PPT2Poster using Diffusion Models

Swapnil Bag
3rd Year Undergraduate
Computer Science and Engineering, Engineering Science
Indian Institute of Technology, Hyderabad
Email: es22btech11034@iith.ac.in

Advisor:
Prof. Vineeth N Balasubramanian

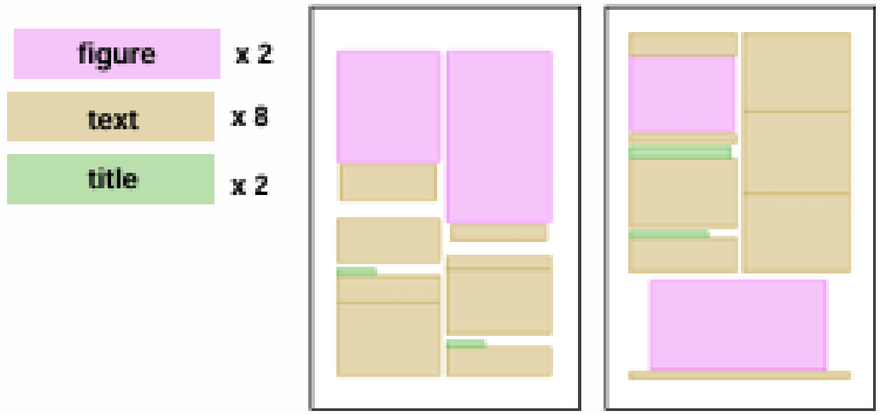
Motivation

- A tool to convert a PowerPoint to a poster has innumerable applications in real life. One such use case in research conferences is to condense your PPT to a research poster.
- Can be used in professional settings like converting business presentations to posters to portray your information in a concise manner
- There are similar tools online but require manual designing (Canva, PowerPoint)
- This is a research problem that has not been explored much directly and has a good application

Proposed Model

Content Summarisation : The PPT is parsed and the infographics like text images, graphs, tables, etc are extracted. This is fed to an LLM to summarise the content and present it in a predefined format consisting of (elements)

Layout Generation: Use Diffusion models (D3PM) techniques to predict a layout conditoned by the elements generated by the LLM.



Diffusion model Input

- The input can be represented as $x = \{(c_1, x_1, y_1, h_1, w_1), \ldots, (c_l, x_l, y_l, h_l, w_l)\}$
- C is the content type that can be images, text, equations, etc (categorical)
- The values of the bounding box can be quantized into integer bins or continuous range of $[0, 1]^4$

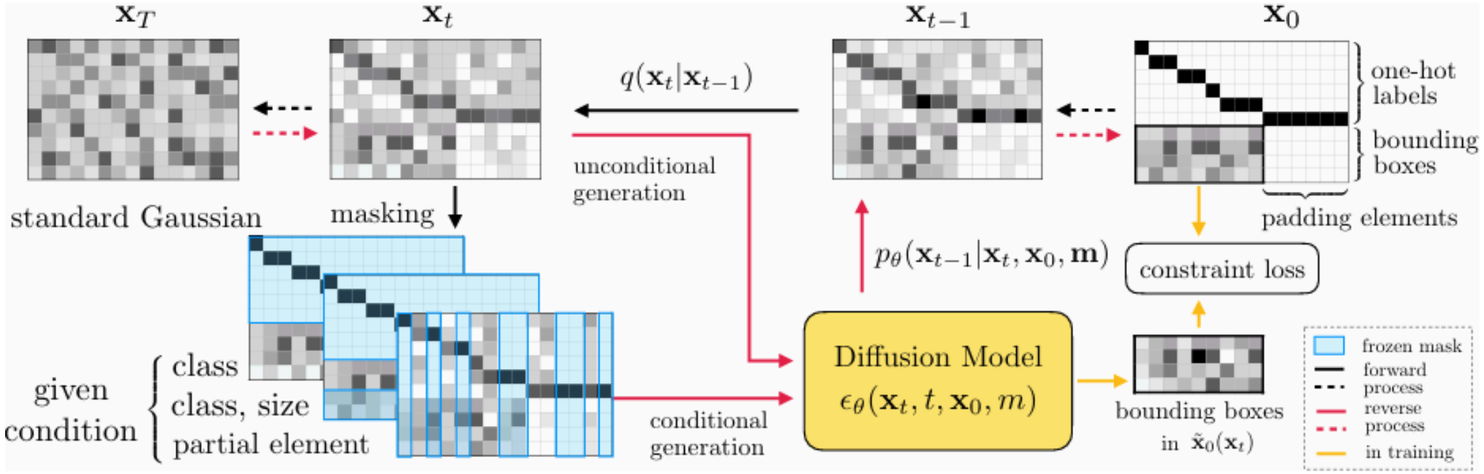
- These models are trained for unconditional generation of layout however they can be conditioned with Content+size and generate the position of the elements
- During the revese diffusion we sample from the augmented latent

