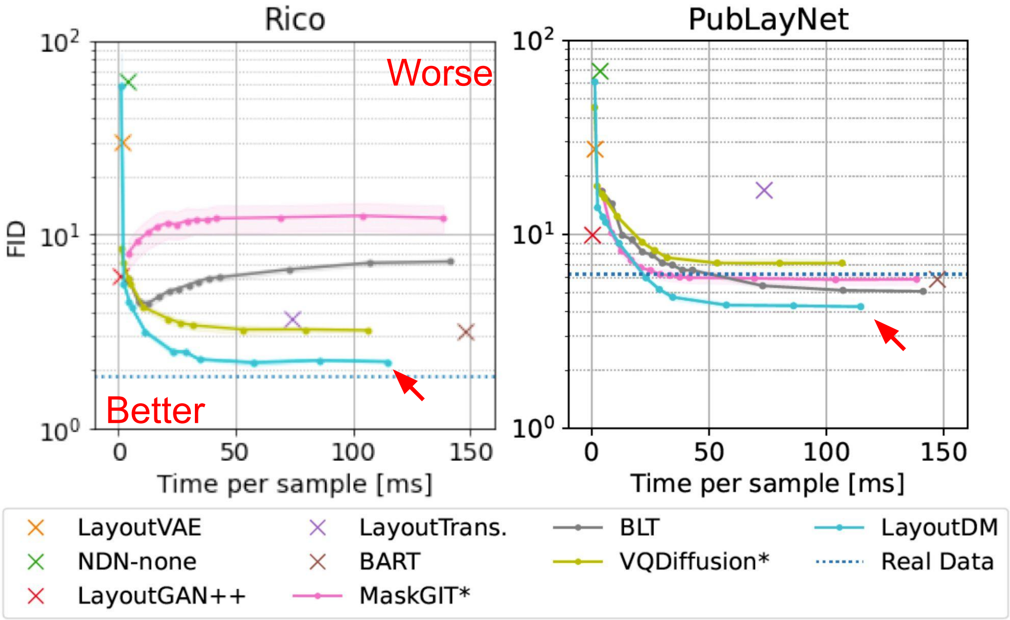
Quantitative Evaluation (in category + size → position)

LayoutDM achieves the best speed-quality tradeoff



Quantitative Evaluation (in category + size → position)

LayoutDM achieves the best speed-quality tradeoff



Summary

- A discrete diffusion model tamed for layout generation
- Training-free algorithm to inject various conditions during inference
- Favorable performance against task-specific/agnostic baselines

Check codes and more results at

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

WACV WAIKOLOA HAWAII JAN 3-7 • 2023

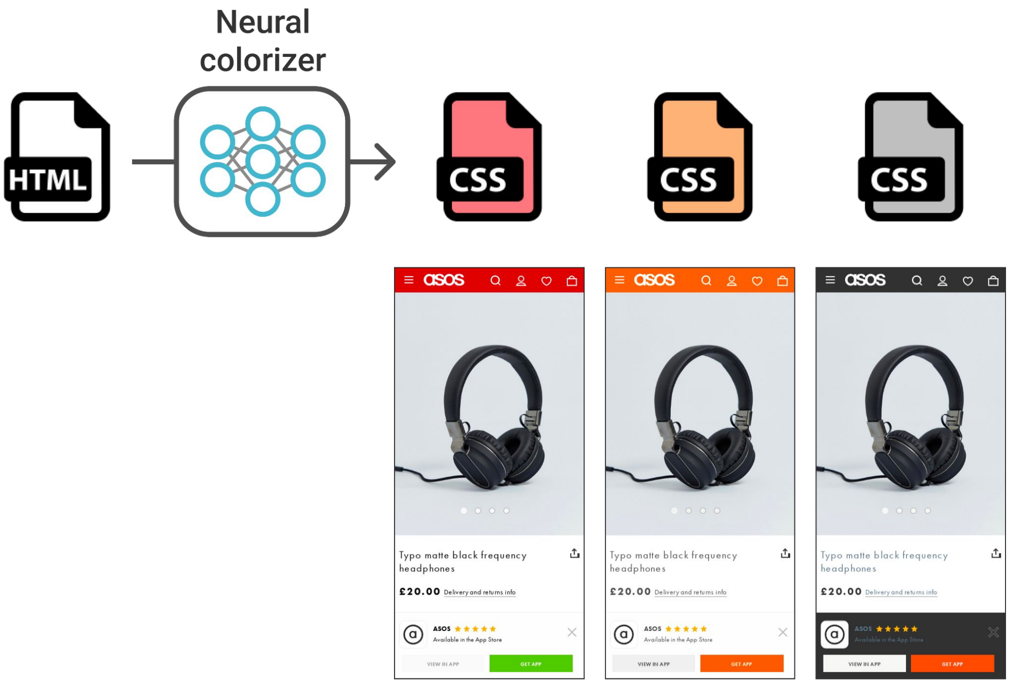
Generative Colorization of Structured Mobile Web Pages

