Visual Layout Composer: Image-Vector Dual Diffusion Model for Design Layout Generation

Mohammad Amin Shabani[1]* Zhaowen Wang[2] Difan Liu[2] Nanxuan Zhao[2]
Jimei Yang[2] Yasutaka Furukawa[1]
[1]Simon Fraser University [2]Adobe Research

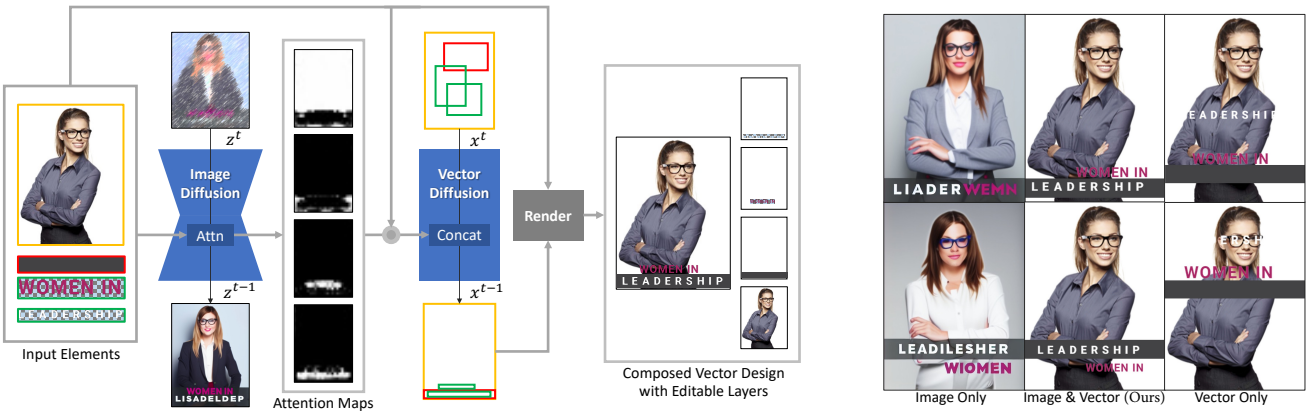

Figure 1. Overview of our method. Left: our image-vector dual diffusion model architecture where both image and vector diffusion are conditioned on input elements and exchanging information with attention maps. Right: a comparison showing the advantage of dual-domain (image and vector) over traditional single-domain approaches.

Abstract

This paper proposes an image-vector dual diffusion model for generative layout design. Distinct from prior efforts that mostly ignores visual information of elements and the whole canvas, our approach integrates the power of a pre-trained large image diffusion model to guide layout composition in a vector diffusion model by providing enhanced salient region understanding and high-level inter-element relationship reasoning. Our proposed model simultaneously operates in two domains: it generates the overall design appearance in the image domain while optimizing the size and position of each design element in the vector domain. The proposed method achieves the state-of-the-art results on several datasets and enables new layout design applications. Project webpage: https://aminshabani.github.io/visual_layout_composer.

1. Introduction

Layout design is essential in creative tasks, determining the spatial arrangement of texts, images, and other visual elements, which are composed into a visual design, such as book and magazine covers, posters showcasing new products, or invitation cards for holiday parties. Graphic designers rely on their keen artistic sensibility in meticulously optimizing the arrangements of the visual elements while ensuring that the composited image conveys the intended message with aesthetic appeals.

Machine learning techniques have been successful in directly regressing the layout vector parameters, i.e., the bounding box coordinates of elements. Both transformer-based models [1] and more recent diffusion-based models [18] have achieved promising results that capture basic geometrical properties such as alignment and symmetry. However, most of these models ignore the visual information from the input elements, including image content, text typeface, and color/shape of vector arts. This leads to unsatisfactory designs as illustrated by the *Vector Only* results in Figure 1. The lack of layout datasets with associated images also contributes to the negligence of visual information. The most commonly used RICO [8] and PubLayNet [45] datasets only contain box representation of design elements. The Crello [40] dataset provides element-level image rendering for over 23K design templates, whose visual features are used in the FlexDM [18] model. However, the interplay of element appearance when rendered on

*This work was done during the first author's internship at Adobe.

the same canvas has never been studied at any large scale.