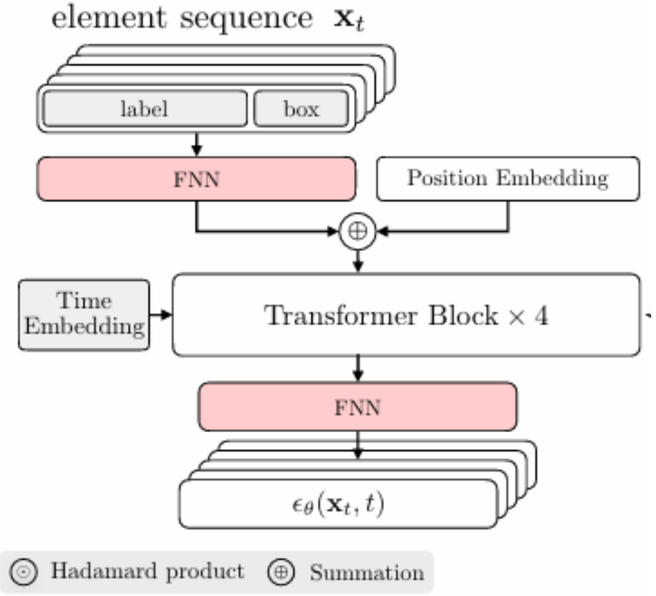
Layout Generation Through Diffusion

LACE

- Diffusion in the continuous space with aesthetic constraints (Most models follow D3PM architecture)
- Introduces alignment and overlap constraints in addition to MSE

$$\mathcal{L}_{\text{rec}} = \text{MSE}(\tilde{\mathbf{x}}_0, \mathbf{x}_0) + \omega_t \cdot (\mathcal{C}_{\text{alg}}(\tilde{\mathbf{x}}_0(\mathbf{x}_t), \mathbf{x}_0) + \mathcal{C}_{\text{olp}}(\tilde{\mathbf{x}}_0(\mathbf{x}_t))) ,$$





Chen et al. (2023). Towards aligned layout generation via diffusion model with aesthetic constraints.