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

CyberAgent AI Lab

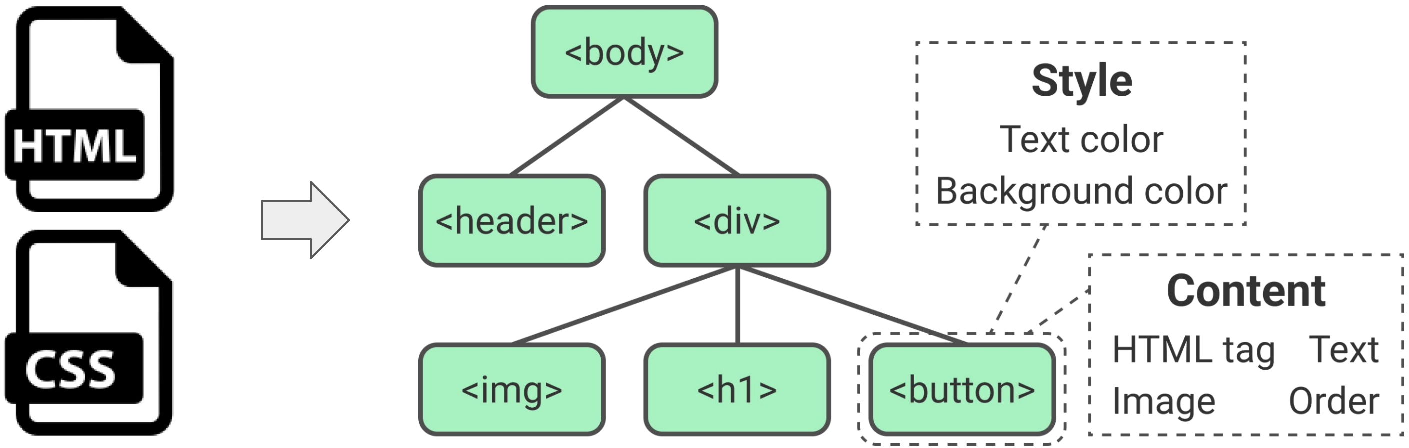
WASEDA University
早稲田大学

Web page colorization



Data structure of web page

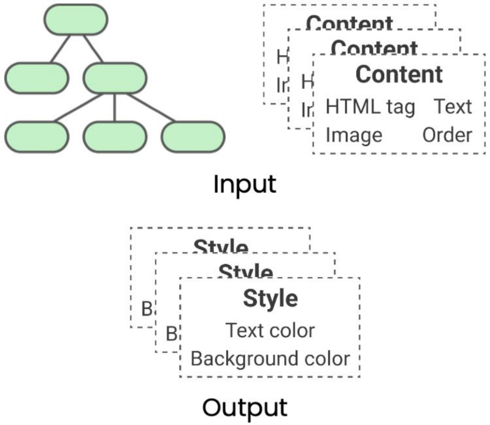
= Tree structure where each element has style and content information