The recent success of image generative models [6, 9, 30, 32, 41] has enabled the generation of layout design in the image space. Trained on millions of high quality design images from web, these large generative models produce visually diverse and compelling layouts as raster images, demonstrating sophisticated design principles such as the placement of major subjects, style coherence, and perceptual contrast. See the *Image Only* results as examples in Figure 1. These visual properties, vital to layout design, are learned in a holistic way, not straightforward by previous vector domain approaches. Despite high visual quality, the layout designs in image format forbid further editing and customization. Moreover, current generative models are controlled only by text prompts, and cannot use given visual assets as required in many real-world applications.

Motivated by the complementary strengths of the vector and image domains, we propose a combined approach that generates layout designs in both image and vector domains simultaneously. An image branch is a pre-trained text-to-image stable diffusion model [30], which refines (i.e., denoises) a design as an image, given a variable number of design elements with different media types. A vector branch refines per-element bounding box parameters conditioned on attention maps from the UNet of the image branch. Ground truth element masks are used to apply loss directly to these attention maps, whose manipulation at generation time will allow users to control the positioning of elements. This dual diffusion model refines the visual design in the vector and image domains simultaneously. The final vector output assembles the input elements and renders the visual design while allowing further editing by users. The overall proposed method is illustrated on the left of Figure 1.

For training large diffusion models, we have collected a large-scale design Poster dataset from public design web pages by extracting both images and metadata. This effort yields 140K image-layout paired samples as our own dataset, more than six times of the current largest visual layout dataset Crello with 23K samples. We conduct experiments on both datasets to validate the effectiveness of our dual diffusion model. Our data collection tools and processing pipeline can facilitate future efforts to curate even larger dataset for exploring visual design.

Our key contributions are summarized below:

- A novel dual diffusion model that generates layout design in both image and vector domains conditioned on input design elements. The two diffusion processes interact by exchanging element attention maps, where the final results attain both vector editability and visual quality.
- Creative design applications, enabling element-wise user controls, such as canvas resizing and design variation.
- State-of-the-art layout generation results on the existing datasets and our Poster dataset.

2. Related Works

2.1. Vector-only Layout Generation

Generative models for layout generation have seen significant developments. Yamaguchi [40] introduced a VAE-based architecture for unconditional generation of vector graphic documents. Following this, Li et al. [26] proposed LayoutGAN, which uses a relational generator and a wireframe renderer for training with a pixel-based discriminator. Kikuchi et al. [20] further refined this approach with LayoutGAN++, incorporating user-specified constraints into layout generation. Chai et al. [4] proposed using DDPM for layout generation. Inoue et al. proposed LayoutDM [17], a discrete diffusion-based model for layout designs. Zhang et al. [42] and Hui et al. [16] both contributed to this domain with LayoutDiffusion and LDGM, respectively, each based on discrete diffusion models. Cheng et al. [7] proposed a unique approach using guidelines as input conditions for the latent diffusion model.

2.2. Visual-guided Layout Generation

Zhou et al. [46] tackled image-composition-aware layout generation with a multi-stage GAN-based method, an approach further improved by Xu et al. [39] through the use of a pixel-level discriminator. Zheng et al. [44] proposed a GAN based approach using image-based representation for layouts, which requires subsequent post-processing. Lin et al. [28] introduced a multi-step approach dedicated to the cleansing and retargeting of advertising posters. Hsu et al. [15] introduced DS-GAN, a CNN-LSTM-based conditional GAN method for poster layout generation. Inoue et al. [18] proposed Flex-DM, utilizing a transformer to predict masked attributes in design. Shimoda et al. [34] and Tang et al. [36] expanded the field by proposing new models for typography generation and layout generation, respectively. Our work differs by focusing on enhancing visual information and consistency through a large image diffusion model.

2.3. Image-based Diffusion Models

Image diffusion model [35] was proposed initially for unconditional image generation. With the rapid development of the large-language model (LLM) [3], it has shown groundbreaking ability for text-to-image generation tasks [14]. To enhance the computational efficiency, rather than performing diffusion steps at the pixel level, latent diffusion model (LDM) [30] conducted operations on latent space, further boosting the generation ability of diffusion models. It has been widely used in various applications, including image-to-image translation [31], style transfer [23, 37], and condition-guided editing (*e.g.*, sketch, layout, depth map, etc.) [27, 43]. To enable multi-object composition closer to our work, Xiao et al. [38] introduced

Localization Loss for free text-to-image generation with multiple subjects, while Sarukkai et al. [33] developed Collage Diffusion for creating realistic photos, focusing on harmonization and fidelity. Balaji et al. [2] training an ensemble of diffusion models specialized for different parts of the diffusion model to improve the overall text alignment of the model. Goel et al. [10] proposed Pair Diffusion to edit object properties in images via structure and appearance decomposition. Our model stands out by not needing predefined element positions and scales, handling multiple subjects, maintaining visual consistency, and providing enhanced editing flexibility.