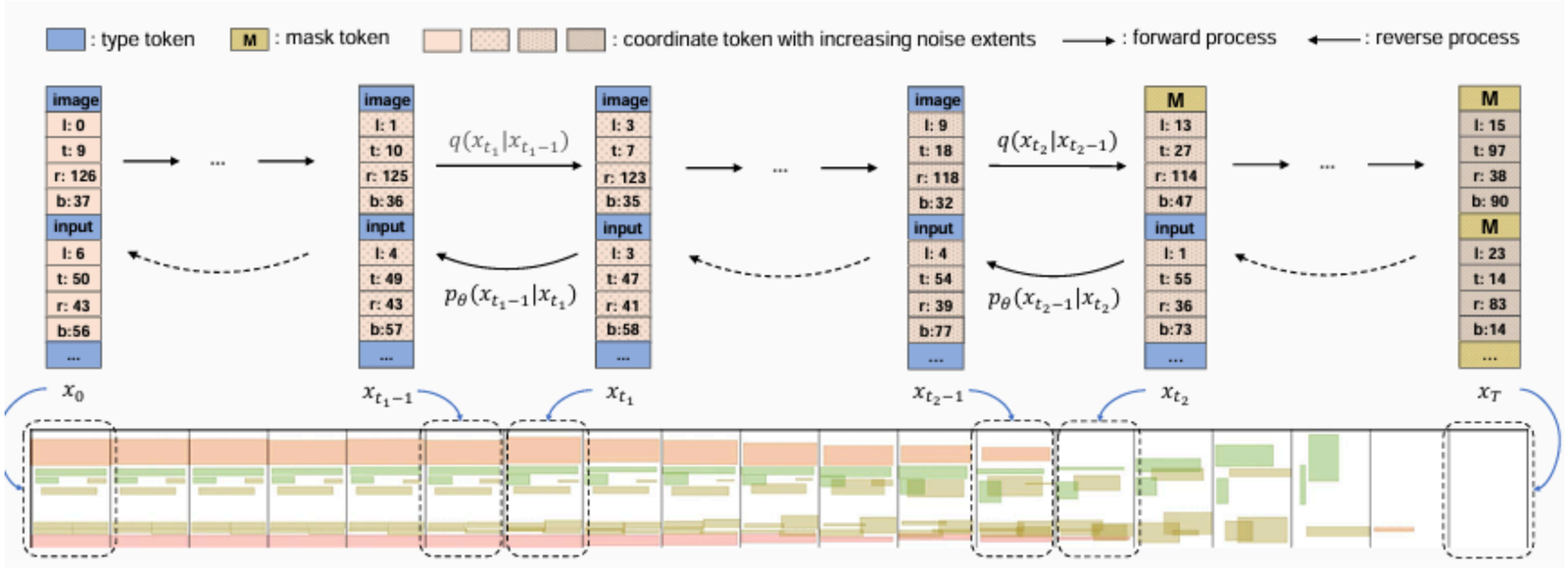
Layout Diffusion

Based on D3PM , has a quantized bin input

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}),$$

$$q(x_t|x_{t-1}) = x_t \mathbf{Q}_t x_{t-1}.$$

$$q(x_t|x_0) = \mathbf{x}_t^\top \overline{Q}_t \mathbf{x}_0$$



$$\mathbf{Q}_t = \begin{bmatrix} \mathbf{Q}_t^{\text{coord}} & & \\ & \mathbf{Q}_t^{\text{type}} & \\ & & \mathbf{Q}_t^{\text{spec}} \end{bmatrix}$$

$$\mathbf{Q}_t^{\text{type}} = \begin{bmatrix} 1-\gamma_t & 0 & \cdots & 0 \\ 0 & 1-\gamma_t & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_t & \gamma_t & \cdots & 1 \end{bmatrix}$$

$$[\mathbf{Q}_t^{\text{coord}}]_{ij} = \begin{cases} \dfrac{\exp\left(-\frac{4|i-j|^2}{(K-1)^2\beta_t}\right)}{\sum_{n=-(K-1)}^{(K-1)} \exp\left(-\frac{4n^2}{(K-1)^2\beta_t}\right)}, \\ 1 - \sum_{l=0, l\neq i}^{(K-1)} [\mathbf{Q}_t^{\text{coord}}]_{il}, \end{cases}$$

$$\mathcal{L} = \mathcal{L}_{\text{VLB}} - \lambda \log p_\theta(\mathbf{x}_0|\mathbf{x}_t).$$