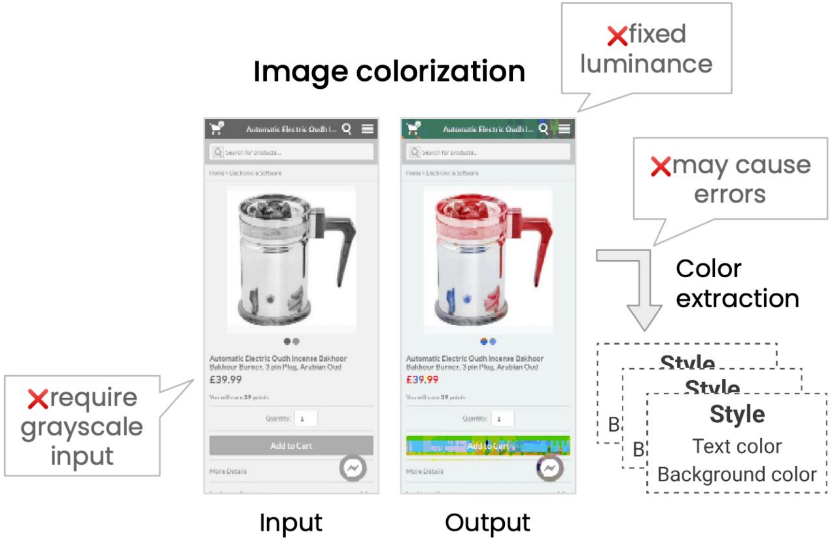


Structured color prediction

Generate color styles given content and hierarchical structure of elements

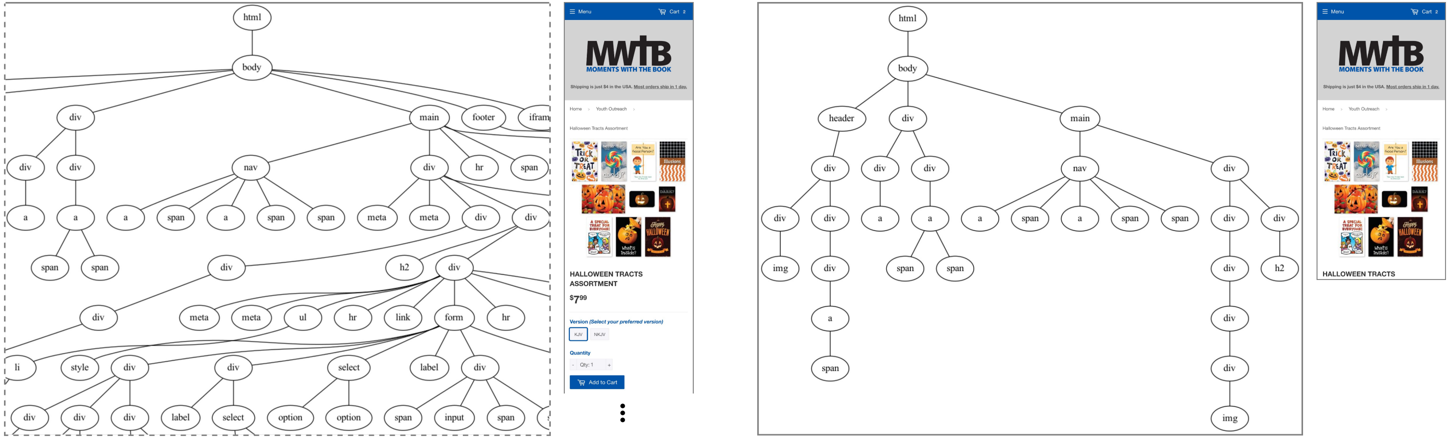
Web page colorization (ours)





New dataset for web page colorization

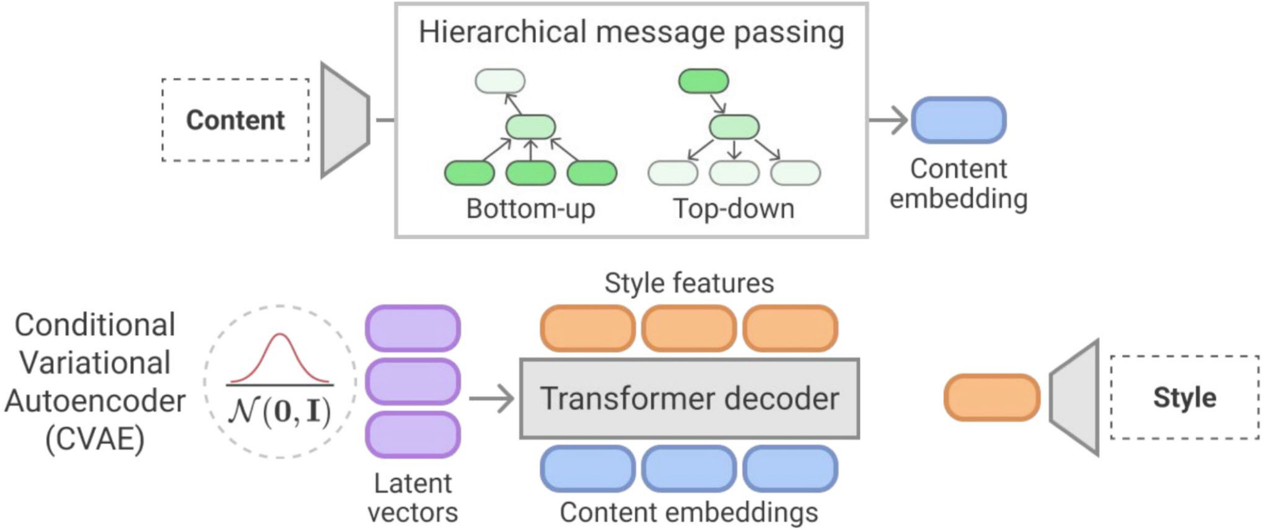
- E-commerce mobile web pages adapted from [Hotti+, arXiv'21]
- Convert to a tractable data format



e.g., keep only elements that contribute to the first view (Avg. elements: 1656 → 61)

