Unsupervised Domain Adaption with Pixel-level Discriminator for Image-aware Layout Generation Supplementary Material

Chenchen Xu[1,2*] Min Zhou[2] Tiezheng Ge[2] Yuning Jiang[2]
Weiwei Xu[1†]
[1]State Key Lab of CAD&CG, Zhejiang University [2]Alibaba Group
xuchenchen@zju.edu.cn, {yunqi.zm, tiezheng.gtz, mengzhu.jyn}@alibaba-inc.com, xww@cad.zju.edu.cn

1. Demonstration of Advertising Poster Design with Our Layout Results

Designers have applied graphic layouts generated by our PDA-GAN to design aesthetic advertising posters. As shown in Fig. 1, our model generates graphic layouts (middle) with multiple elements conditioned on product images (left). Designers can utilize graphic layouts to make visually pleasing advertising posters (right). More demonstration of advertising posters designed with our layout results can be seen in Fig. 9 and Fig. 10.

Text bounding boxes without underlays in Fig. 1, Fig. 9, and Fig. 10 tend to appear in relatively simple areas of backgrounds. When the areas to place texts are complex, PDA-GAN will simultaneously generates underlay elements to enhance the text readability. Meanwhile, the layout elements are managed to avoid occluding subjects/products for fully representing products.

Figure 1. Demonstration of advertising poster design with our layout results.

2. Domain Gap Visualization

Different marginal distributions across domains are termed as domain gap [2]. In this work, paired images and layouts in the existing dataset [8] are collected by inpainting [7] and annotating posters, respectively. There is a domain gap between inpainted posters (source domain data) and clean product images (target domain data).

To illustrate the domain gap, Fig. 2 shows the details of source and target domain images. We select three clean product images from the target domain and draw graphic elements on them to make posters. We then inpaint these posters to from source domain data. The inpainted regions become distorted and blurred, and the pixel-level domain gap is formed.

3. Metrics

For quantitative evaluations, we follow [8] and divide layout metrics into composition-relevant part and graphic part. The composition-relevant metrics include R_{com}, and R_{shm}, R_{sub}, which measure background complexity, subject occlusion, and product occlusion, respectively. These composition-relevant metrics are more important for this research topic. The graphic metrics include R_{ove}, R_{und}, and R_{ali}, which measure layout overlap, underlay overlap, and

*Work done during an internship at Alibaba Group
†Corresponding author

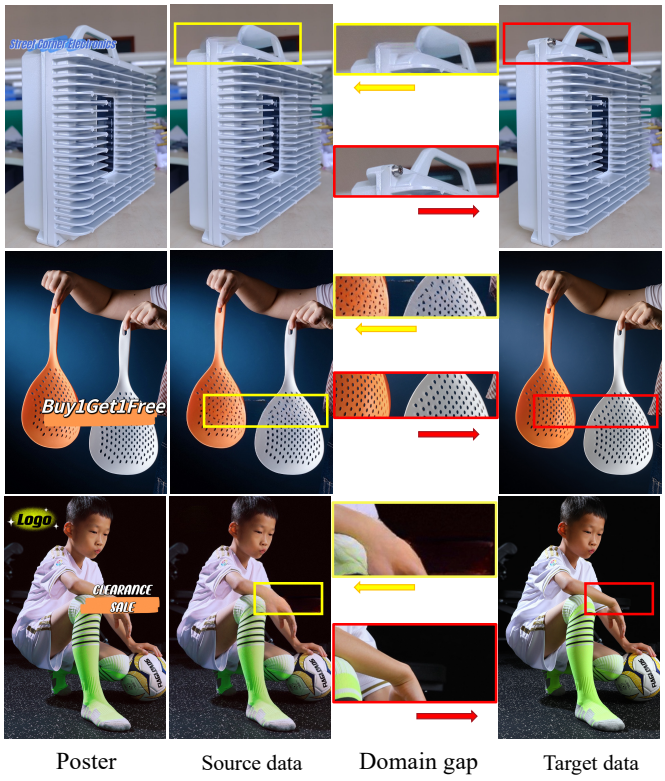

Figure 2. **Domain gap visualization.** To display the domain gap, we manually design several posters based on clean product images (target data), and inpaint the graphic elements to get the source data. Visual contents in inpainted areas (framed with yellow boxes) are distorted and blurred compared with the original contents (framed with red boxes).

layout alignment degree, respectively. In this section, we will briefly interpret the formal definitions of these metrics and utilize them to verify the effectiveness of our model.

