Two-stage Content-Aware Layout Generation for Poster Designs

Shang Chai
chaishang@mail.ustc.edu.cn
University of Science and Technology of China

Liansheng Zhuang*
lszhuang@ustc.edu.cn
University of Science and Technology of China

Fengying Yan
fengying@tju.edu.cn
Tianjin University

Zihan Zhou
zzhou@ist.psu.edu
Manycore Tech Inc.

ABSTRACT

Automatic layout generation models can generate numerous design layouts in a few seconds, which significantly reduces the amount of repetitive work for designers. However, most of these models consider the layout generation task as arranging layout elements with different attributes on a blank canvas, thus struggle to handle the case when an image is used as the layout background. Additionally, existing layout generation models often fail to incorporate explicit aesthetic principles such as alignment and non-overlap, and neglect implicit aesthetic principles which are hard to model. To address these issues, this paper proposes a two-stage content-aware layout generation framework for poster layout generation. Our framework consists of an aesthetics-conditioned layout generation module and a layout ranking module. The diffusion model based layout generation module utilizes an aesthetics-guided layout denoising process to sample layout proposals that meet explicit aesthetic constraints. The Auto-Encoder based layout ranking module then measures the distance between those proposals and real designs to determine the layout that best meets implicit aesthetic principles. Quantitative and qualitative experiments demonstrate that our method outperforms state-of-the-art content-aware layout generation models.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design process and methods**; • **Computing methodologies** → **Computer vision problems**.

KEYWORDS

layout generation, graphic design, conditional diffusion model

ACM Reference Format:
Shang Chai, Liansheng Zhuang, Fengying Yan, and Zihan Zhou. 2023. Two-stage Content-Aware Layout Generation for Poster Designs. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612275

*Corresponding Author

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612275

1 INTRODUCTION

Automatic layout generation has become an essential tool in modern design and engineering. It offers a faster, more efficient, and more accurate alternative, enabling designers to focus on the creative aspects of their work. Benefiting from the excellent generation performance of deep generative models, recent automatic layout generation models can generate numerous layouts in seconds. However, most current models [1, 3, 9, 15–17, 20–22, 27, 32] define the layout generation problem as arranging elements with different attributes on a blank canvas. Although this setting is suitable for some traditional design situations, such as document typesetting or web page layout, it is not ideal for the layout design of advertising posters or magazine covers, where an image is used as background. Content-aware layout generation, also known as visual-textual presentation [33] layout generation, considers the task of placing texts and embellishments on a background image. Following [36], we generally call this kind of task poster layout generation. It is challenging to generate high-quality poster layouts, since we need to consider not only the relationships between layout elements, but also the relationships between layout elements and the background image.

In recent years, various methods have been proposed to address the challenges posed by poster layout generation. Template-based methods [14, 33] directly blend the text templates and the background image, which often fail to generate flexible and various layout results due to the lack of highly professional design templates. Early template-free poster layout generation methods [34] first use a layout proposal algorithm to generate candidate text regions, then use a deep scoring network to assess the aesthetic quality of the candidate results. Due to the heavy computation, these methods are only suitable for generating simple poster layouts. With the development of deep learning, using deep generative models such as VAEs [19] and GANs [7] to generate poster layouts has gained more research interests [2, 8, 35, 36]. By learning the distribution of real layouts, these models can generate decent poster layouts without the need for templates or manually-designed rules. However, current deep generative models still have obvious shortcomings or room for improvement. First, most GAN-based methods [2, 35, 36] directly use the visual features of background images as conditional inputs to control the generation process. However, studies [4, 13] have found that mode collapse [4] easily appears in these strongly conditional generative models, which results in similar or degraded samples. Second, these methods do not exploit existing layout aesthetic principles to control the generation process, which results in the inability to guarantee the quality of the generated layouts. Therefore, our research goal is to explore a new method that can

fully utilize the generative power of deep generative models while considering both explicit and implicit aesthetic principles.

Considering the superiority of the diffusion models in generation quality and diversity, we choose LayoutDM [3] as our base layout generation model. However, two major challenges need to be addressed before using it to generate poster layouts that meet aesthetic principles. First, the existing diffusion models [6, 10, 11, 23, 28] lack a mechanism that can utilize an explicit objective function, such as overlap ratio and alignment metrics, to control the generation process. Second, existing models neglect implicit aesthetic principles such as canvas space allocation and page balance. How to effectively model them remains an open problem.

In this paper, we propose a two-stage poster layout generation framework that considers both explicit and implicit aesthetic principles. Given an input background image and custom aesthetic constraints, our framework tends to generate high-quality poster layouts that satisfy both explicit and implicit design aesthetics. The first stage aims to generate high-quality layout proposals with explicit aesthetic constraints through a layout generation module based on LayoutDM [3]. Specifically, we incorporate explicit aesthetic constraints into the diffusion reverse process by utilizing optimization in the latent space during our iterative layout denoising process. Generated layout proposals are then used as inputs for the second stage. The second stage aims to select layouts that meet implicit aesthetic principles (such as better canvas space allocation and page balance) through a layout ranking module. To this end, we first extract the salient object from the background image using a salient object detection (SOD) [31] network, and combine it with the layout proposals generated in the first stage to construct novel composite layouts. Then, an Auto-Encoder based layout ranking module pretrained on real poster designs is used to rank these composite layouts with the reconstruction loss. The key insight here is that the Auto-Encoder can efficiently reconstruct human-designed poster layouts which accurately meet layout aesthetic principles. Therefore, if the composite layout is far from the distribution of the real poster designs, the reconstruction loss will be relatively large. For simplicity, we name the proposed layout generation framework as CA-LayoutDM. Extensive experiments show that our method outperforms state-of-the-art poster layout generation methods.

