AesthetiQ: Enhancing Graphic Layout Design via Aesthetic-Aware Preference Alignment of Multi-modal Large Language Models

Supplementary Material

Sohan Patnaik* MDSR Adobe
Rishabh Jain* MDSR Adobe
Balaji Krishnamurthy MDSR Adobe
Mausoom Sarkar MDSR Adobe

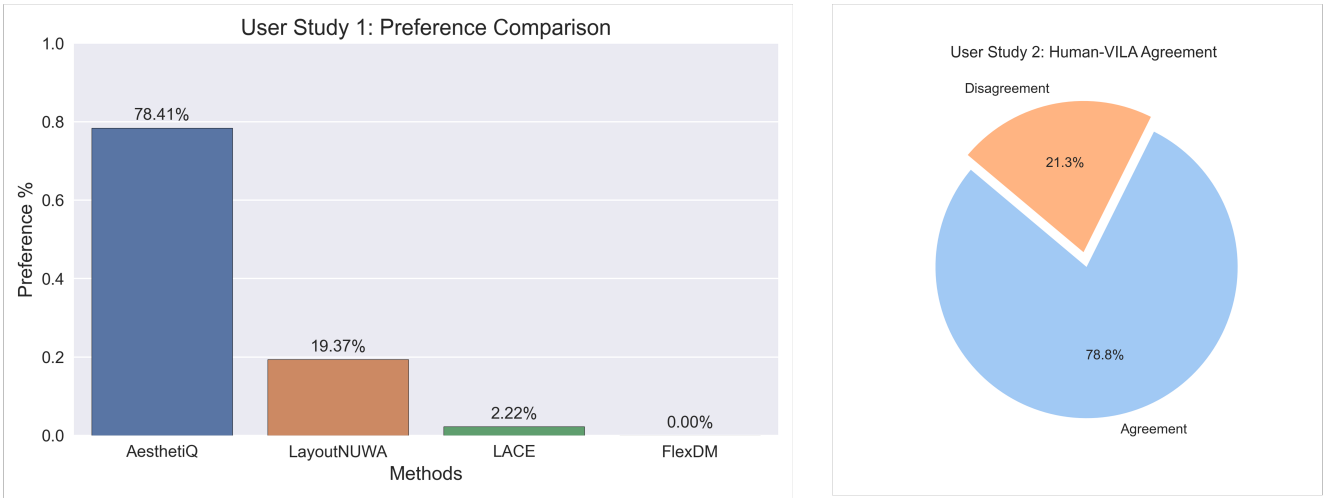

Figure 1. Comparison of results from two user studies evaluating layout aesthetics. The bar plot (left) shows the preference of layout predictions across four methods: AesthetiQ, LayoutNUWA, LACE, and FlexDM, with AesthetiQ significantly outperforming others. The pie chart (right) evaluates alignment between human preferences and the VILA model, achieving a substantial agreement rate of 78.8%. Together, these results highlight the superiority of AesthetiQ in generating aesthetically pleasing layouts and the reliability of VILA as an evaluator.

1. User Study

To evaluate the aesthetic quality of layouts generated by various methods and validate the alignment of human preferences with our approach, we conducted two user studies involving 22 diverse volunteers. The participants were selected to represent a broad spectrum of demographics, including variations in age, gender, occupation, and religion, ensuring a well-rounded and inclusive evaluation. Each participant was presented with a total of 30 questions, designed to capture their aesthetic preferences and opinions on the generated layouts.

User Study 1: Participants were shown layout predictions from four methods: AesthetiQ, LayoutNUWA, LACE, and FlexDM. They were instructed to select the better layout based on aesthetics, alignment and overlap between text and images, and whether the text in the layout made sense. The results are as follows:

- **AesthetiQ**: 78.41%
- **LayoutNUWA** [5]: 19.37%
- **LACE** [1]: 2.22%
- **FlexDM** [2]: 0.00%

These results highlight the significant preference for layouts generated by AesthetiQ compared to the baselines, underscoring its ability to produce more visually appealing and coherent designs.