3.1. Composition-relevant Measures

We follow the description in [8] to formally define the composition-relevant metrics here. The metric R_{com} is used to measure the background image complexity of the areas placing of text elements (without underlays), and the higher value indicates the lower readability of texts. To calculate R_{com}, we first compute the gradients along x and y directions at each pixel using the Sobel operator, and then obtain the root mean square of these two gradients as the final edge response at each pixel. Finally, R_{com} is computed as the average gradients of all pixels covered by predicted text-only bounding boxes. It can be defined as follows:

$$R_{com} = \frac{\sum_i^N \boldsymbol{S}(R_i)}{N} \tag{1}$$

where $\boldsymbol{S}$ means the gradients are calculated by the Sobel operator. The number of text elements in this layout is N. R_i is the region of the ith text element.

The metric R_{shm} measures how the graphic elements occlude the main subject or product in a background image. Since an aesthetic advertising poster should be able to clearly present subjects or products, we expect this metric to be small for a high quality advertising poster design. To calculate this metric, we respectively feed the salient images with or without masked layout regions into a pretrained VGG16 [6], and calculate the distance L_2 between their output logits.

$$R_{shm} = L_2[\boldsymbol{VGG}(x^{sal}), \boldsymbol{VGG}(x^{sal}, y^l)]. \tag{2}$$

The salient image x^{sal} is sent into a pretrained $\boldsymbol{VGG}$ conditioned on the mask of layout y^l or not.

To calculate R_{sub}, we get attention maps of promoted products (queried by their category tags extracted from product pages) in test images by CLIP[1] [1, 5] and sum the attention values within layout regions.

$$R^s_{sub} = \frac{\sum_i^N \boldsymbol{Rgn}(\boldsymbol{CLIP}(Map^s))}{N}, \tag{3}$$

where R^s_{sub} is the value of R_{sub} for the sample of s. We extract product categories from product pages and use $\boldsymbol{CLIP}$ to get attention Map^s. The symbol $\boldsymbol{Rgn}$ refers to layout bounding box regions. Both R_{shm} and R_{sub} can reflect the occlusion degree of subjects with layouts, and the lower values indicate a layout of better quality.

3.2. Graphic Measures

As mentioned in the paper, advertising poster graphic layouts include four types of elements: logos, texts, underlays, and embellishments. The overlap and alignment of graphic measures used to measure how these types of elements overlap each other are described in [3, 4, 8]. Specifically, the underlay elements are allowed to overlap with any other elements to improve the readability of texts, since it is possible that the desirable color of text is not salient on a background image. The layout also allows a embellishment to overlap with other elements, except other embellishment elements, since the embellishments are usually small for the purpose of decorating a text or logo box.

We follow the equation in [4] to calculate the layout overlap metric as follows:

$$R_{\mathbf{e}} = \sum_{i \in \mathbf{e}} \sum_{(j \in \mathbf{e}) \neq i} \frac{a_i \cap a_j}{a_i}, \tag{4}$$

a_i means the area of the ith box in the set of elements bounding boxes $\mathbf{e}$. In this work, $\mathbf{e}$ refers to the

[1] https://github.com/hila-chefer/Transformer-MM-Explainability

set of elements bounding boxes of the $\{logo, text\}$ or $\{embellishment\}$.

$$R_{ove} = R_{\mathbf{e}}^{\{logo,text\}} + R_{\mathbf{e}}^{\{embellishment\}} \tag{5}$$

The underlay, as a background element, is mainly used to emphasize or highlight another element. Thus, when an underlay element appears, at least another type of element over it should appear simultaneously. A higher R_{und} of the underlay overlap degree metric means better performance. We adapt Eq. 4 to calculate R_{und} as follows:

$$R_{und} = \sum_{i \in \mathbf{e}_1} \max_{j \in \mathbf{e}_2} \frac{a_i \cap a_j}{a_i} \tag{6}$$

$$\mathbf{e}_1 = \{underlay\}$$

$$\mathbf{e}_2 = \{logo, text, embellishment\},$$

where a_i means the area of the ith box in $\mathbf{e}_1$, and $\mathbf{e}_1$ and $\mathbf{e}_2$ represent the set of elements boxes of underlay or other types of elements respectively.