3. Method

Our Visual Layout Composer (VLC) model consists of two diffusion models that operate in parallel: one in the vector domain and the other in the image domain. The two models exchange features at intermediate stages, enabling them to produce outputs that are not only consistent but also leverage the information from both domains. In the following subsections, we first describe our problem formulation, followed by the specifics of the diffusion models for each domain, and how they collaborate to enhance output consistency. Figure 2 shows the model overview.

3.1. Preliminary

Our input is a set of N design elements $E = \{e_1, e_2, \ldots, e_N\}$. Each element e_i is associated with a class category c_i indicating its asset type (such as text, image, shape), an index o_i indicating its layer order relative to other elements, and an RGBA image $\mathbf{I}_i$ of its rendering. On a designated 2D canvas with aspect ratio r_c, the task is to predict an upright bounding box for each element e_i, defined by the 2D coordinates of the top left and bottom right corners as a 4D vector x_i. The goal is to generate a layout design with high visual quality by composing each element e_i on the canvas with the position and size specified by x_i.

3.2. Condition Embedding

After resizing each element image $\mathbf{I}_i$ of aspect ratio r_i to 224×224, we employ CLIP [29] followed by a linear layer to derive its feature representation in the model's hidden dimension d. In parallel, we encode the vector attributes of each element by a vector attributes encoder. The vector condition is constructed by concatenating the element image ratio r_i, canvas image ratio r_c, one-hot encoded version of the element's type c_i and the order of the element o_i. A linear layer maps the concatenated feature to the model's dimension d. The image feature and the vector feature are concatenated and processed by a linear layer to form the final condition feature of size $N \times d$. For each image or vector diffusion model, this condition feature is fed into a transformer-based condition processing module (Figure 2), yielding outputs of sizes $N \times 768$ for image condition feature and $N \times d$ for vector condition feature.

3.3. Network Architecture

The diffusion model in the image domain utilizes a pretrained Stable Diffusion V1.5 model [30], retaining all layers frozen except for the attention modules. During a given time step t, a noisy latent image z^t, representing the target layout rendering, will be passed to the U-Net. Contrary to the original Stable Diffusion, which uses text prompts, we pass the processed image condition feature as the input to the cross-attention layers.

The diffusion model in the vector domain produces the bounding box positions as a 4D vector x_i. We embed the noisy input coordinates x^t to a $N \times d$ feature vector, concatenate with the vector condition feature, and map to the model's dimension d by a linear layer. These features are fed into a transformer-based diffusion denoiser to estimate the corresponding noise for each element x_i.

Information exchange between the image and vector domains is challenging, where we found that direct feature space sharing as in ControlNet [43] is not effective. We use the processed attention scores from U-Net as a shared medium. The advantages of this method are: (i) it aligns the number of attention scores with the number of input elements, enabling a straightforward concatenation of each image mask to its respective vector model feature token; (ii) it provides a general indication of each element's location in the image domain, simplifying the interpretation for the vector domain; and (iii) it allows the vector domain to adjust these scores to enhance the U-Net's focusing capability, with a mechanism similar to Attention Modulation [21].

Concretely, we extract a set of attention scores from the cross-attention module of the U-Net's middle block. Given the scores with 8 heads and size 8×8 for each design element, we flatten the tensors to form a set of vectors with the total size of $N \times 512$. This feature is embedded to $N \times 1024$ and concatenated with each input element in the vector domain prior to the transformer-based diffusion denoiser. The transformer model produces an output of $N \times 2048$, which is passed through linear projection layers to yield denoised estimation of x^{t-1} and the attention modulation residual of size $N \times 512$. We add the residual values to the original attentions to form the refined attention scores. These scores are sent back to the U-Net, promoting interactive and iterative information exchange between the two domains.

During the diffusion process, repetitive elements can lead to omissions or misalignments in the image. To mitigate this, we adopt the cross-attention localization loss [38]. This loss refines the attention scores of each region, ensuring a more accurate correspondence to its respective element. We apply this loss within the U-Net's attention modules to further enhance the quality of the attention scores.