In summary, our main contributions are as follows:

- We propose a novel two-stage content-aware layout generation framework to generate poster layouts. Compared to existing models, our method generates layouts that adhere to aesthetic principles while maintaining quality and diversity.
- We propose a novel aesthetics-guided layout denoising process conditioned on explicit aesthetic principles to generate layout proposals, and an Auto-Encoder based layout ranking module to select the layout that best satisfies overall layout aesthetic principles.
- Extensive experiments demonstrate that our method outperforms state-of-the-art models in terms of visual quality on both content-agnostic and content-aware metrics.

2 RELATED WORK

Content-agnostic layout generation has been a long-standing research topic, involving the arrangement of graphic layout elements on a blank canvas. Early methods embed design rules into handcrafted energy functions [24, 25], but often fail to generate complex and diverse layout results. LayoutVAE [16] and LayoutGAN [22] are the first to introduce deep generative networks to layout generation, facilitating data-driven approaches to layout generation tasks. Subsequent works [1, 9, 15, 20, 21, 27, 32] improve the quality of generated layouts by developing different models based on generative networks such as VAEs [19] and GANs [7]. LayoutDM [3] is a recently proposed layout generation model that leverages the generative performance of diffusion models, and has improved quality and diversity on content-agnostic layout generation. As the development process continues, a research trend is to impose more constraints on the models to get desired results. NDN [21], for instance, represents the relative positional relationships of layout elements as a complete graph and uses graph convolutional neural networks to generate graphic layouts that satisfy these relationships. Similarly, CLG-LO [17] designs user-specific constraints based on aesthetics and element relationships, improving the degree of control over the results through constrained optimization in the latent space. Finally, BLT [20] represents the layout as a discrete sequence, generating layouts conditioned on the categories and size of layout elements via iterative decoding operations.

Content-aware layout generation, on the other hand, requires considering visual information in the background image when placing texts, logos, and embellishments on it. Early methods [14, 33] rely on templates and heuristic rules designed by designers, which can quickly synthesize layouts but lack diversity and versatility. To overcome these shortcomings, template-free methods are developed. SmartText [34] generates text-anchor proposals for text bounding boxes and then ranks them using a binary classifier based scoring network. Vinci [8] uses a VAE to learn the multimodal distribution of product images and corresponding design sequences, then samples from it. ContentGAN [35] and ICVT [2] extract visual features from images using a visual backbone, and CGL [36] further considers the salient regions of the background. Essentially, these three models are directly conditioned on the visual features. ContentGAN relies on concatenating-based conditioning, while ICVT and CGL utilize Cross-Attention based conditioning. Although there are some improvements in terms of content-aware metrics, such as saliency overlap, the coarse-grained and strongly conditional nature of these approaches still leads to a noticeable loss of diversity and quality in the generated results.

3 OUR METHOD

3.1 Problem Formulation

A poster layout $\mathbf{l}$ consists of several layout elements with geometric parameters and attributes. It can be depicted as a variable-size set:

$$\mathbf{l} = \{(g_1, f_1), (g_2, f_2), ..., (g_i, f_i), ..., (g_N, f_N)\}$$

where N is the number of elements in the layout, $g_i = [x_i, y_i, w_i, h_i]$ is a vector representing the geometric parameters (center coordinates and size) of i-th element in the layout, f_i is the attributes of i-th element, which might be category labels or other features. We use sequences $\mathbf{g} = (g_1, g_2, ..., g_N)$ and $\mathbf{f} = (f_1, f_2, ..., f_N)$ to represent the geometric parameters and attributes of all elements in a layout respectively. Given a background image $\mathbf{bg}$ and a layout attributes