The metric R_{ali} used to manifest that elements in an aesthetic graphic layout tend to align in one dimension. We first follow the equation in [4] to calculate the alignment distance of ith bounding box, which is as follows:

$$d_i = \min\left(\blacktriangledown x_i^l, \blacktriangledown x_i^c, \blacktriangledown x_i^r, \blacktriangledown y_i^t, \blacktriangledown y_i^c, \blacktriangledown y_i^b\right) \tag{7}$$

$\blacktriangledown x_i^l, \blacktriangledown x_i^c, \blacktriangledown x_i^r, \blacktriangledown y_i^t, \blacktriangledown y_i^c, \blacktriangledown y_i^b$ represent the minimum distance between the ith bounding box and other bounding boxes in the left, horizontal midpoint, right, top, vertical midpoint, bottom dimensions respectively. $\blacktriangledown x_i^* \ (* = l, c, r)$ refers to [4] as:

$$\blacktriangledown x_i^* = \min_{j \neq i} \left| x_i^* - x_j^* \right| \tag{8}$$

$\blacktriangledown y_i^* \ (* = t, c, b)$ can be calculated similarly. Thus, d_i is the minimum distance between box i and all other boxes in the above six dimensions. The alignment metric R_{ali} can be defined using d_i as follows:

$$R_{ali} = \sum_{i=1}^{N} \begin{cases} d_i, & d_i - d_t < 0 \\ 0, & d_i - d_t \geq 0 \end{cases} \tag{9}$$

When d_i is larger than the threshold d_t, d_i is set to 0, and N means the total number of predicted elements. This is because, in an advertising poster graphic layout, an element such as a logo often appears in the image's corner and is often far from other elements.

4. More Evaluations

In this section, we will show more quantitative and qualitative evaluations to demonstrate the effectiveness of PDA-GAN.

4.1. Eliminating Domain Gap

As shown in Fig. 4, we select four clean product images x_t from the target domain data and add graphic layout elements to these images into advertising posters x_p. Inpainting the regions of elements in posters to obtain inpainted images x_i. Due to inpainted areas, there is a domain gap between x_t (target domain) and x_i (source domain).

To demonstrate that PDA-GAN can effectively eliminate the domain gap, we input x_i and x_t to CGL-GAN and PDA-GAN to generate layouts. The mean difference values of the shallow-level feature maps, fusion feature maps, and deep-level feature maps generated by CGL-GAN between x_i and x_t of the above four samples as input are 0.0610, 0.1289, and 0.1263. The corresponding mean values calculated by PDA-GAN are 0.0293, 0.0726, and 0.1160, respectively. Compared with CGL-GAN, PDA-GAN has less difference in the generated feature maps between the source and target domain data.

From the perspective of generated results, layouts generated by CGL-GAN conditioned on different domains of the same product images with significant differences. In particular, CGL-GAN tends to generate layout elements in distorted and blurred areas of inpainted regions. In comparison, layouts generated by PDA-GAN with x_s and s_t as inputs are more similar. The above analysis demonstrates that PDA-GAN can effectively eliminate the domain gap caused by inpainting.

4.2. More Qualitative Comparisons with SOTA Methods

To enhance the paper's quantitative evaluations of composition-relevant metrics, more qualitative analyses and comparisons between PDA-GAN and existing methods are presented here. Tab.1 and Tab.2 in the paper show that PDA-GAN performs best on the background complexity metric R_{com}. Correspondingly, as shown in columns 1, 2, and 9 of Fig. 5, bounding boxes of text elements generated by PDA-GAN are more likely to appear in simple background areas, which improves the readability of the text information. As shown in other columns, when the background of the text element is complex, PDA-GAN will generate an underlay bounding box to replace the complex background to enhance the readability of text information.

The paper also shows PDA-GAN achieves the SOTA performance on the occlusion subject degree metric R_{shm}. From the Fig. 6, layout bounding boxes generated by PDA-GAN avoid subject regions nicely. Thus the generated posters better express the information of subjects and layout elements. In particular, it should be noted that bounding boxes generated by PDA-GAN can effectively avoid the critical regions of the subject, such as the human head or face, as shown in columns 1 and 3 of Fig. 6.