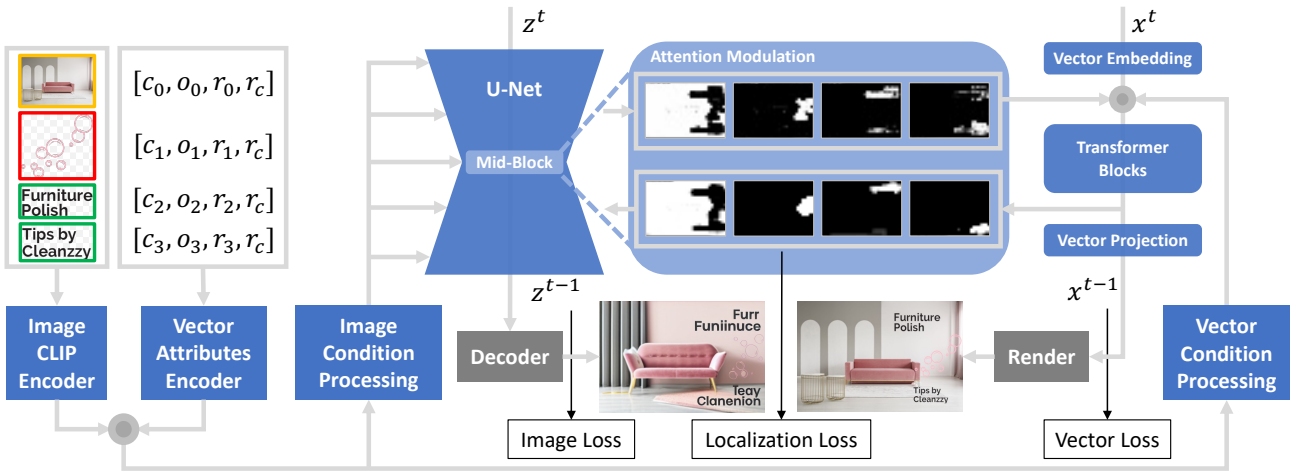

Figure 2. Illustration of the proposed model. Given a set of element images and additional attributes, the model simultaneously employs dual diffusion processes to generate a cohesive image layout and the corresponding bounding boxes for each element.

Figure 3. Generated image/vector layout without or with canvas masking. By masking intermediate outputs of the diffusion model in image domain, we optimize the image output for the canvas size and increase the consistency between the vector-image domains.

The training losses for the image and vector diffusion models are the same as the original DDPM [13].

3.4. Controllable Element Composition

Our experiments showed that fine-tuning the image domain led the model to disregard the specified canvas aspect ratio, favoring the generation of square-shaped layout designs. This tendency not only reduces the quality and diversity of the generated designs but also increases the discrepancy between the image and vector domains, particularly given that the vector domain generally follows the canvas's aspect ratio. To address this problem, we introduce a straightforward yet efficient strategy: masking the noisy latent feature map of z^t at each time step to have zero value outside the canvas region, which ensures image contents related to input condition only appear inside the canvas. The homogeneous region outside the canvas is cropped from the final generated image, maintaining the original image aspect ratio. Figure 3 shows the effect of this process on both domains.

4. Experiments

Implementation details: In our implementation, we use the pretrained weights of StableDiffusion v1-5 [30] for image domain, and OpenAI's clip-vit-large-patch14 vision model for embedding the image of each element to latent space. For training, we run each model for 250,000 iterations, employing a starting learning rate of 5e-5 which decreases by a factor of 0.5 every 100k steps. The vector model has a hidden dimension of 2048, accommodating a wide range of features including attention scores, vector inputs, and image information. Additionally, we incorporate classifier-free guidance into our process, randomly dropping the image inputs of the elements in 20% of the times. Training is conducted on 8 NVIDIA A100 GPUs, with a total batch size of 128. We pad each layout to square shape and normalize the coordinates to $[-1, 1]$. For both domains, we employ the DDPM scheduler with 1000 steps for training and 50 steps for inference to ensure high-quality layout generation.

Datasets: Our experiments use two groups of datasets.

- **Poster and Crello [40] datasets**: demonstrating our method's ability to capture intricate visual information and the interplay between design elements. These datasets, characterized by their rich graphic content, are instrumental in showcasing the method's effectiveness in a visually complex environment. We gathered Poster dataset, which consists of 137,781 layout designs. From this, we reserved 1,000 layouts for validation and 3,000 for testing. Each layout is represented as an RGBA image and is classified into one of five categories: Frame, Canvas, Shape, Mix, and Text. We have limited the number of elements in any given layout to a maximum of 20 to maintain structural simplicity. Given the intricate nature of

these layouts, we have also provided the sequential order of the elements within each design. The Crello datasets includes 18,768 training, 2,316 validation, and 2,331 test layouts including categories such as vector shape, image, or text placeholder.