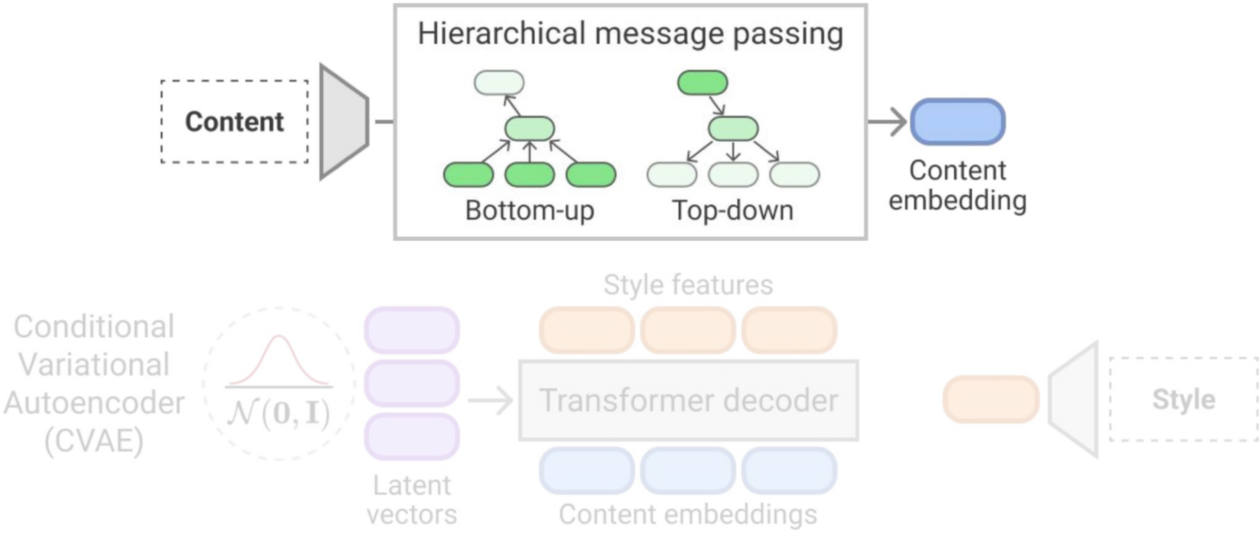
Proposed Hierarchy-aware CVAE model

Generate color styles given content and hierarchical structure of elements



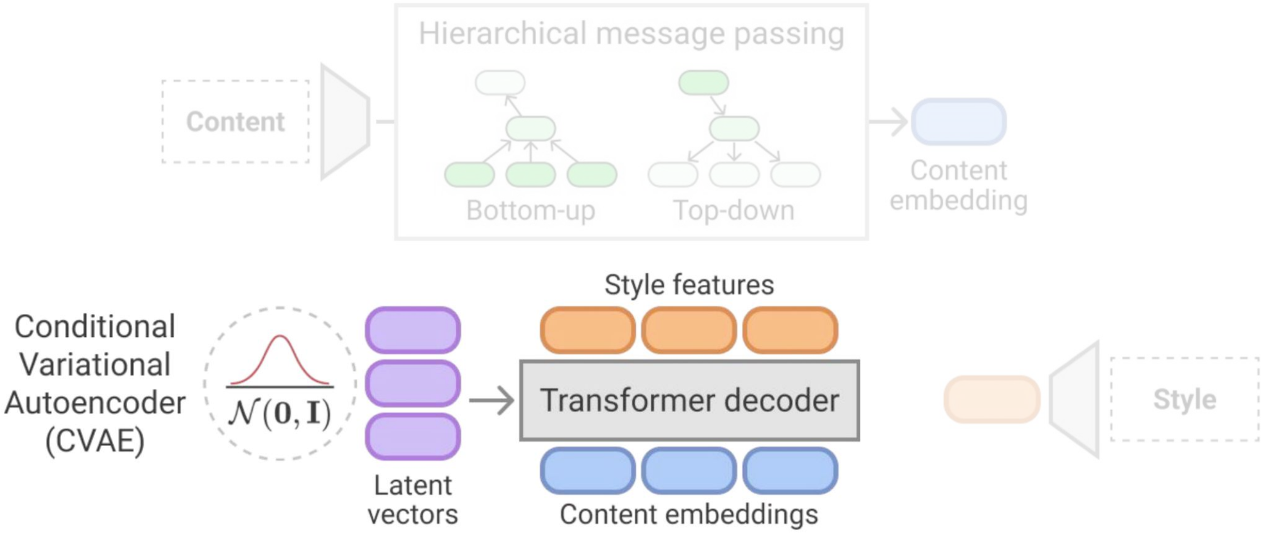
Proposed Hierarchy-aware CVAE model

Embed content with message passing to capture hierarchical relationships



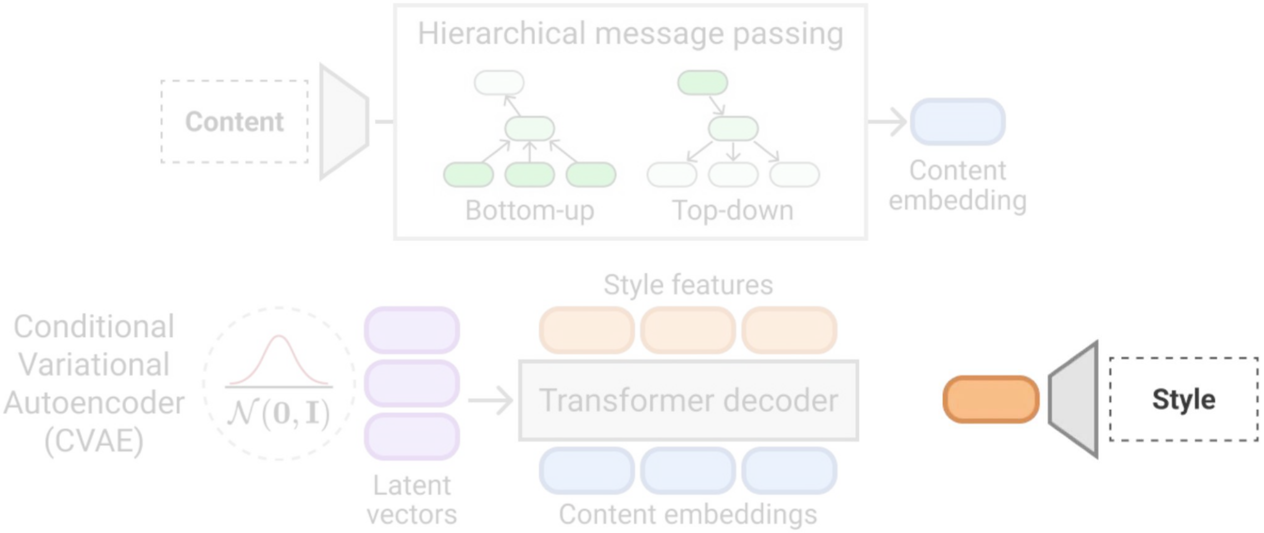
Proposed Hierarchy-aware CVAE model