Meanwhile, from the paper, PDA-GAN performs better

Model	$R_{com} \downarrow$	$R_{shm} \downarrow$	$R_{sub} \downarrow$	$R_{ove} \downarrow$	$R_{und} \uparrow$	$R_{ali} \downarrow$
Ours$'$	34.07	15.13	0.800	0.0350	0.9259	0.0108
Ours	**33.55**	**12.77**	**0.688**	**0.0290**	**0.9481**	**0.0105**

Table 1. **Quantitative ablation study on PD.** Ours$'$ refers to our model of PDA-GAN without PD module.

Model	$R_{com} \downarrow$	$R_{shm} \downarrow$	$R_{sub} \downarrow$	$R_{ove} \downarrow$	$R_{und} \uparrow$	$R_{ali} \downarrow$
Ours*	36.71	20.14	1.036	0.0475	0.9376	**0.0068**
Ours	**33.55**	**12.77**	**0.688**	**0.0290**	**0.9481**	0.0105

Table 2. **Quantitative ablation study on Gaussian blur.** Ours* means the model of PDA-GAN with Gaussian blur for input image.

than all existing methods on occlusion product degree metric R_{sub}. Compared with existing methods, PDA-AGN generates layout bounding boxes on regions with lower thermal values to avoid occluding products. For example, in columns 2, 4, and 5 of Fig. 7, layout bounding boxes generated by PDA-GAN effectively avoid the region with high thermal values of products, which present clothing products information pretty well.

4.3. Ablations

Effects of PD. Compared with the model without the PD module in the first row of Tab. 1, under the same configuration, the model with the PD module achieves better results in all metrics. Benefiting from PD effectively eliminating the domain gap, as demonstrated in Sec. 4.1, the model with PD module can generate high-quality image-aware graphic layouts for advertising posters.

Affects of Gaussian Blur. In Tab. 2 and Fig. 8, based on the model of PDA-GAN, we both quantitatively and qualitatively analyze the effect of the Gaussian blur on the layout generation. When PDA-GAN utilizes Gaussian blur to generate the graphic layout, R_{com} is increased from 33.55 to 36.71. Boxes 2, 6, and 10 in Fig. 8 show that the background of the text bounding box generated by the model with Gaussian blur is more complex, reducing the readability of the text information. In comparison, boxes 7 and 11 show that the model without Gaussian blur generates the text bounding box with simple background or simultaneously generates an underlay bounding box to replace the complex background.

Tab. 2 shows that R_{shm} and R_{sub} of the model with Gaussian blur are increased from 12.77 to 20.14 and 0.688 to 1.036, respectively. As shown in Fig. 8, layout bounding boxes generated by the model with Gaussian blur are more likely to occlude the subject or product regions. These layouts will diminish the presentation of subjects and layout elements information in advertising posters. These quantitative and qualitative analyses demonstrate that the lost image details caused by Gaussian blur will degrade the quality of the generated image-aware graphic layout.

Figure 3. **Layouts under user constraints.** Left: Input images. Middle: Images with user constraints. Right: Our results.

4.4. User constraints on images.

We can input the user constraints to PDA-GAN in the same way as CGL-GAN such that PDA-GAN can output reasonable layouts according to user constraints on images. Therefore, for a given image, the model can generate various reasonable layouts according to different input constraints. As shown in the 2nd row of Fig. 3, our model can add an underlay for a text box to mitigate the impact of complex backgrounds, and the 3rd row shows that it can also put a text box on a user-input underlay constraint.

References

[1] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, pages 387–396. IEEE, 2021. 2

[2] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. *CoRR*, abs/2010.03978, 2020. 1

[3] Jianan Li, Jimei Yang, Aaron Hertzmann, Jianming Zhang, and Tingfa Xu. Layoutgan: Synthesizing graphic layouts with vector-wireframe adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(7):2388–2399, 2021. 2

[4] Jianan Li, Jimei Yang, Jianming Zhang, Chang Liu, Christina Wang, and Tingfa Xu. Attribute-conditioned layout GAN for automatic graphic design. *IEEE Trans. Vis. Comput. Graph.*, 27(10):4039–4048, 2021. 2, 3

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[7] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 3172–3182. IEEE, 2022. 1

[8] Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Composition-aware graphic layout GAN for visual-textual presentation designs. In *IJCAI*, pages 4995–5001. ijcai.org, 2022. 1, 2

Figure 4. **Layouts generated by different models with source/target domain data.** Inpainted images (source data) and clean images (target data) are both fed into PDA-GAN (ours) and CGL-GAN. The results of PDA-GAN are relatively stable, which indicates that the features of two domains are better aligned by our method.