- **RICO [8] and PubLayNet [45] datasets**: having limited visual information, which challenges our method. To compensate for this, we assign distinct colors to each category and depict layouts with these colors, simplifying the visuals to fixed-color blocks representing each element. This approach tests our method's ability in generating layouts based primarily on structure rather than rich visual details. The Rico dataset consists of user interface (UI) designs sourced from various Android applications, spanning up to 25 categories of UI elements including text bars, icons, and buttons. PubLayNet includes 360,000 document layout instances from academic papers, categorized into five primary layout elements: text, titles, lists, tables, and figures. We use the processed datasets by [17], obtained by keeping only those layouts containing a maximum of 25 elements each. This results training/validation/test sets of sizes 35,851/2,109/4,218 respectively for Rico and 315,757/16,619/11,142 for PubLayNet.

4.1. Comparisons

For the baselines, we compare our method with recent state-of-the-art methods in Layout Design Generation [5, 11, 12, 17, 19, 20, 22, 24, 25]. We compare our method to the LayoutDM [17] as the state-of-the-art for layout generation. This model only uses element categories, adhering to their original design in discrete diffusion space. Our vector domain model operates in a continuous diffusion space and incorporates each element's image embedding. Finally, our dual-domain model combines element image embeddings with the capabilities of the image diffusion model to get the highest quality.

4.1.1 Quantitative Evaluation

Metrics: We follow the previous work [17] and use metrics including Maximum IoU [20] and FID over extracted features of the layouts [20, 24] for RICO and PubLayNet datasets. The previous methods and measurements are effective when there are no visual data or input images for layout elements. However, assessing layout designs in Poster and Crello datasets is more complex. Even layouts with well-aligned bounding boxes might look poor due to the textures and visual details of each element. Elements can also overlap for shadow or visual effects. Inoue et al. [18] approached this as a reconstruction task, using metrics like cosine similarity between the actual and predicted values. This approach works for simple designs with few elements but limits diversity and learning on more complex datasets. Instead, we propose using FID values over three outputs: bounding boxes, feature embeddings from RGB images created by composing elements within generated bounding boxes, and the directly produced RGB images. Although this is our best option, we realize that FID can be biased to image quality and content similarity which should be removed from the consideration for layout evaluation. Finding better visual metric for layout can be a future direction.

Dataset	Models	Bounding Box	Composed Image	Generated Image
Crello	LayoutDM [17]	10.20	34.43	—
	VLC (Vector only)	0.21	5.74	—
	VLC (Dual-domain)	0.21	5.83	11.42
Poster	LayoutDM [17]	1.81	20.00	—
	VLC (Vector only)	0.19	5.82	—
	VLC (Dual-domain)	0.09	3.75	6.00

Table 1. FID scores of bounding boxes, composed RGB layouts obtained from the bounding boxes, and the directly generated layouts from the image domain. Our vector-only method outperforms LayoutDM by effectively using image information, with our multi-domain approach yielding the best results.

Models	**RICO**		**PubLayNet**	
	FID (↓)	mIoU (↑)	FID (↓)	mIoU (↑)
LayoutVAE [19]	33.3	0.249	26.0	0.316
NDN-none [24]	28.4	0.158	61.1	0.162
LayoutGAN++ [20]	6.84	0.267	24.0	0.263
LayoutTrans [12]	5.57	0.223	14.1	0.272
MaskGIT [5]	26.1	0.262	17.2	0.319
BLT [22]	17.4	0.202	72.1	0.215
BART [25]	3.97	0.253	9.36	0.320
VQDiffusion [11]	4.34	0.252	10.3	0.319
LayoutDM [17]	3.55	0.277	7.95	0.310
VLC (Vector only)	**2.38**	**0.418**	**5.49**	**0.348**
VLC (Dual-domain)	5.60	0.376	5.74	0.331
Validation Set	1.85	0.691	6.25	0.438

Table 2. Quantitative comparison of our method with the baselines using RICO and PubLayNet datasets. Our method achieves superior results compared with the baselines even with no visual information in input conditions.