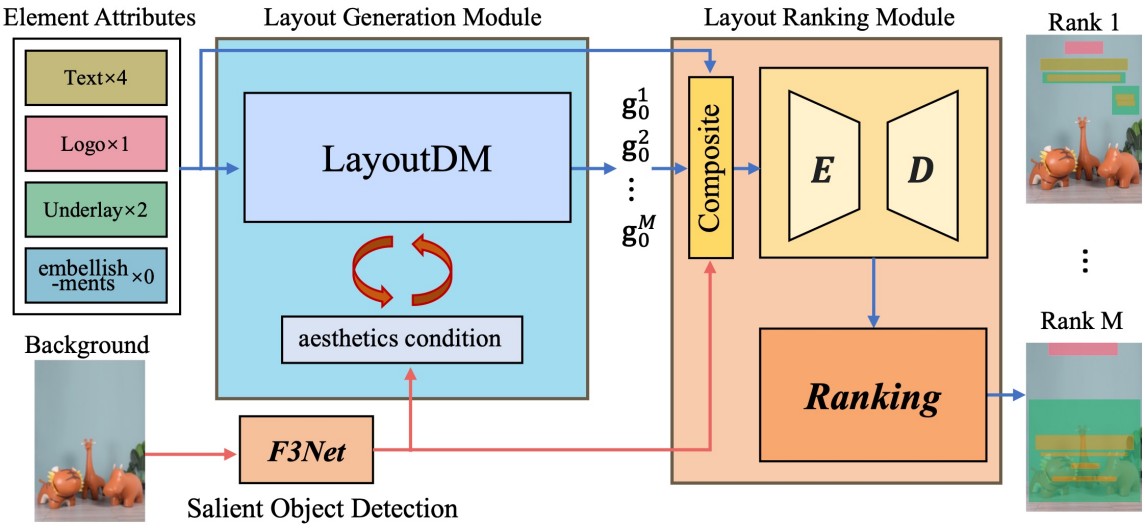


Figure 1: The architecture of our content-aware poster layout generation framework. It consists of a layout generation module and a layout ranking module. Layout generation module takes the network architecture of LayoutDM [3].

sequence $\mathbf{f}$ as input, our goal is to generate decent element geometry parameters $\mathbf{g}$ to construct a high-quality and visually-pleasing poster layout.

3.2 Architecture Overview

Figure 1 illustrates the architecture of our content-aware layout generation framework. It consists of two primary components: a layout generation module and a layout ranking module. The DDPM-based [12] layout generation module has the same architecture as LayoutDM [3]. It generates layout proposals conditioned on user-specific aesthetic constraints and the provided element attributes. The layout ranking module is a transformer-based Auto-Encoder that assesses the generated layout proposals considering both the layout elements and the input background image.

Formally, denote the layout generation module, user-specific aesthetic constraints, and element attributes in the layout as PG, $const$ and $\mathbf{f}$ respectively. PG generates M layout proposals conditioned on $const$ and $\mathbf{f}$:

$$\{\mathbf{g}_0^1, \mathbf{g}_0^2, \cdots, \mathbf{g}_0^i, \cdots, \mathbf{g}_0^M\} = PG(\{\mathbf{g}_T^i\}_{i=1}^M, \mathbf{f}, const) \tag{1}$$

where M is the number of the generated proposals, $\mathbf{g}_T^i$ is the random Gaussian noise used to generate i-th layout proposal, and $\mathbf{g}_0^i$ is the generated geometric parameters of all the layout elements in i-th proposal. Then, we use the layout ranking module to rank the proposals:

$$\{(\overline{\mathbf{g}}_0^j, \mathbf{f}, \mathbf{sal})\}_{j=1}^M = LRanker(\{(\mathbf{g}_0^i, \mathbf{f}, \mathbf{sal})\}_{i=1}^M) \tag{2}$$

where $LRanker$ is our layout ranking module, $\mathbf{sal}$ is the saliency map of the input background image. We have the following layout rank order after layout ranking:

$$rank(\overline{\mathbf{g}}_0^1, \mathbf{f}, \mathbf{sal}) < rank(\overline{\mathbf{g}}_0^2, \mathbf{f}, \mathbf{sal}) \cdots < rank(\overline{\mathbf{g}}_0^M, \mathbf{f}, \mathbf{sal}) \tag{3}$$

where $\overline{\mathbf{g}}_0^j \in \{\mathbf{g}_0^i\}_{i=1}^M$, $j = 1, 2, \cdots, M$, we ultimately select the layout proposal with the highest rank, i.e., $(\overline{\mathbf{g}}_0^1, \mathbf{f})$ as our final result.

3.3 Salient Object Detection

We use a pretrained salient object detection network *F3Net* [31] to detect the position of the salient element in the background image. *F3Net* takes the background image $\mathbf{bg}$ as input and output a gray image $\mathbf{sal}$ whose pixel values represent the degree of saliency.

3.4 Layout Generation

We train our layout generation module following the instruction in layoutDM and use the DDIM [30] sampler to further speed up the sampling process, reducing the original 1000-step iteration to 50 steps. This leads to the following iterative denoising process:

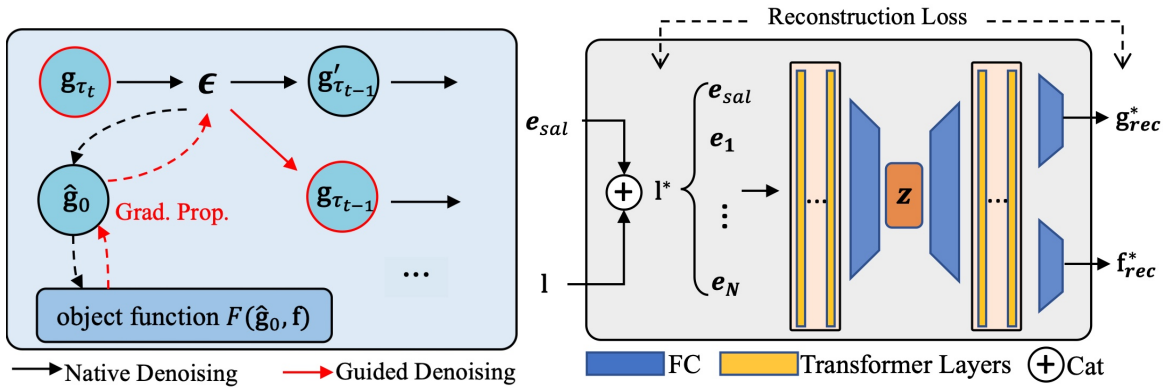
$$\mathbf{g}_{\tau_{t-1}} = \sqrt{\alpha_{\tau_{t-1}}}\left(\frac{\mathbf{g}_{\tau_t} - \sqrt{1-\alpha_{\tau_t}}\epsilon_{\tau_t}^\theta(\mathbf{g}_{\tau_t}, \mathbf{f})}{\sqrt{\alpha_{\tau_t}}}\right) + \sqrt{1-\alpha_{\tau_{t-1}}}\epsilon_{\tau_t}^\theta(\mathbf{g}_{\tau_t}, \mathbf{f}) \tag{4}$$