User Study 2: Participants were shown pairs of model-generated layouts and asked to select the better one using

Method	Mean IoU (%)	$\mathcal{M}_{\text{judge}}$ Win Rate (%)
Desigen [6]	15.36	4.81
LACE [1]	17.88	5.27
PosterLLaVa [7]	30.19	14.73
LayoutNUWA [5]	32.16	15.28
AesthetiQ-1B	38.47	19.29
AesthetiQ-2B	41.42	21.87
AesthetiQ-4B	44.16	22.74
AesthetiQ-8B	**48.29**	**24.48**

Table 1. Comparison of AesthetiQ with baseline methods on the WebUI dataset, evaluated using Mean IoU (%) and $\mathcal{M}$judge Win Rate (%). The results demonstrate the superior performance of AesthetiQ across all model scales, with notable gains in aesthetic and structural alignment metrics.

the same instructions as in the first study. The same layouts were also evaluated by VILA, a Vision-Language Model (VLM) judge. We measured the agreement between human preferences and VILA's outputs, which yielded an alignment score of 78.8%.

This result demonstrates that VILA's aesthetic judgment aligns well with human preferences, further validating its use as an aesthetic evaluator in our framework.

2. Complete Results on WebUI

Table 1 provides a detailed comparison of our approach, AesthetiQ, against baseline methods, including Desigen, LACE, PosterLLaVa, and LayoutNUWA. While previous methods like PosterLLaVa and LayoutNUWA achieve decent performance, they fall short in terms of both structural coherence and aesthetic alignment. In contrast, AesthetiQ shows consistent improvements across all metrics, achieving the highest Mean IoU and $\mathcal{M}_{\text{judge}}$ Win Rate.

We observe a clear trend of performance scaling with model size. The Mean IoU improves progressively from 38.47% for the 1B model to 48.29% for the 8B model. Similarly, the Judge Win Rate increases from 19.29% to 24.48%, showcasing the model's alignment with human aesthetic preferences as the scale grows. In the main paper, due to space constraints, we presented results for only the 1B and 8B variants of AesthetiQ. Here, we include results for the 2B and 4B variants to offer a comprehensive analysis of our model's performance across different scales. The full results underscore the scalability and effectiveness of our approach, particularly in leveraging aesthetic preferences to optimize layout quality.These findings highlight the robustness of AesthetiQ in addressing the challenges of layout generation, establishing a new benchmark for performance on the WebUI dataset.

In the paper, we primarily focus on showcasing qualitative results on the Crello dataset, as it contains individual elements, allowing for detailed analysis and visualization. In contrast, the WebUI dataset only includes category labels and their positions, making it impossible to generate the final rendered templates. For the AAPA evaluation, we render the bounding boxes of the predicted elements on a background, similar to the approach used in Desigen [6]. These renderings are then evaluated by the judge VLM, which selects the layout it deems superior between the two.

AAPA $\mathcal{A}_{\text{judge}}$	Mean IoU (%) ↑			Eval $\mathcal{M}_{\text{judge}}$	$\mathcal{M}_{\text{judge}}$ Win Rate (%) ↑
	All	Single	Multiple		
VILA (Paper)	42.83	52.67	40.64	Vila (Paper) Gpt4o	17.19 14.27
Gpt4o	**44.79**	**55.81**	**43.28**	Vila Gpt4o	**19.41** **15.74**

Table 2. Comparison of AesthetiQ-8B with different training ($\mathcal{A}_{\text{judge}}$) & eval ($\mathcal{M}_{\text{judge}}$) judges. mIoU is independent of $\mathcal{M}_{\text{judge}}$.

3. Stronger MLLM training

We chose VILA-7B as the judge for its open-source, license-friendly nature. Training and evaluation with GPT-4o (Tab 2) improved all metrics with consistent trends in $\mathcal{M}_{\text{judge}}$ win rate across ablations (Paper Fig. 4).

Justification for MLLM as judge: We conduct a user study to measure VILA's correlation with human aesthetic preferences, finding 78.8% agreement. GPT-4o achieves **88.6%** correlation, & AesthetiQ-8B performs better with a stronger judge (Tab 2).

SFT:AAPA split (%)	Mean IoU (%) ↑			Vila Win Rate (%) ↑
	All	Single	Multiple	
100:0 (p=0.5)	40.81	51.82	38.51	16.13
90:10 (p=0.5)	41.17	51.96	39.23	16.4
75:25 (p=0.25)	40.79	50.94	40.64	15.19
75:25 (p=0.5)	42.83	52.67	40.64	17.19
75:25 (p=0.75)	41.88	52.05	40.77	16.74
60:40 (p=0.5)	**43.11**	**53.28**	**41.09**	**17.87**
30:70 (p=0.5)	42.07	52.19	40.04	16.95
0:100 (p=0.5)	40.14	51.5	38.63	15.13

Table 3. Ablation for training Aesthetiq-8b with different data split ratio of crello for supervised and AAPA losses after pretraining