Compute style features with the content embeddings and latent vectors



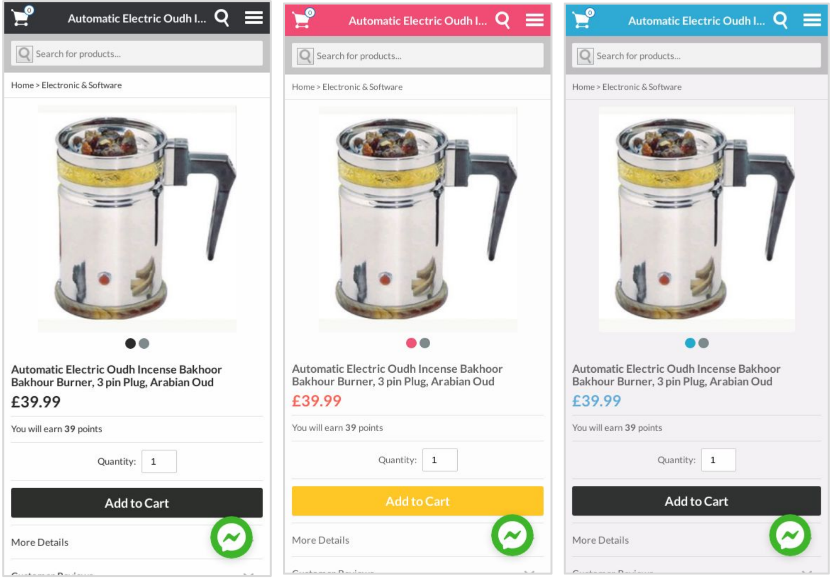
Proposed Hierarchy-aware CVAE model

Predict color style for each element based on the style feature



Experimental Results

Method	Accuracy RGB	Accuracy Alpha
Dataset statistics	0.621	0.821
Image colorization [1]	0.285	0.411
Ours (CVAE)	**0.771**	**0.929**



Ours (CVAE)