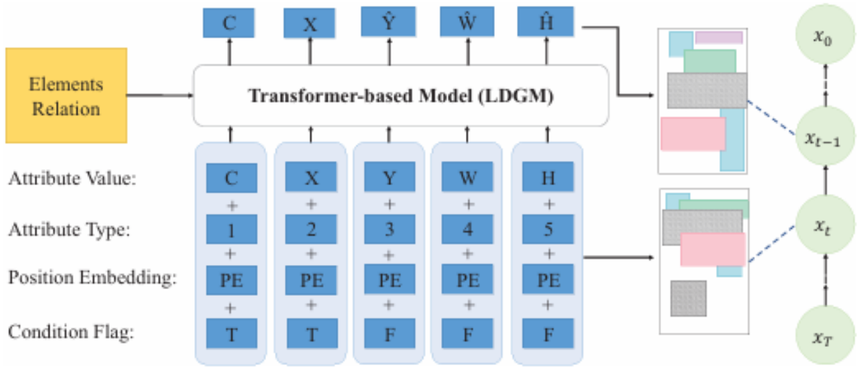
A transformer encoder is used tominimise the additional loss term in the reverse diffusion process

$$\{\text{emb}(x_{t,i}) + p_i + \text{emb}(t)\}_{i=1}^{M}$$

LDGM

- Similar to the idea of Layout diffusion, uses seperate matrices for content, (x,y) (h,w)
- It has a decoupled corruption strategy, uniform noicse is applied to content matrix and discrete gaussian noise to position matrices

$$\mathcal{L}_{rec} = -\log p_\theta(\hat{x}|x, \boldsymbol{g}(x))$$

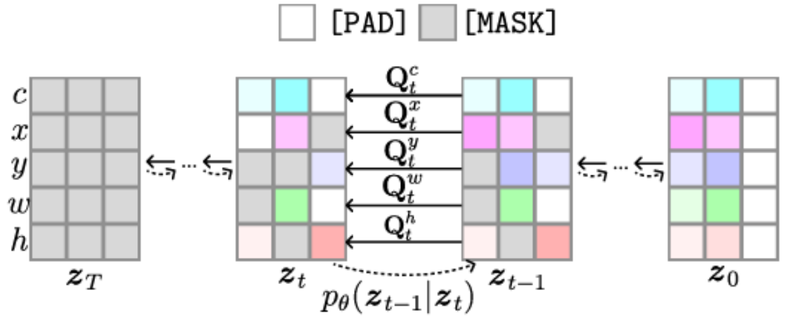


$$e_{i,j} = \frac{(\boldsymbol{x}_i W^Q + [V_r^Q]_{i,j})(\boldsymbol{x}_j W^K + [V_r^K]_{i,j})}{\sqrt{d}},$$

Layout DM

Uses 5 seperate transition matrices, corruption occurs modality wise

$$Q_t^c, Q_t^x, Q_t^y, Q_t^w, Q_t^h$$



$$z_{t-1} = m \odot z_{\text{known}} + (1 - m) \odot \hat{z}_{t-1}$$

$$\log \hat{p}_\theta(z_{t-1}|z_t) \propto \log p_\theta(z_{t-1}|z_t) + \lambda_\pi \pi$$

Hui et al. (2023). Unifying layout generation with a decoupled diffusion model.
Inoue et al. (2023). LayoutDM: Discrete diffusion model for controllable layout generation.

Content Summarisation

Method 1

- Extract the infographic elements like text images, equations, and tables from the PPT.
- Store the embeddings of the text extracted in a vector store. For images, equations, tables, and graphs, create a detailed summarization and store its embeddings.
- We can use a RAG pipeline to retrieve the information using a structured prompt to the LLM
- The prompt retrieves information for sections of the poster like title, introduction, methodology, results, and discussions/ conclusions (research type)
- These sections consists of elements like paragraphs, bullet points, heading, images , and graphs. This is basically the content type (C)

$$x = \{(c_1, x_1, y_1, h_1, w_1), \ldots, (c_l, x_l, y_l, h_l, w_l)\}$$

- We can fix the size of the canvas, and ask the LLM to predict the approximate size of each element. This (c,h,w) pair is passed as a conditional mask to the diffusion model and it predicts the center coordinates x,y

$$\hat{x}_t = m \circ x_0 + (1 - m) \circ x_t$$

How to ensure text and its related images are grouped together ?