Figure 5. **Qualitative evaluations of background complexity for different models.** Layouts in each column generated by different models are conditioned with the same product image. DAP-GAN (ours) tends to place text-only elements on relatively simple regions and add underlays for texts on relatively complex regions.

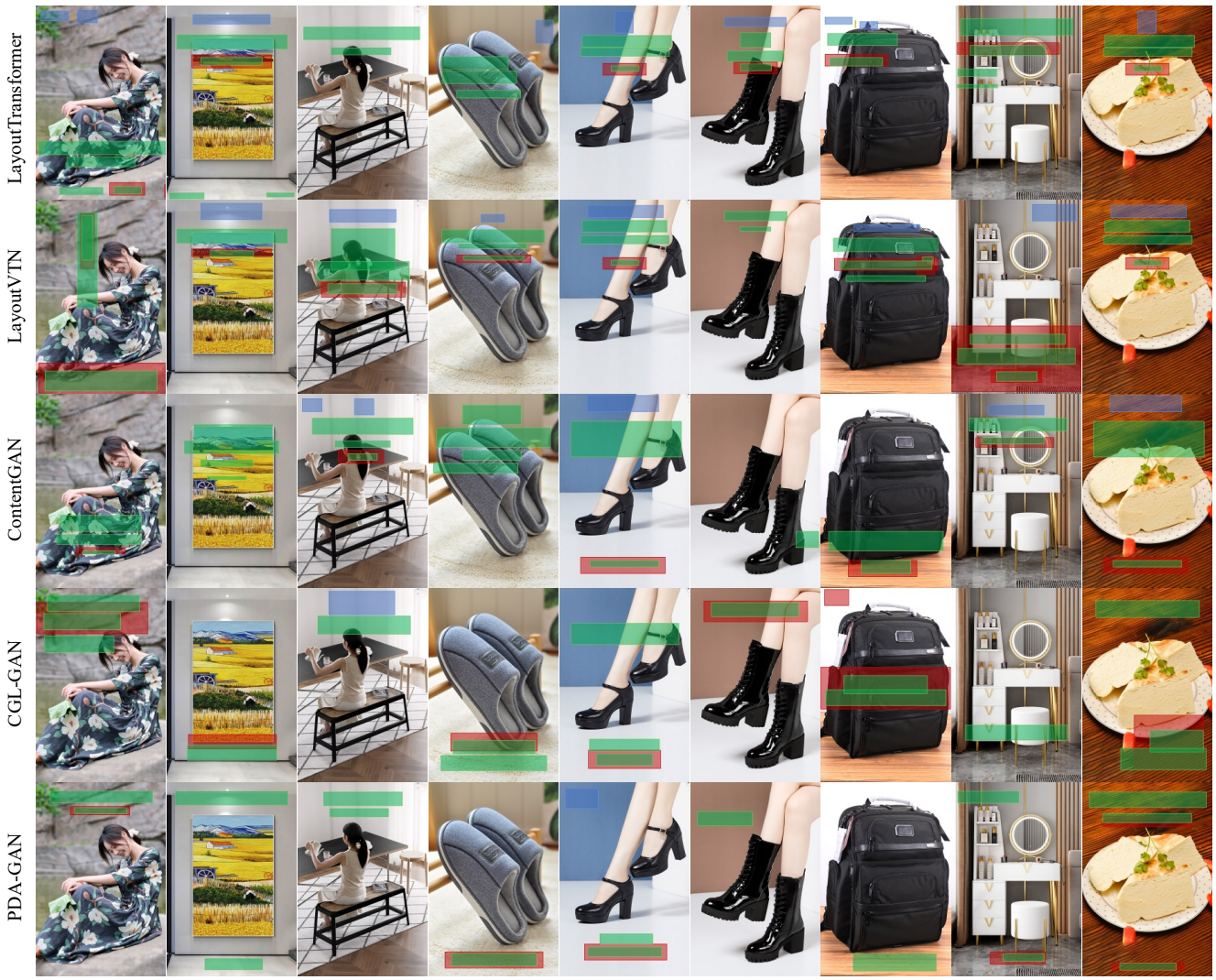

Figure 6. **Qualitative evaluations of occlusion subject degree for different models.** Layouts in each column generated by different models are conditioned the same product image. DAP-GAN (ours) avoids subject occlusion when managing element placements.