Summary

- New dataset for web page colorization
- Generate color styles given content and hierarchical structure of elements
- Our hierarchy-aware CVAE model performs better than baselines

Dataset, code, and pre-trained models are available!

https://github.com/CyberAgentAILab/webcolor

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

WED-PM-185

Towards Flexible Multi-modal Document Models

(Highlight)

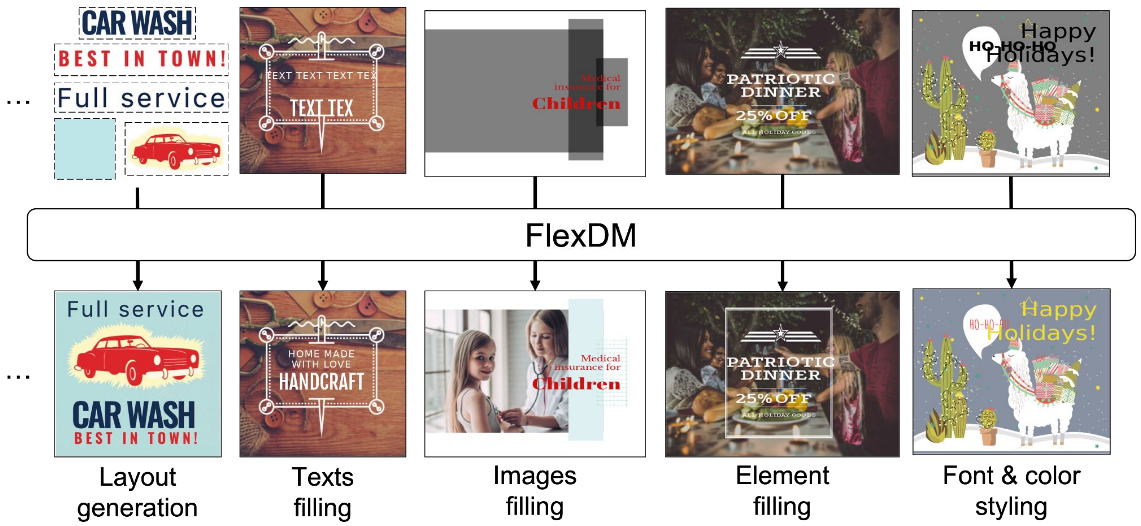
Naoto Inoue Kotaro Kikuchi Mayu Otani
Edgar Simo-Serra Kota Yamaguchi

CyberAgent AI Lab

WASEDA University
早稲田大学

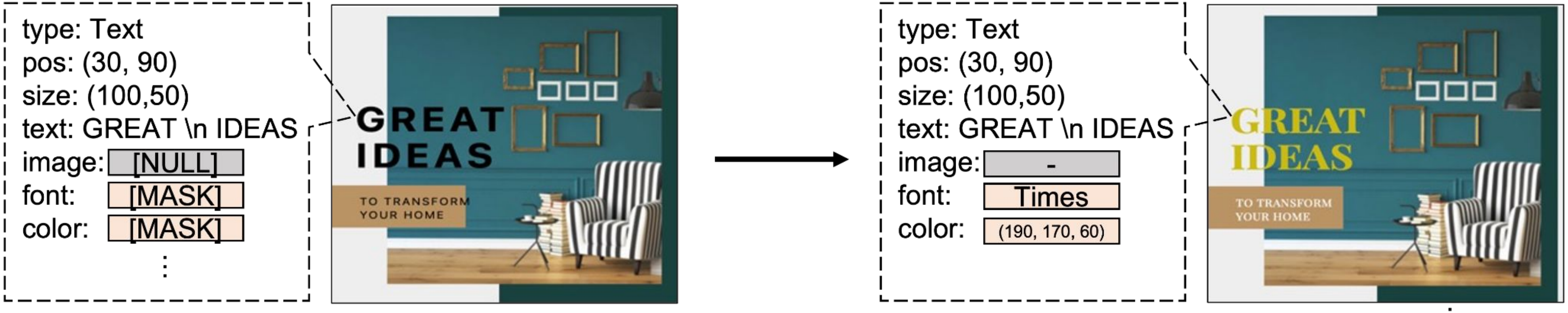
Flexible Document Model (FlexDM)