where τ is a subsequence of $[1, 2, \cdots, T]$ of length S, $\alpha_{1:T} \in (0, 1]$ is a decreasing sequence to parameterize the Gaussian transitions described in [30], and $\epsilon_{\tau_t}^\theta(\mathbf{g}_{\tau_t}, \mathbf{f})$ is a conditional noise predictor modeled by a neural network.

Unlike the native reverse diffusion process, we guide the denoising process with explicit aesthetic constraints to further improve the quality of the sampled layout proposals. To this end, we first define an objective function F to model the user-specific aesthetic constraints. F should be a scalar function differentiable with respect to $\mathbf{g}$, which consists of a series of differentiable constraint terms:

$$F(\mathbf{g}, \mathbf{f}) = w_a R_{align}(\mathbf{g}, \mathbf{f}) - w_u R_{und}(\mathbf{g}, \mathbf{f}) + w_o R_{ovrlp}(\mathbf{g}, \mathbf{f}) + w_s R_{sal}(\mathbf{g}, \mathbf{f}) \tag{5}$$

where R_{align}, R_{und} and R_{ovrlp} are the functions that compute the Alignment, Underlay-overlap and Overlap metrics. R_{sal} is a function that computes the overlap between layout elements and the background salient element. We describe the methods to compute these terms in Sec. 4.1.2. w_a, w_u, w_o and w_s are the corresponding weights. These weights are hyperparameters fixed to 1, 0.6, 0.1 and 0.3, respectively in all the experiments.



(a) Aesthetics-guided layout denoising

(b) Architecture of the layout ranking module

Figure 2: a) Illustration of the aesthetics-guided layout denoising. We perform optimization in the ϵ-space at each denoising step to make the result satisfy explicit aesthetic constraints. b) Architecture of our Auto-Encoder based layout ranking module.

Then, we take F as the objective function and perform gradient descent in the ϵ-space at each denoising step, so as to generate samples that ultimately minimize F. In particular, we have:

$$\hat{\mathbf{g}}_0 = \frac{\mathbf{g}_{\tau_t} - \sqrt{1-\alpha_{\tau_t}}\boldsymbol{\epsilon}}{\sqrt{\alpha_{\tau_t}}} \tag{6}$$

$$\hat{F}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^0, \mathbf{f}) = F(\hat{\mathbf{g}}_0, \mathbf{f}) + \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^0\|^2 \tag{7}$$

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon} - \nabla_{\boldsymbol{\epsilon}} \hat{F}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}^0, \mathbf{f}) \tag{8}$$

where F is the differentiable objective function we defined, $\boldsymbol{\epsilon}^0$ is the output of noise predictor at time step τ_t, and $\boldsymbol{\epsilon}$ are initialized with $\boldsymbol{\epsilon}^0$. We take several gradient descent steps at each denoising step, and continue the denoising process using formula 4 with the modified $\hat{\boldsymbol{\epsilon}}$. Intuitively, we estimate $\mathbf{g}_0$ using formula 6 at each denoising step t, and then perform gradient descent in the $\boldsymbol{\epsilon}$-space, to search for a $\hat{\boldsymbol{\epsilon}}$ with a smaller estimated $\hat{F}$ value. Regularization term $\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^0\|^2$ is used to keep the sampling process following the original diffusion flow, preventing degrading the layout quality.

We iteratively perform S aesthetics-guided denoising steps to generate a $\mathbf{g}_0$, and repeat this process M times to obtain M layout proposals. These proposals serve as the input for our layout ranking module. We illustrate the aesthetics-guided layout denoising process described above in Figure 2a and Algorithm 1.

Algorithm 1: Aesthetics-guided layout proposal sampling

Require: pretrained layoutDM with a layout denoiser $\boldsymbol{\epsilon}_t^\theta$
Input: Objective function $\hat{F}$ regarding user-specific aesthetic constraints, layout attributes $\mathbf{f}$, proposal numbers M, gradient step number $Gradstep$
Output: Layout proposals set $\mathbf{P}$ with M proposals

```
P = ∅;
for i = 1, 2, ··· , M do
    g_T ~ N(0, I);
    for t = S, ··· , 1 do
        ε = ε^0 = ε^θ_{τ_t}(g_{τ_t}, f);
        for r = 1, 2, ··· , Gradstep do
            ĝ_0 = (g_{τ_t} − √(1−α_{τ_t}) ε) / √α_{τ_t};
            Take gradient descent step:
                ε = ε − ∇_ε F̂(ε, ε^0, f)
        end
        g_{τ_{t−1}} = √α_{τ_{t−1}} ( (g_{τ_t} − √(1−α_{τ_t}) ε) / √α_{τ_t} ) + √(1 − α_{τ_{t−1}}) ε;
    end
    P = P ∪ (g_0, f);
end
return P
```