4. Ablation on AAPA mixture:

In our paper, we used 25% of the data for AAPA training. We use equal probability (p=0.5) for either comparing 2 model-generated layouts or a model-generated layout with the ground truth. We extended our experiments to 0, 10, 40, 70, and 100% AAPA mixtures and test different probabilities (p=0.25, 0.75) for selecting between the two settings to further validate AAPA's efficacy. We used two model-predicted layouts in AAPA to promote diversity

and guide the model's internal distribution. Following SFT, we apply RL-based AAPA, as this sequence is most effective for MLLM training in literature. Also, to assess diversity, we measure the average mIoU between two model-generated layouts on the test dataset, with lower mIoU indicating greater diversity. We observe a decrease in mIoU from 72.36 to 67.59 after AAPA, indicating increased diversity.

5. Detailed Experimental Results

This section provides the complete experimental results referenced in the main paper, presented in Table 4. The table details the performance of our models across various configurations, highlighting the effects of scaling, pretraining, VILA alignment, and quality filtering on layout generation tasks. Metrics include All IoU, Single Text IoU, Multiple Text IoU, and Judge Win Rate. These results support the analysis presented in Section 5 of the main paper, showcasing the effectiveness of Aesthetic-Aware Preference Alignment (AAPA) and other components in enhancing the quality and alignment of generated layouts.

6. Direct Preference Optimization

Direct Preference Optimisation (DPO) [4] emerged as an alternative approach to Reinforcement Learning using Human Feedback (RLHF) [3], eliminating the requirement of training a reward model. While RLHF relies on a reward model to evaluate LLM outputs for fine-tuning through reinforcement learning to achieve human preference alignment, DPO takes a different approach. It converts the reward-function loss into a loss over the LLM policy, enabling implicit reward optimization through policy loss optimization. This is achieved using human preference data that pairs two LLM-generated outputs, where one is designated as the winner candidate - y_w and the other as the loser candidate - y_l. Using a static dataset structured as $\mathcal{D} = \{x, y_w, y_l\}$, where x represents the input, the loss is formulated as:

$$\mathcal{L}_R = -log[\sigma(r(x, y_w) - r(x, y_l))] \quad (1)$$

$$r(x, y) = \beta log(\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}) \quad (2)$$

Here, $\pi_\mathcal{Z}(y|x)$ denotes the probability of generating y given input x for model $\mathcal{Z} \in \{\mathcal{M}_{ref}, \mathcal{M}_\theta\}$, where $\mathcal{M}_{ref}$ typically represents the instruction fine-tuned model for LLMs to maintain policy proximity to the initial model, and $\mathcal{M}_\theta$ represents the LLM policy being optimized through DPO. Additionally, σ represents the sigmoid activation, and β is a parameter controlling the deviation extent from the reference model. In essence, this algorithm trains the LLM to develop output preferences among candidates without explicitly modeling rewards. Our algorithm *Aesthetic-Aware Preference Alignment* (AAPA) draws motivation from DPO and carries out preferential training across different layout configurations. For a more comprehensive understanding on DPO, readers are directed to the original publication [4].

7. Prompt Templates for Layout and Judge VLMs

The following prompt template was used as input to our layout generation model $\mathcal{M}_{\text{layout}}$ to guide the generation of aesthetic poster layouts. The template specifies the canvas dimensions and provides a structured description of the elements to be placed, including their type (e.g., text or image), content, and category. This format allows the model to interpret the spatial constraints and semantic attributes of each element effectively, enabling systematic exploration of layout generation. The `<image>` token in the prompt is replaced with a sequence of image embeddings corresponding to the input images, ensuring that the model processes visual information in a compact and meaningful way. By explicitly defining these attributes, the template facilitates reproducibility and evaluation of layout designs. The prompt template is shown below:

```
Consider the image <image> with
height and width of {canvas_height}
and {canvas_width}.  The following
elements need to be placed on the image
to obtain an aesthetic poster layout.

  Element 1:
Text:  LOREM IPSUM
Category:  text

  Element 2:
Image:  <image>
Category:  image
  ...
```

This structured input format ensures that the model can accurately process both visual and textual elements while adhering to aesthetic principles, making it particularly suitable for tasks in computer vision and graphics.