Our work: solve many design tasks in a single model



Key Idea of FlexDM

Multi-modal masked field prediction as a unified interface



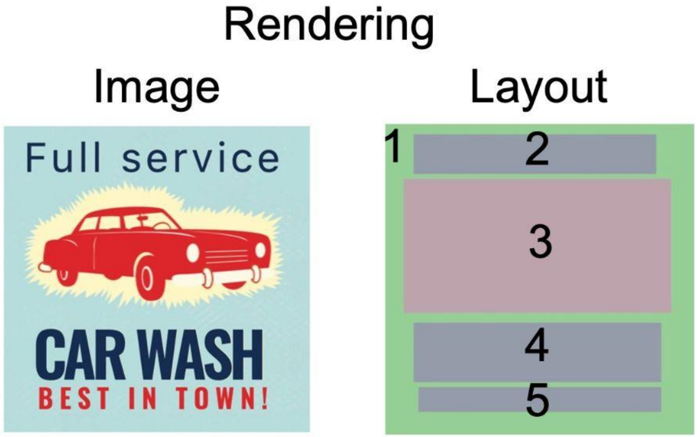
FlexDM Results

Input

Output

Vector Graphic Document

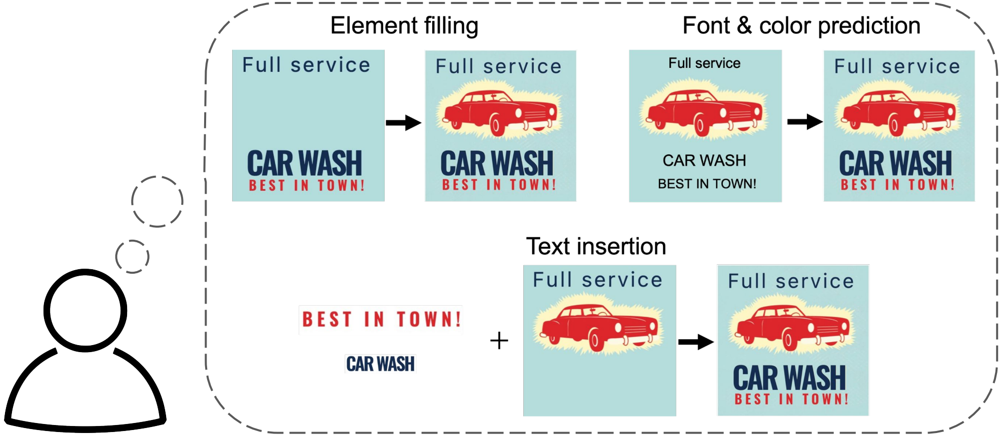
- A data format for making visual design (e.g., banner by Photoshop)
- Consists of a set of visual elements (+ global info) [Yamaguchi+, ICCV'21]
- Scalable, editable, human-interpretable



Vector graphic format

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

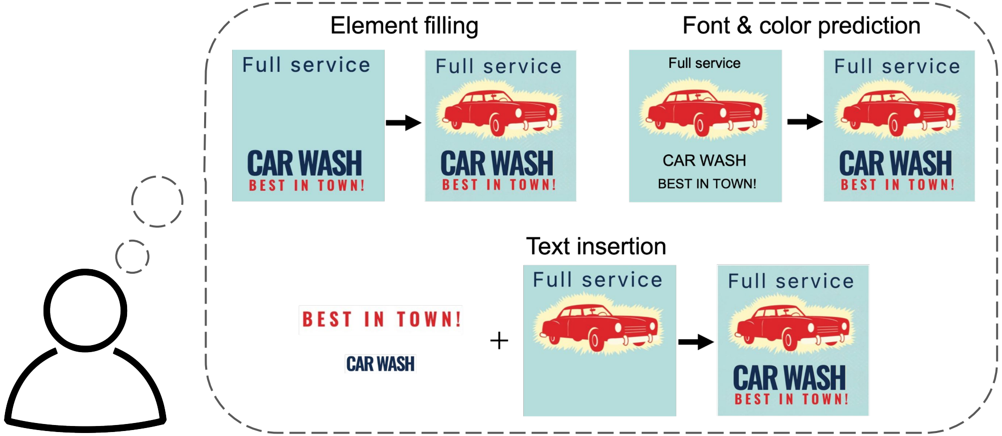
Design Tasks in Iterative Design Process



Design Tasks in Iterative Design Process

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

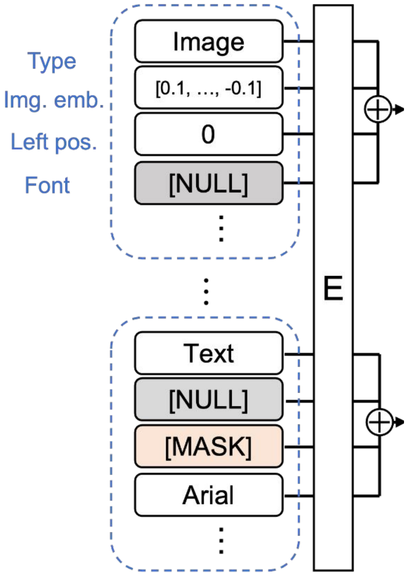
→ We handle design tasks in a principled manner