3.5 Layout Ranking

The layout ranking module ranks the layout proposals generated by the layout generation module by measuring the distribution distance between generated layouts and high-quality human designs. We first take a threshold operation to get the binarized saliency mask **salmask** from the saliency map **sal**.

$$\mathbf{salmask}_{ij} = \begin{cases} 1 & \mathbf{sal}_{ij} > thred \\ 0 & \mathbf{sal}_{ij} \le thred \end{cases} \tag{9}$$

we use the minimum bounding rectangular to bound the saliency mask, and use the geometric parameters of the bounding box and a new category to define a special type of layout element $\mathbf{e}_{sal} = (g_{sal}, f_{sal})$. We add the salient element $\mathbf{e}_{sal}$ into the original layout proposal $\mathbf{l}$ to construct a novel composite layout $\mathbf{l}^* = (\mathbf{g}^*, \mathbf{f}^*)$, which not only contains element information, but also contains salient object information in the background image.

$$\mathbf{l}^* = \{(g_{sal}, f_{sal}), (g_1, f_1), (g_2, f_2), ..., (g_i, f_i), ..., (g_N, f_N)\} \tag{10}$$

Following the above composite layout representation, we build a novel composite layout dataset based on the real poster dataset. After creating the composite layout dataset, we train a transformer-based Auto-Encoder on it. The Auto-Encoder is optimized to learn a compressed representation of the distribution of composite layouts, including both background salient object and layout elements. The architecture of the layout ranking module is illustrated in Figure 2b.

Our layout ranking module takes inspiration from autoencoder-based out-of-distribution detection (OOD) methods [5, 29, 37]. It is

Table 1: Comparison with content-agnostic methods. The values of Balance are multiplied by 1000x for visibility. "Occ." means the ratio of non-empty layouts generated by models.

Models	Occ.↑	Overlap↓	Underlay Ovlp↑	Alignment↓	FID↓	Saliance Ovlp↓	Balance↓
LayoutTranformer [9]	100	0.0156	0.9516	0.0049	**5.34**	0.175	73.8
VTN [1]	99.9	0.0130	0.9698	0.0047	10.3	0.190	70.8
LayoutGAN++ [17]	100	0.0404	0.9859	**0.0015**	25.91	0.159	73.7
LayoutDM [3]	100	0.0176	0.9728	0.0033	5.6	0.158	74.6
Ours	100	**0.0102**	**0.9922**	0.0027	13.62	**0.092**	**58.5**
Real Data	100	0.0004	0.9946	0.0035	0.68	0.026	59.6

Table 2: Comparison with content-aware methods.

Models	Occ.↑	Overlap↓	Underlay Ovlp↑	Alignment↓	FID↓	Saliance Ovlp↓	Balance↓
ContentGAN [35]	93.4	0.0397	0.8626	0.0071	26.64	0.165	71.4
CGL [36]	99.7	0.0256	0.9413	0.0098	36.24	**0.077**	62.8
Ours	100	**0.0102**	**0.9922**	**0.0027**	**13.62**	0.092	**58.5**
Real Data	100	0.0004	0.9946	0.0035	0.68	0.026	59.6

trained to reconstruct in-distribution samples (high-quality human-designed poster layouts), which allows it to effectively reconstruct inlier data samples while distorting out-of-distribution samples that violate implicit aesthetic principles for poster layouts. Thus, reconstruction loss is an appropriate metric for measuring the distribution distance between generated layouts and high-quality human-designs. We rank the layout proposals, with layouts having lower reconstruction loss receiving a higher rank.

3.6 Training and Inference

We train our layout generation module following the instruction in LayoutDM [3] and optimize random term L_t, which are the KL divergences between $p_\theta(\mathbf{g}_{t-1}|\mathbf{g}_t, \mathbf{f})$ and forward process posteriors. We train our layout ranking module using the reconstruction loss on both continuous domains (geometric parameters) and discrete domains (element category attributes). The loss function is constructed as follows:

$$L_{rec}(\Theta) = \|\mathbf{g}^* - \mathbf{g}^*_{rec}\|^2 + \lambda_{ce} \cdot CrossEntropy(\mathbf{f}^*, \mathbf{f}^*_{rec}) \quad (11)$$

where Θ is the trainable parameters of our layout ranking module, $\mathbf{g}^*$ and $\mathbf{f}^*$ are the geometric parameters and attributes of all the elements in the composite layout $\mathbf{l}^*$ defined in Sec. 3.5, $\mathbf{g}^*_{rec}$ and $\mathbf{f}^*_{rec}$ are the reconstructed values. λ_{ce} is the hyper-parameter to weigh the cross-entropy loss term.

4 EXPERIMENTS

4.1 Experimental Settings

4.1.1 Dataset. **CGL** [36] dataset collects posters from e-commerce platforms. It comprises 60K advertising posters collected from various product categories such as cosmetics, electronics and clothing. The layout elements are manually classified into four categories (logos, texts, underlays and embellishments) and annotated with a bounding box to present their position. We use 51,818 poster-layout pairs for training, 8,730 pairs for validation and 1,000 pure product images for testing. Note here that other public datasets like **Crello** [32] and **Magazine** [35] only contain visual information for image-type elements, which is different from our setting where an image is used as the background. Thus, all of our experiments are performed on the **CGL** dataset.