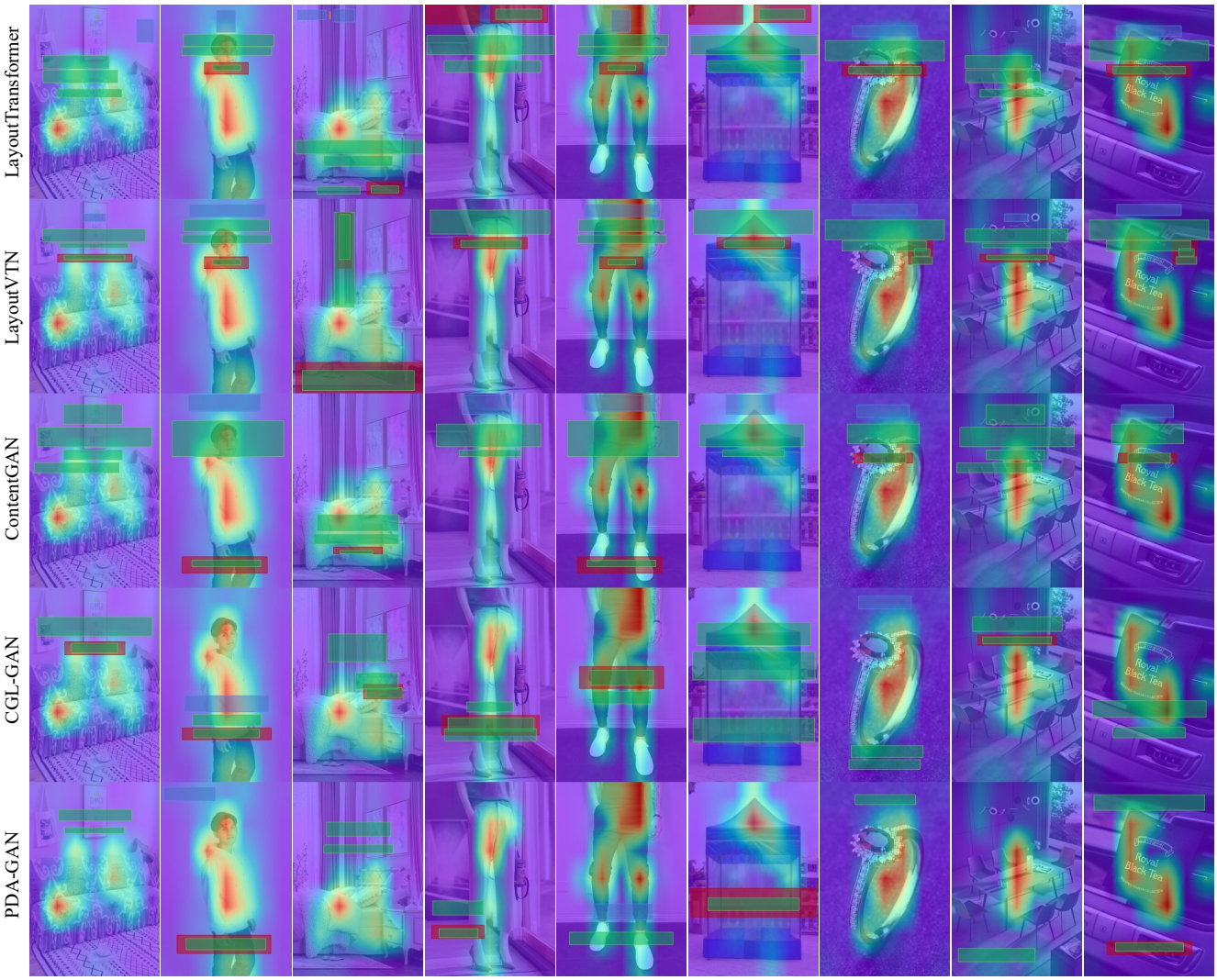

Figure 7. **Qualitative evaluations of occlusion product degree for different models.** Layouts in each column generated by different models are conditioned the same product image. DAP-GAN (ours) avoids product occlusion when managing element placements.