Results: We evaluate our VLC method using visual information as shown in Table 1. Our dual-domain method stands out for its ability to create visually appealing images by using a large image model. This method is a significant improvement over the LayoutDM [17]. LayoutDM uses only element categories as input, which, while effective in creating quality vectors, often fails to align with the actual visual content of each element. This mismatch results in a

Figure 4. Qualitative comparison of our method with baselines on Crello dataset.

drop in the overall quality of the final image. Our vector domain approach is the most similar one to the transformer-based reconstruction used in FlexDM [18]. It includes both image information and other input conditions, and produces higher-quality images compared to LayoutDM, but has some drawbacks such as occasionally covering faces or not managing color contrast well when elements overlap. Our vector only model and dual-domain model perform similarly on the Crello dataset. The reason can be attributed to the larger number of elements in the dataset, which leads to the content mismatch between image and vector domain such that vector model cannot follow the guidance from image correctly.

In Table 2, we compare our method with existing baselines on the RICO and PubLayNet datasets. The baseline results are reported by Inoue et al. [17]. Remarkably, even when the input elements are limited to only categorical data without any visual information, our method demonstrated its capability to produce high-quality layouts in both vector-only and dual-domain settings. In the vector-only domain, we observed a trend where training appeared less stable. However, despite this instability, the results often surpassed those generated by discrete diffusion models [11, 17]. In addition, the vector domain model tends to align with the performance of our dual-domain model. This alignment is particularly notable as our dual-domain method achieves better results than most of the baselines, underscoring its effectiveness even in the absence of visual input. It is important to note that while our method is slower than the baselines duo to the amount of computation, this trade-off is compensated by its high-quality output and robustness in processing visual information with its diffusion formulation. Overall, these findings suggest that our method not only achieves state-of-the-art performance but also offers promising potential for generating high-quality layouts under varied and challenging conditions.

Figure 5. Qualitative comparison of our method with baselines on Poster dataset.

4.1.2 Qualitative Evaluation

The effectiveness of our dual-domain method is further illustrated in Figure 4&5 with visual results from the Crello and Poster datasets. These examples highlight our approach not only improves the layout quality but also ensures that the final compositions are more visually coherent and appealing, especially when compared to other methods.

4.2. Model Analysis

4.2.1 The Effectiveness on Dual-domain Modeling

In addressing the complexities of dual-domain modeling, a crucial challenge lies in maintaining consistency between

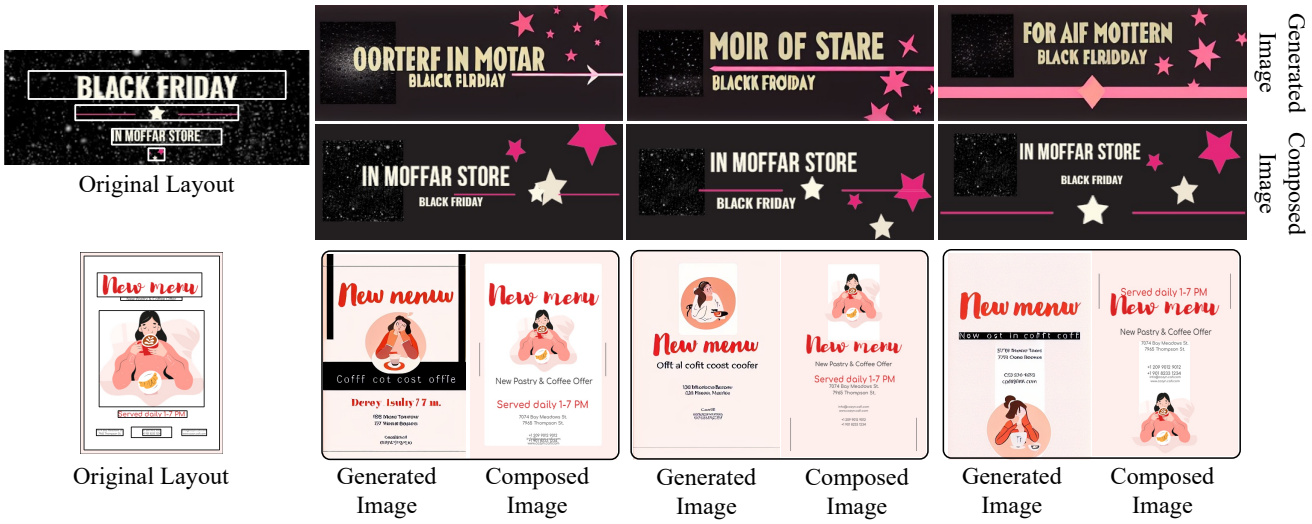

Figure 6. Our approach is capable of producing a wide range of consistent designs in both image and vector domains.