4.1.2 Metrics. We employ four content-agnostic metrics (Overlap, Underlay Overlap, Alignment, FID) and two content-aware metrics (Saliency Overlap, Balance) to evaluate the quality of the generated layouts. **Overlap** and **Alignment** [36] are widely used metrics to evaluate the quality of the layout. In poster layouts, the underlay elements often overlap with text elements, and embellishment elements need not align with other elements. Therefore, we omit related computations as in CGL [36]. **Underlay Overlap** [36] computes the maximum overlap ratio of an underlay element with every element from other categories. **FID** [18] measures the distribution distance between two collections of layouts. We follow the method in [18] and train a classifier to compute FID. **Saliency Overlap** calculates the intersection area of all elements within the layout and the saliency map of the background image. We normalize the value with the area of the union of all elements. **Balance** measures the compositional balance of a composite layout. We calculate the average *L2* distance between the center of a canvas and the center of gravity of the layout elements.

4.1.3 Implement details. We use Pytorch [26] to implement our models. We employ a 50-step DDIM sampler when generating proposals. During the aesthetics-guided layout denoising process, we use an Adam optimizer with a learning rate of 0.01 and take five gradient descent steps at each time step. The number of layout proposals M is set to 32 in all the experiments. Our layout ranking module contains a symmetric encoder and decoder, each having 8 Transformer layers with an *atten dim* of 512. λ_{ce} is set to 1×10^{-3}.

4.2 Quantitative Evaluation

To quantitatively compare the generation performance of our method with state-of-the-art layout generation models, we implement ContentGAN [35], VTN [1], LayoutTranformer [9] and CGL [36] based on official codes and technical details in their papers. Moreover,

Figure 3: Qualitative comparison results. "Trans." denotes "LayoutTransformer" and "GANpp." denotes "LayoutGAN++".

we reimplement two additional layout generation models, LayoutGAN++ [17] and LayoutDM [3], as our baseline models. Note here that LayoutDM, LayoutGAN++ and our framework are conditional and are designed to have element attributes in a layout as additional conditional input. At test time, We align these three models with other unconditional models by randomly sampling element attributes from the layouts in the validation set.

4.2.1 Comparison with content-agnostic methods. The quantitative comparisons with content-agnostic methods are reported in Table 1.

Figure 4: Diversity performance. We show four samples given the same layout element attributes as conditional input. Layouts generated by LayoutDM are shown for reference.

From this table, we can observe: 1) Our method outperforms existing content-agnostic models on Overlap and Underlay Overlap, and achieves plausible results on Alignment. This demonstrates that our method tends to generate layouts that satisfy explicit aesthetic constraints better, benefiting from the aesthetics-guided layout denoising in the layout generation module. 2) Our method outperforms content-agnostic methods by a large margin on Saliency Overlap and Balance metrics, this validates the effectiveness of our method in modeling aesthetic rules such as saliency non-overlap and page balance accurately. 3) Our method performs slightly weaker on the FID metric, particularly when compared to LayoutDM, which is our layout generation module. This is because our method generates layouts that satisfy explicit and implicit aesthetic constraints through optimization and ranking, which likely results in less distribution coverage than content-agnostic methods.

4.2.2 Comparison with content-aware methods. The quantitative comparisons with content-aware methods are reported in Table 2. This table shows the following: 1) Our method significantly outperforms the other two methods on content-agnostic metrics. This proves that our method generates higher quality layouts, thanks to its robust generation ability, aesthetics-guided layout sampling, and layout ranking. 2) Our method achieves comparable results on Balance and slightly inferior results on Saliency Overlap compared to the SOTA model CGL. This is likely because we only consider the saliency information of the background image and do not fully leverage the visual information and semantic information contained in an RGB image using a visual backbone, as CGL does.

4.3 Qualitative Results

4.3.1 Generation quality comparison. The quantitative comparison results are shown in Figure 3. One can see that: 1) Compared with content-agnostic methods (first four rows), content-aware methods (last three rows) generate more harmonious and coherent layout results that blend seamlessly with the underlying background images. For instance, text elements are positioned in a way that does not occlude the salient object in the background, such as people or products, thus enhancing the overall aesthetic appeal of the generated layouts. This is because content-aware methods can capture the relationships between layout elements and background images. 2) Among the content-aware methods, our proposed approach demonstrated superior performance in generating high-quality layouts, especially in terms of alignment, when compared to the state-of-the-art CGL method. Additionally, our method outperformed ContentGAN in generating more visually pleasing layouts. Overall, the results validate the effectiveness and superiority of our approach.

Table 3: Ablation of primary components in our framework. "w/o $\mathcal{R}$" and "w/o $\mathcal{G}$" denotes model without layout ranking module and without aesthetics conditioning. "Full" denotes our full model. "None" denotes LayoutDM.

	Ovrlp↓	Und.↑	Align.↓	FID↓	Sal.↓	Balance↓
None	0.0176	0.9728	0.0033	5.6	0.158	74.6
w/o $\mathcal{R}$	0.0094	0.9873	0.0022	12.05	0.139	75.2
w/o $\mathcal{G}$	0.0117	0.9778	0.0034	11.70	0.096	59.5
Full.	0.0102	0.9922	0.0027	13.62	0.092	58.5

4.3.2 Diversity performance. The diversity of generation results is essential in evaluating layout generation methods, as designers may want to generate diverse and high-quality layouts for reference. However, we discover that the SOTA poster layout generation method CGL [36] produces no diversity results. CGL can only generate one unique layout result when given a specific background image. In contrast, our approach is capable of generating a wide range of diverse layout results, as demonstrated in Figure 4. Specifically, we generate multiple layout results conditioned on the same layout element attributes randomly sampled from the validation set. As shown, our method maintains high-quality results while also incorporating desired diversity.