Figure 8. **Affects of Gaussian blur.** Layouts in each row are generated by models with the same image as input. And layouts in each column generated with different inputs. Ours* means using Gaussian blur to process input data. The middle boxes with blue numbers are the enlargement of boxes with yellow numbers on images. The left part of the vertical dotted line is presented with input images, and the right part with product attention heatmaps.

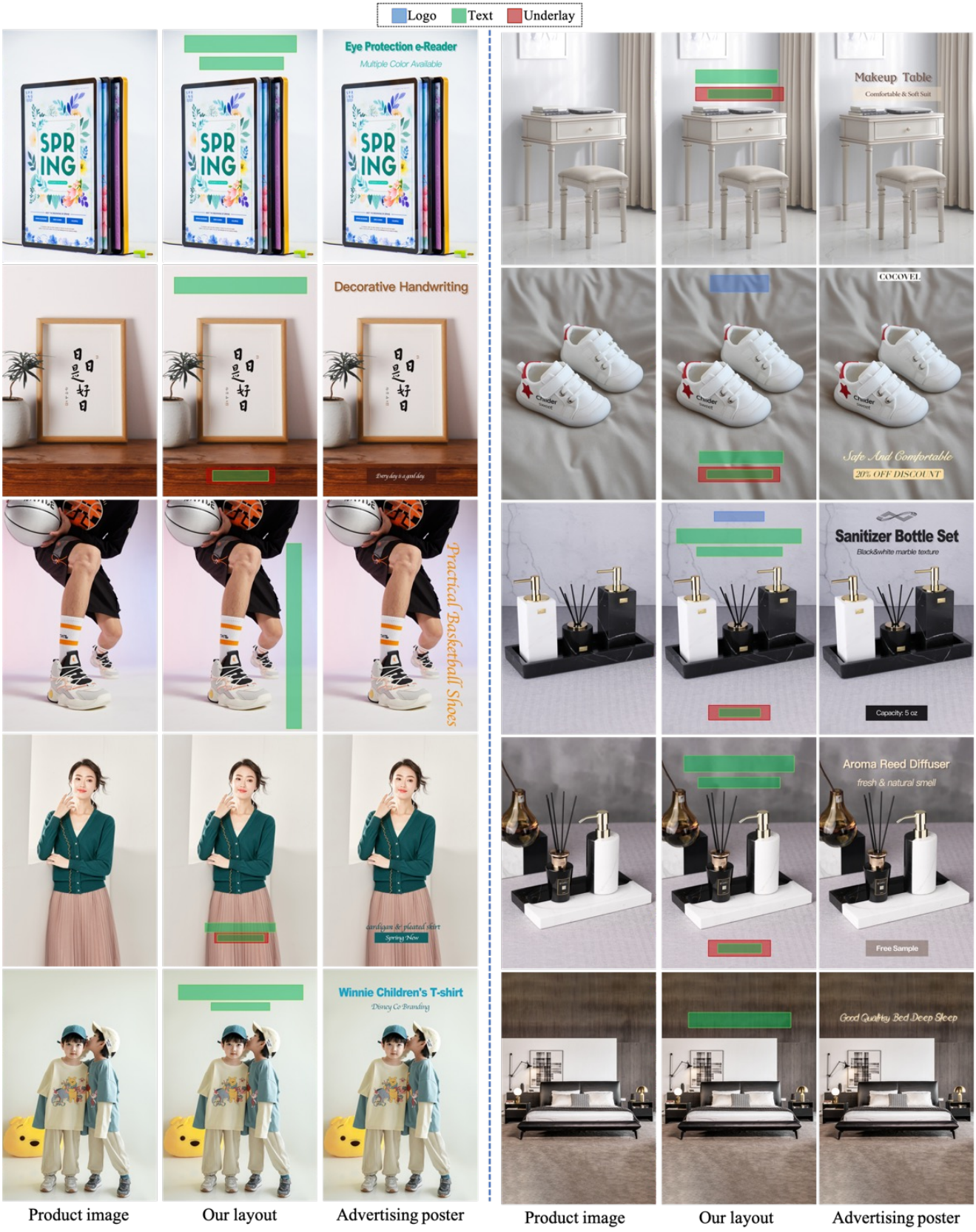

Figure 9. Demonstration of advertising posters designed with graphic layouts generated by PDA-GAN conditioned on product images.