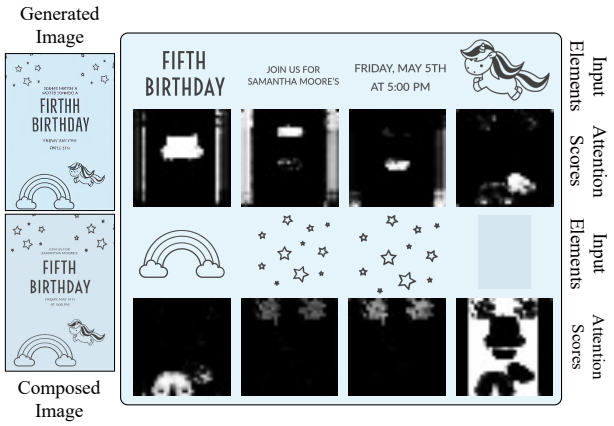

Figure 7. Attention scores are assigned to various elements in the image domain.

Figure 8. Through attention masking, we can decompose all the elements in the latent image space, and include them one by one to create a layered image representation. The bottom right image shows the original design.

the image and vector domains. This is vividly illustrated in Figure 3, which demonstrates that when these two domains adhere to a consistent denoising path, the result quality is significantly enhanced. Conversely, deviation from this path leads to a noticeable drop in quality. Complementing this, Figure 6 highlights the efficacy of our method, showcasing that both the image and vector domains can generate results that are not only consistent with each other but also retain the diversity of the generated layouts.

4.2.2 The Effect of Attention Score Maps

The attention score maps are the key for the visual compositing capability of our model. We visualize the clear spatial correlation between attention scores and corresponding elements in the image domain in Figure 7. Our vector domain model leverages these spatial information to predict vector layouts by adaptively tracking each element and resolving the ambiguity between similar objects.

By manipulating the attention scores, we can also control the element generation in the image domain. For example, we can remove one element from canvas while keeping the rest layout unchanged by masking out its attention map. If we do this sequentially to each element, we can decompose image output into a layered representation organized by elements, as shown in Figure 8. This results in a virtual vector layout format, which may be used for similar purpose as our vector domain model.

4.3. Applications

With its two domain outputs and high controllability through attention maps, our model can be used for a range

Figure 9. Layout controlled visual style variation. Given input layout with bounding boxes, our image model can generate visual results with the same layout and different styles.

Figure 10. Canvas resize with different aspect ratios. Our model can adapt layout under various aspect ratios.

Figure 11. Controlled layout generation. Our model allows users to add conditions on design elements to customize the result. In this example, the user specifies the location of the balloons.

of design applications.

In Figure 9, we produce various visual designs with the same spatial layout conforming to given element boxes by setting the attention scores for each element accordingly. This unlocks the power of the pretrained large image model for visual style variation.

In Figure 10, we generate layouts of different canvas sizes by changing the input condition and masking the intermediate latent states during the denoising process.

In Figure 11, when a user specifies the position of a certain element (the balloons of digits "20") to a desired position, we fix the attention map for this element during inference, and only optimize the layout for the remaining elements.

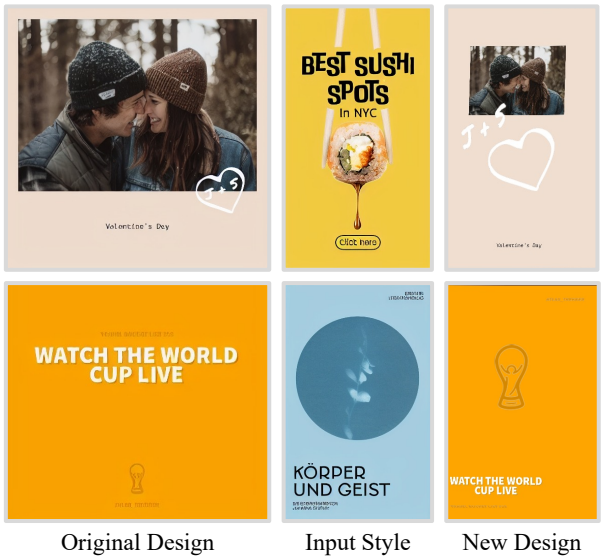

Figure 12. Layout style transfer. Our model can transfer the style from the reference design to the input design.