4.3.3 Layout proposals with different ranks. To demonstrate the effectiveness of our layout ranking module, we present layout proposals with varying ranks given the same input background image in Figure 5. We observe that the highly-ranked layout proposals effectively highlight the salient object, maintain the balance of the page, and preserve the overall quality. However, the lowly-ranked proposals often occlude the salient object in the background, indicating the importance of our proposed layout ranking module. The second stage, with layout ranking, improves the adherence to aesthetic principles of the layout results. The original advertising poster layout dataset contains well-balanced and aesthetically pleasing layouts where the salient object in the background is not occluded. Hence, layout proposals that closely resemble these human-designed layouts rank highly.

4.4 Ablation Study

We conducted ablation studies to demonstrate the effect of two key components in our framework: the aesthetics-conditioned layout generation module and the layout ranking module. To verify the effect of the layout generation module, we replace the first stage with native layout sampling. To verify the effect of the layout ranking model, we skip the second stage and only generate one layout proposal. Quantitative results are presented in Table 3. We discover that the aesthetics conditioning significantly enhances explicit aesthetic metrics such as overlap and alignment, while the

Figure 5: Layout proposal examples. We show the three proposals with the highest ranks and the lowest ranks.

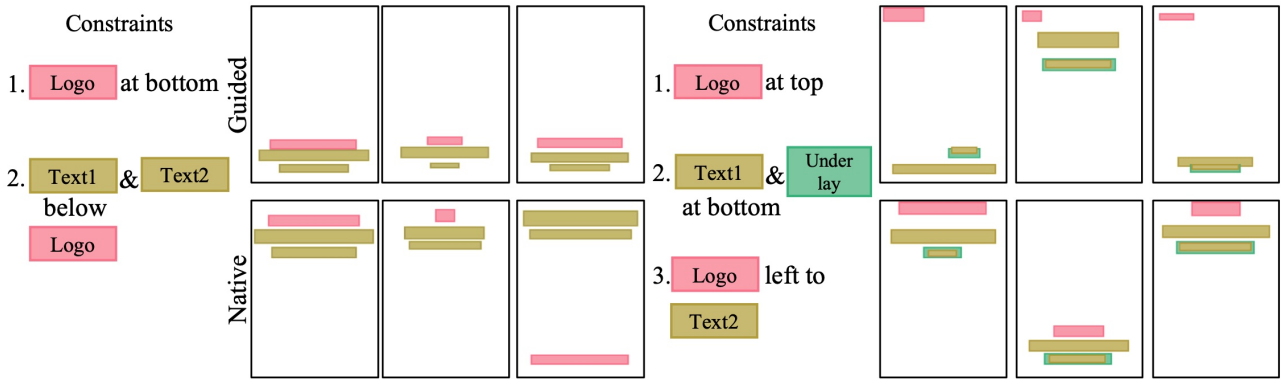


Figure 6: Relationship-constrained layout generation. First row: layout samples generated using our method with relationship constraints. Second row: layout samples generated using native DDIM sampler.

layout ranking module further improves the model's performance on content-aware metrics.

4.5 Extended Layout Generation Tasks

Our layout generation module can incorporate not only aesthetic constraints, but also other user-specific constraints. Inspired by NDN [21] and CLG-LO [17], we use a complete graph to represent the relative positional relationships between elements in a poster layout. Then we design an objective function F_{loc} to calculate the loss when the specified relationships are unsatisfied. The results are shown in Figure 6. We can see that, in most cases, the results generated by the relationship-guided layout sampling perfectly adhere to the given constraints of relative element positions. In contrast, the native sampling exhibits greater randomness.

5 LIMITATIONS

When modeling the relationships between layout elements and background, our method solely considers the background's saliency information, failing to leverage its semantic information fully. It's also worth noting that our method is much slower than other approaches. This is because our method has to run the neural network multiple times and perform additional optimization and ranking.

6 CONCLUSION

In this paper, we propose a novel two-stage content-aware layout generation framework. We first parameterize layouts and detect the salient object in the input background images. Then we employ a diffusion model based layout generation model to generate layout proposals conditioned on aesthetic constraints and element attributes. Finally, we rank the generated proposals to select the result that best satisfies aesthetic principles. Quantitative and qualitative results demonstrate that our method outperforms the state-of-the-art content-aware layout generation methods.

ACKNOWLEDGMENTS

This work was supported in part to Dr. Liansheng Zhuang by NSFC under contract No. U20B2070 and No. 61976199, in part to Dr. Fengying Yan by NSFC under contract No. 42341207, and in part by the Key R&D Program of Zhejiang Province (No. 2022C01025).