Masked Field prediction (MFP)

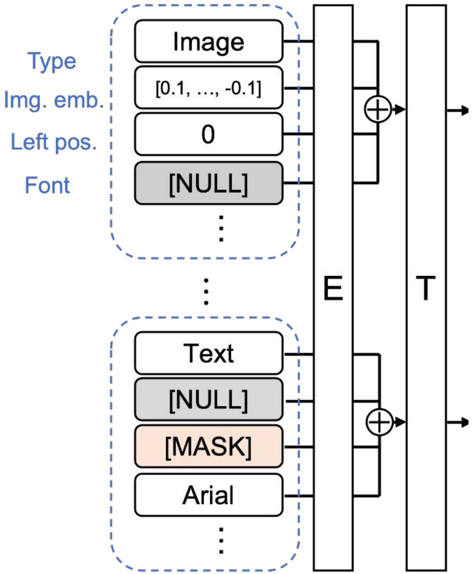
- **Predicting arbitrary number of fields hidden by [MASK]**
- **Challenges**
  - How to encode/decode various type of fields?
  - How to handle larger number of fields?

Network for Masked Field Prediction (MFP)



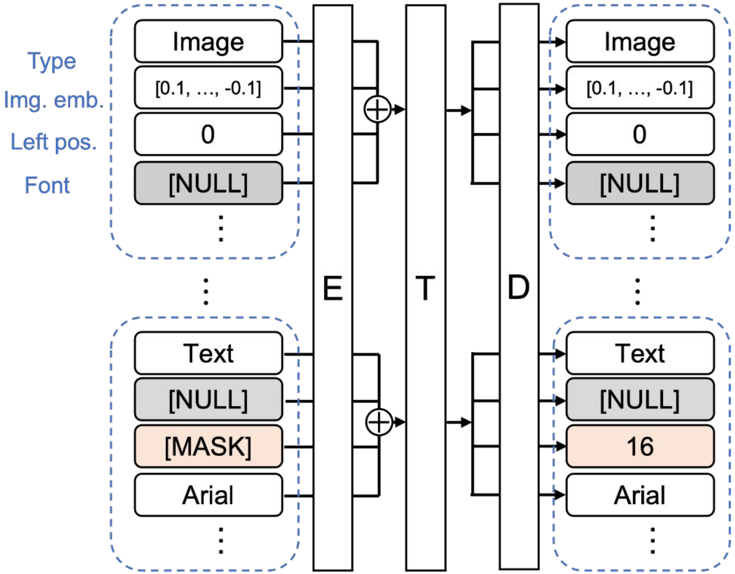
E: encoder

Network for Masked Field Prediction (MFP)



E: encoder, T: Transformer encoder

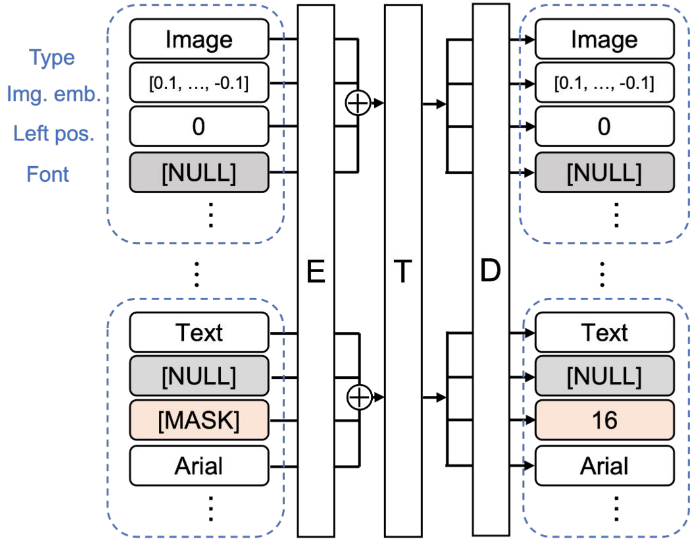
Network for Masked Field Prediction (MFP)



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



CyberAgent AI Lab

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/

Conclusion

Summary

- **Graphic design = multi-modal data**
- **Formulate many generation tasks using sequence-like data structure**
- **Many challenges remaining**
    - End-to-end generation including texts and images, or some alternative?
    - ChatGPT-like one-model-fits-all moment for design generation?

Acknowledgement

Kotaro Kikuchi
(CyberAgent)

Edgar Simo-serra
(Waseda University)

Mayu Otani
(CyberAgent)

Kota Yamaguchi
(CyberAgent)