In Figure 12, we transfer the layout from one design image to a new set of elements by propagating the attention maps from an existing design to a new one. The image model takes a given reference image as the initial noisy latent state, accompanied by a new set of elements. During inference, the image model aligns the visual content to input elements while preserving the initial reference layout. The vector model utilizes the aligned attentions to create a new design incorporating both the input elements and reference layout.

5. Conclusion

Our paper introduces a new approach to layout design, combining vector and image domains through a dual diffusion model. Compared to previous methods solely based on the vector domain, this model has shown notable improvement on layout visual quality while maintaining the flexibility of vector editing on several datasets. Our model also shows flexible controllability via element attention manipulation, enabling a set of design applications across image and vector domains.

While we have made strides in integrating visual elements into layout design, our model is an initial step in addressing this complex task. Our findings offer a new perspective to the layout generation problem, and point out a potential direction for future exploration.

References

[1] Diego Martin Arroyo, Janis Postels, and Federico Tombari. Variational transformer networks for layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13642–13652, 2021. 1

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 3

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[4] Shang Chai, Liansheng Zhuang, and Fengying Yan. Layoutdm: Transformer-based diffusion model for layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18349–18358, 2023. 2

[5] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 5

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[7] Chin-Yi Cheng, Forrest Huang, Gang Li, and Yang Li. Play: Parametrically conditioned layout generation using latent diffusion. *arXiv preprint arXiv:2301.11529*, 2023. 2

[8] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854, 2017. 1, 5

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[10] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023. 3

[11] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 5, 6

[12] Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1004–1014, 2021. 5

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[14] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022. 2

[15] Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6018–6026, 2023. 2

[16] Mude Hui, Zhizheng Zhang, Xiaoyi Zhang, Wenxuan Xie, Yuwang Wang, and Yan Lu. Unifying layout generation with a decoupled diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1942–1951, 2023. 2

[17] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. LayoutDM: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10167–10176, 2023. 2, 5, 6

[18] Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Towards flexible multi-modal document models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14287–14296, 2023. 1, 2, 5, 6

[19] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. Layoutvae: Stochastic scene layout generation from a label set. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9895–9904, 2019. 5

[20] Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. Constrained graphic layout generation via latent optimization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 88–96, 2021. 2, 5

[21] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *ICCV*, 2023. 3

[22] Xiang Kong, Lu Jiang, Huiwen Chang, Han Zhang, Yuan Hao, Haifeng Gong, and Irfan Essa. Blt: bidirectional layout transformer for controllable layout generation. In *European Conference on Computer Vision*, pages 474–490. Springer, 2022. 5

[23] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 2

[24] Hsin-Ying Lee, Lu Jiang, Irfan Essa, Phuong B Le, Haifeng Gong, Ming-Hsuan Yang, and Weilong Yang. Neural design network: Graphic layout generation with constraints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 491–506. Springer, 2020. 5

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and

Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 5

[26] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. Layoutgan: Synthesizing graphic layouts with vector-wireframe adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2388–2399, 2020. 2

[27] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 2

[28] Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. Autoposter: A highly automatic and content-aware design system for advertising poster generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1250–1260, 2023. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 4

[31] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 2

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[33] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. *arXiv preprint arXiv:2303.00262*, 2023. 3

[34] Wataru Shimoda, Daichi Haraguchi, Seiichi Uchida, and Kota Yamaguchi. Towards diverse and consistent typography generation. *arXiv preprint arXiv:2309.02099*, 2023. 2

[35] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[36] Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. Layoutnuwa: Revealing the hidden layout expertise of large language models. *arXiv preprint arXiv:2309.09506*, 2023. 2

[37] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 2

[38] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023. 2, 3

[39] Chenchen Xu, Min Zhou, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Unsupervised domain adaption with pixel-level discriminator for image-aware layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10114–10123, 2023. 2

[40] Kota Yamaguchi. Canvasvae: Learning to generate vector graphic documents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5481–5489, 2021. 1, 2, 4

[41] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2

[42] Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, and Dongmei Zhang. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. *arXiv preprint arXiv:2303.11589*, 2023. 2

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3

[44] Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2

[45] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 1, 5

[46] Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Composition-aware graphic layout gan for visual-textual presentation designs. *arXiv preprint arXiv:2205.00303*, 2022. 2