REFERENCES

[1] Diego Martín Arroyo, Janis Postels, and Federico Tombari. 2021. Variational Transformer Networks for Layout Generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13637–13647. https://doi.org/10.1109/CVPR46437.2021.01343

[2] Yunning Cao, Ye Ma, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. 2022. Geometry Aligned Variational Transformer for Image-Conditioned Layout Generation. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 1561–1571. https://doi.org/10.1145/3503161.3548332

[3] Shang Chai, Liansheng Zhuang, and Fengying Yan. 2023. LayoutDM: Transformer-based Diffusion Model for Layout Generation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. arXiv eprint. https://doi.org/10.48550/arXiv.2305.02567

[4] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil Bharath. 2017. Generative Adversarial Networks: An Overview. *IEEE Signal Processing Magazine* 35 (10 2017). https://doi.org/10.1109/MSP.2017.2765202

[5] Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. 2018. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765* (2018).

[6] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 8780–8794. https://proceedings.neurips.cc/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[8] Shunan Guo, Zhuochen Jin, Fuling Sun, Jingwen Li, Zhaorui Li, Yang Shi, and Nan Cao. 2021. Vinci: An Intelligent Graphic Design System for Generating Advertising Posters. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 577, 17 pages. https://doi.org/10.1145/3411764.3445117

[9] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry Davis, Vijay Mahadevan, and Abhinav Shrivastava. 2021. LayoutTransformer: Layout Generation and Completion with Self-attention. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 984–994. https://doi.org/10.1109/ICCV48922.2021.00104

[10] Jonathan Ho. 2022. Classifier-Free Diffusion Guidance. *ArXiv* abs/2207.12598 (2022).

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851. https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[13] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022. ProDiff: Progressive Fast Diffusion Model for High-Quality Text-to-Speech. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 2595–2605. https://doi.org/10.1145/3503161.3547855

[14] Ali Jahanian, Jerry Liu, Qian Lin, Daniel Tretter, Eamonn O'Brien-Strain, Seungyon Claire Lee, Nic Lyons, and Jan Allebach. 2013. Recommendation System for Automatic Design of Magazine Covers. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, California, USA) *(IUI '13)*. Association for Computing Machinery, New York, NY, USA, 95–106. https://doi.org/10.1145/2449396.2449411

[15] Zhaoyun Jiang, Shizhao Sun, Jihua Zhu, Jian-Guang Lou, and Dongmei Zhang. 2022. Coarse-to-Fine Generative Modeling for Graphic Layouts. *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (06 2022), 1096–1103. https://doi.org/10.1609/aaai.v36i1.19994

[16] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. 2019. LayoutVAE: Stochastic Scene Layout Generation From a Label Set. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9894–9903. https://doi.org/10.1109/ICCV.2019.00999

[17] Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. 2021. Constrained Graphic Layout Generation via Latent Optimization. In *ACM International Conference on Multimedia (MM '21)*. 88–96. https://doi.org/10.1145/3474085.3475497

[18] Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. 2021. Constrained Graphic Layout Generation via Latent Optimization. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 88–96. https://doi.org/10.1145/3474085.3475497

[19] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2014).

[20] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. 2022. BLT: bidirectional layout transformer for controllable layout generation. In *European Conference on Computer Vision*. Springer, 474–490.

[21] Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B. Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. 2020. Neural Design Network: Graphic Layout Generation with Constraints. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 491–506.

[22] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. 2021. LayoutGAN: Synthesizing Graphic Layouts With Vector-Wireframe Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 7 (2021), 2388–2399. https://doi.org/10.1109/TPAMI.2019.2963663

[23] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. Diffusion-LM Improves Controllable Text Generation. *ArXiv* abs/2205.14217 (2022).

[24] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2015. DesignScape: Design with Interactive Layout Suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1221–1224. https://doi.org/10.1145/2702123.2702149

[25] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014. Learning Layouts for Single-PageGraphic Designs. *IEEE Transactions on Visualization and Computer Graphics* 20, 8 (2014), 1200–1213. https://doi.org/10.1109/TVCG.2014.48

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[27] Akshay Gadi Patil, Omri Ben-Eliezer, Or Perel, and Hadar Averbuch-Elor. 2020. READ: Recursive Autoencoders for Document Layout Generation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020), 2316–2325.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs.CV]

[29] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. 2018. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3379–3388.

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[31] Jun Wei, Shuhui Wang, and Qingming Huang. 2020. F^3Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 12321–12328.

[32] Kota Yamaguchi. 2021. CanvasVAE: Learning to Generate Vector Graphic Documents. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 5461–5469. https://doi.org/10.1109/ICCV48922.2021.00543

[33] Xuyong Yang, Tao Mei, Ying-Qing Xu, Yong Rui, and Shipeng Li. 2016. Automatic Generation of Visual-Textual Presentation Layout. *ACM Trans. Multimedia Comput. Commun. Appl.* 12, 2, Article 33 (feb 2016), 22 pages. https://doi.org/10.1145/2818709

[34] Peiying Zhang, Chenhui Li, and Changbo Wang. 2020. Smarttext: Learning To Generate Harmonious Textual Layout Over Natural Image. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6. https://doi.org/10.1109/ICME46284.2020.9102780

[35] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. 2019. Content-Aware Generative Modeling of Graphic Design Layouts. *ACM Trans. Graph.* 38, 4, Article 133 (jul 2019), 15 pages. https://doi.org/10.1145/3306346.3322971

[36] Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. 2022. Composition-aware Graphic Layout GAN for Visual-Textual Presentation Designs. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 4995–5001. https://doi.org/10.24963/ijcai.2022/692 AI and Arts.

[37] Yibo Zhou. 2022. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7379–7387.