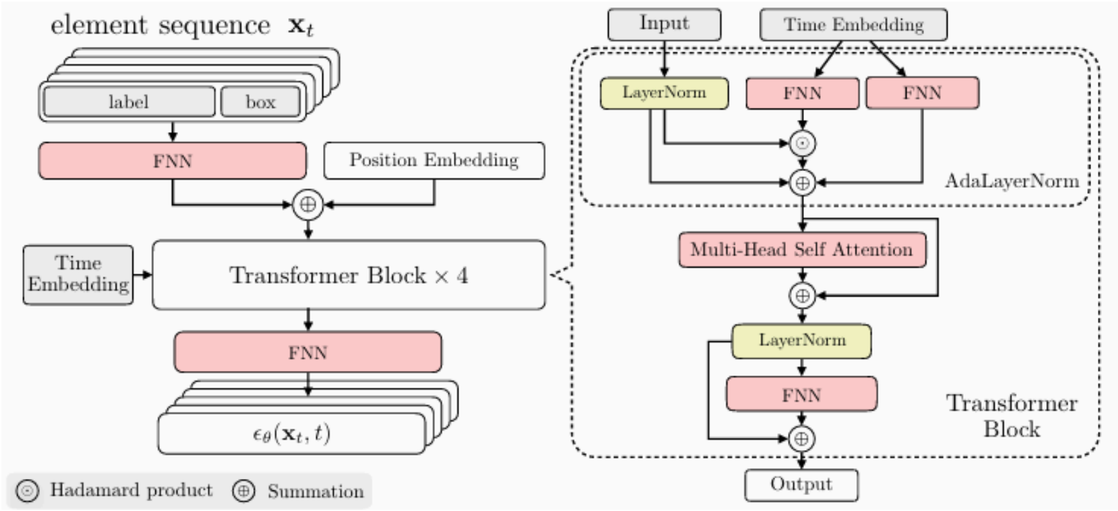
Method 2

- Each element that was decided to be displayed in the Poster has its corresponding embedding, which can be obtained from our vector store $\text{emb}(e_i)$
- Apart from the **aesthetic loss functions** introduced in **LACE**, add a new loss that constraints elements with similar content to appear together (be it image or text)

$$p_i = (x_i, y_i)$$

$$\mathcal{L}_{\text{semantic-proximity}} = \sum_{i \neq j} \|\mathbf{p}_i - \mathbf{p}_j\|_2^2 \cdot \exp\left(-\cos\left(\text{emb}(e_i), \text{emb}(e_j)\right)\right)$$

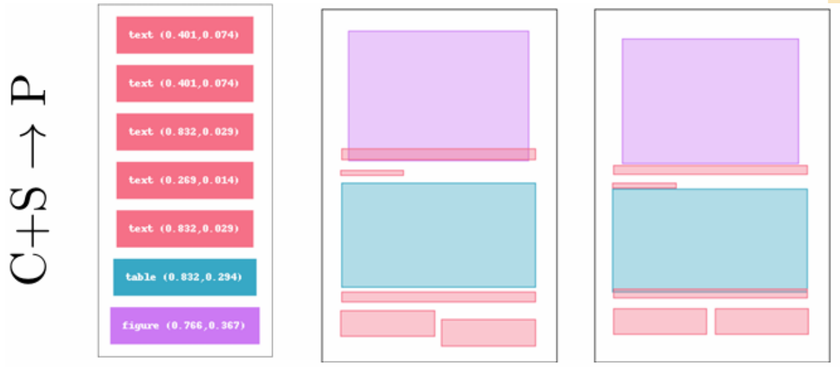
This method would require to retrain the entire LACE model from scratch and introduce the embeddings of elements.These semantic embeddings will be added to the normal input and will be used to compute the cross attention scores



Chen et al. (2023). Towards aligned layout generation via diffusion model with aesthetic constraints.