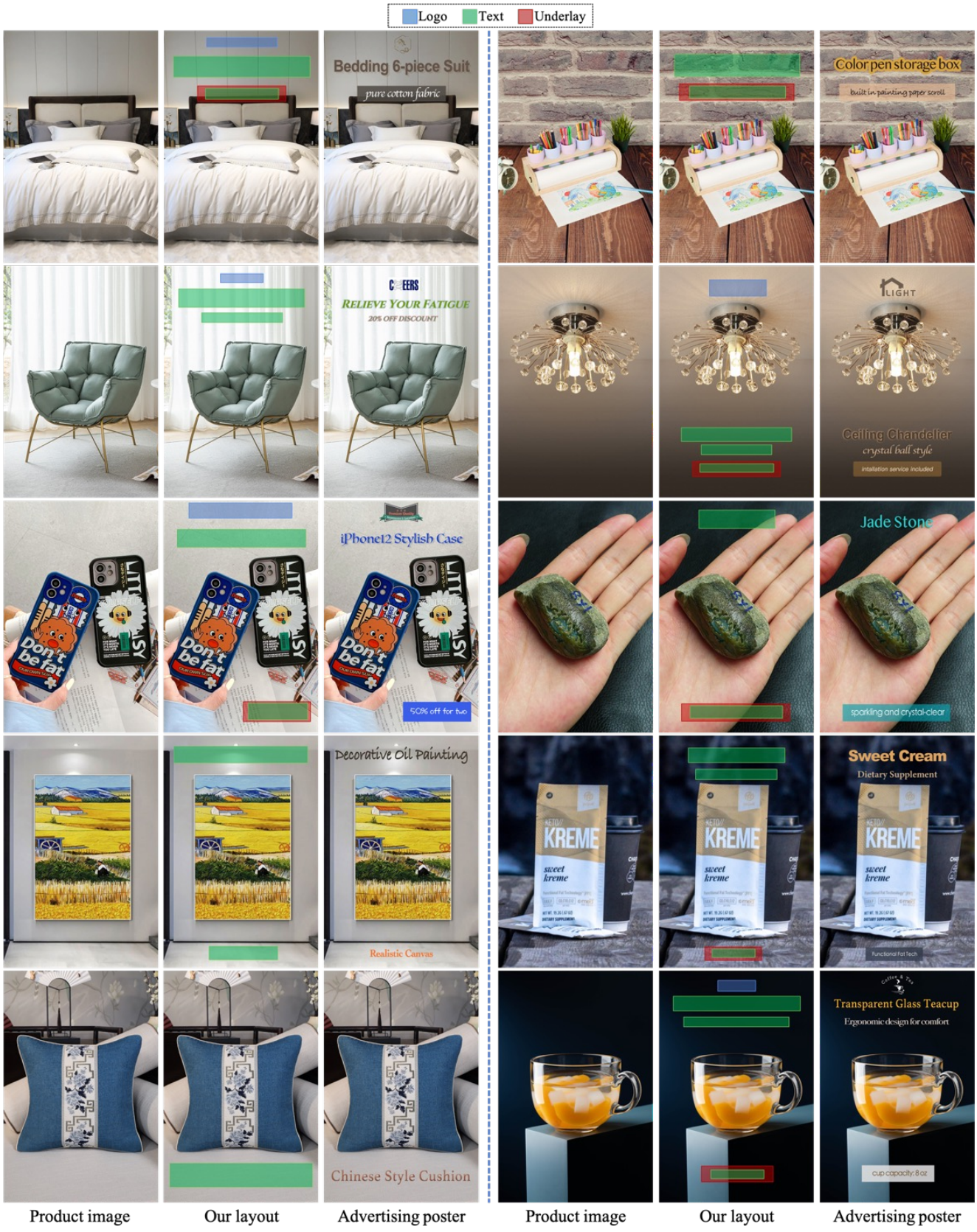

Figure 10. Demonstration of advertising posters designed with graphic layouts generated by PDA-GAN conditioned on product images.