Layout generation

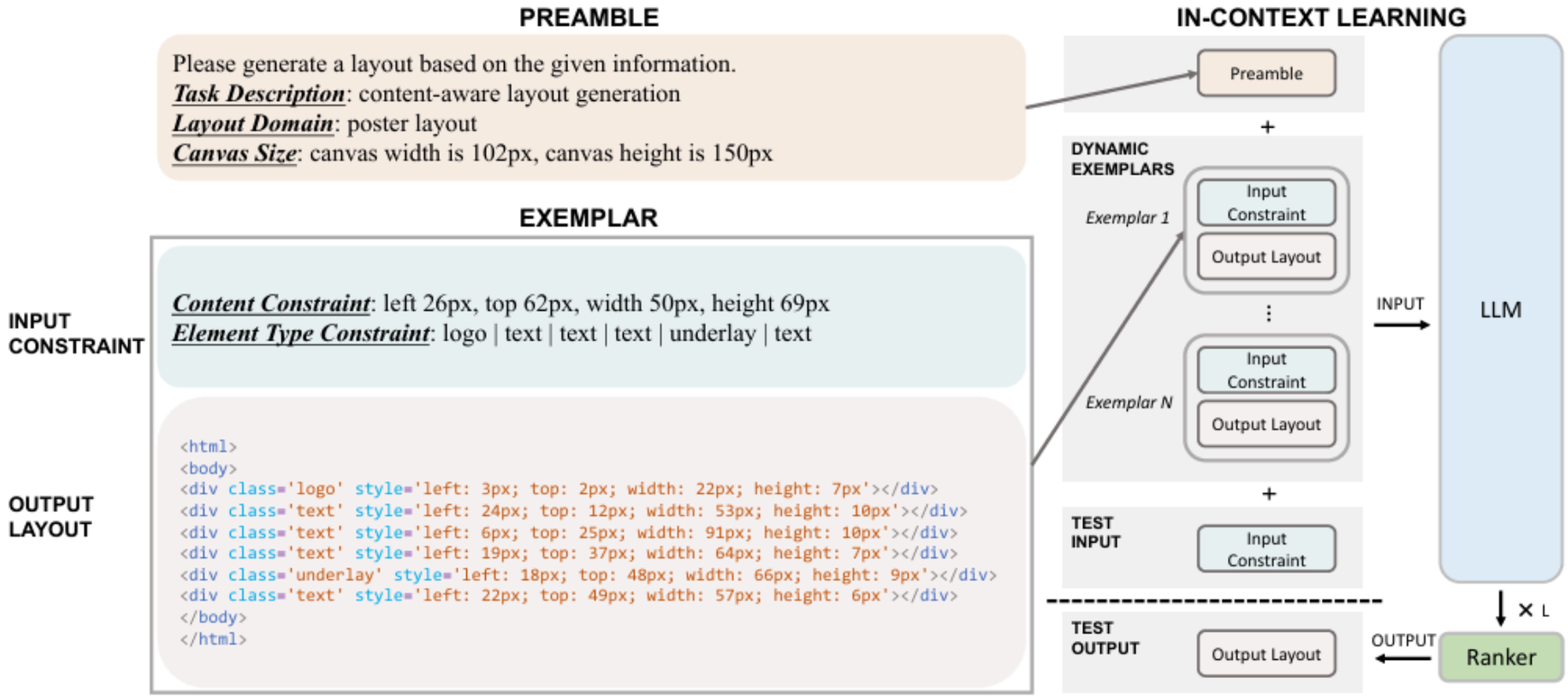
Displaying the final poster can done uses html tags. Each element is uniquely identified by a class name by the tag. The diffusion model pricts the centre cordinates, we can then use that to exactly to display on the poster




```
<div class='logo' style='left: 3px; top: 2px; width: 22px; height: 7px'></div>
```

Layout generation using Prompts

Layoutprompter



Lin et al. (2023). LayoutPrompter: Awaken the design ability of large language models.

References

- LACE
- Layout Diiffusion
- LDGM
- LayoutDM
- layoutPrompter