The following prompt is used as input to the judge visual language model $\mathcal{M}_{\text{judge}}$ to evaluate and compare two visual templates based on predefined criteria: aesthetics, clarity, usability, creativity, and consistency. The model processes these criteria to determine which template is superior and outputs the result in a structured JSON format: {`"better_layout": "answer"`}, where the answer specifies the preferred template (`image_1` or `image_2`). This structured approach ensures objective and standardized evaluation of visual designs. The prompt is shown below:

```
  You are a visual language model
designed to evaluate and rate visual
```

Method	LLM	Pretraining	VILA Alignment	Data Filtering	Mean IoU			$\mathcal{M}_{\text{judge}}$ Win Rate (%)
					All	Single	Multiple	
AesthetiQ -1B	Qwen-0.5b	No	No	No	22.06	40.14	24.88	2.18
AesthetiQ -1B	Qwen-0.5b	Yes	No	No	23.95	42.19	26.93	2.95
AesthetiQ -1B	Qwen-0.5b	No	Yes	No	17.45	35.91	20.76	1.64
AesthetiQ -1B	Qwen-0.5b	No	Yes	Yes	21.62	39.56	25.03	1.93
AesthetiQ -1B	Qwen-0.5b	Yes	Yes	No	20.38	38.24	23.92	2.02
AesthetiQ -1B	Qwen-0.5b	Yes	Yes	Yes	22.85	40.83	26.55	2.43
AesthetiQ -2B	InternLM-1.8b	No	No	No	25.18	43.28	26.94	4.94
AesthetiQ -2B	InternLM-1.8b	Yes	No	No	27.09	44.61	28.94	5.76
AesthetiQ -2B	InternLM-1.8b	No	Yes	No	22.18	41.64	24.14	4.48
AesthetiQ -2B	InternLM-1.8b	No	Yes	Yes	27.35	44.81	29.44	5.08
AesthetiQ -2B	InternLM-1.8b	Yes	Yes	No	24.26	43.83	26.44	4.93
AesthetiQ -2B	InternLM-1.8b	Yes	Yes	Yes	28.19	45.92	30.44	6.13
AesthetiQ -4B	Phi3-3.8b	No	No	No	34.59	47.61	33.19	7.46
AesthetiQ -4B	Phi3-3.8b	Yes	No	No	36.62	48.32	35.47	11.29
AesthetiQ -4B	Phi3-3.8b	No	Yes	No	29.97	44.48	31.14	9.72
AesthetiQ -4B	Phi3-3.8b	No	Yes	Yes	35.82	47.65	34.93	11.48
AesthetiQ -4B	Phi3-3.8b	Yes	Yes	No	33.19	46.94	33.42	12.18
AesthetiQ -4B	Phi3-3.8b	Yes	Yes	Yes	38.16	49.27	37.14	14.74
AesthetiQ -8B	InternLM-7b	No	No	No	37.64	51.01	36.32	13.71
AesthetiQ -8B	InternLM-7b	Yes	No	No	40.81	51.82	38.51	16.13
AesthetiQ -8B	InternLM-7b	No	Yes	No	37.43	48.48	34.18	15.44
AesthetiQ -8B	InternLM-7b	No	Yes	Yes	39.26	51.15	38.11	16.20
AesthetiQ -8B	InternLM-7b	Yes	Yes	No	39.18	50.34	36.42	16.37
AesthetiQ -8B	InternLM-7b	Yes	Yes	Yes	42.83	52.67	40.64	17.19

Table 4. Performance of AesthetiQ across scales (1B, 2B, 4B, 8B) on the Crello dataset, evaluating the effects of pretraining, VILA alignment, and data filtering on IoU metrics and judge win rates. The results demonstrate the scalability and effectiveness of the aesthetic-aware preference alignment method.

```
templates. You are presented with
2 visual templates, and your task
is to choose the better template
between these 2 based on the following
criteria:

Aesthetics: How visually appealing is
the template,
Clarity: How clear and easy to
understand is the template,
Usability: How practical and
user-friendly is the template,
Creativity: How unique and innovative
is the design,
Consistency: How consistent is the
template with design principles and
standards.

Please provide your answer in the
following JSON format and do not
include any other details:

{"better_layout": "answer"}

where answer could either be image_1 or
image_2.
```

8. Qualitative results

Due to limited space, we included only a few examples of comparison in the main paper. In the following pages, we show more examples for a more comprehensive comparison.

References

[1] Jian Chen, Ruiyi Zhang, Yufan Zhou, and Changyou Chen. Towards aligned layout generation via diffusion model with aesthetic constraints. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[2] 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 (CVPR)*, pages 14287–14296, 2023. 1

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, pages 27730–27744. Curran Associates, Inc., 2022. 3

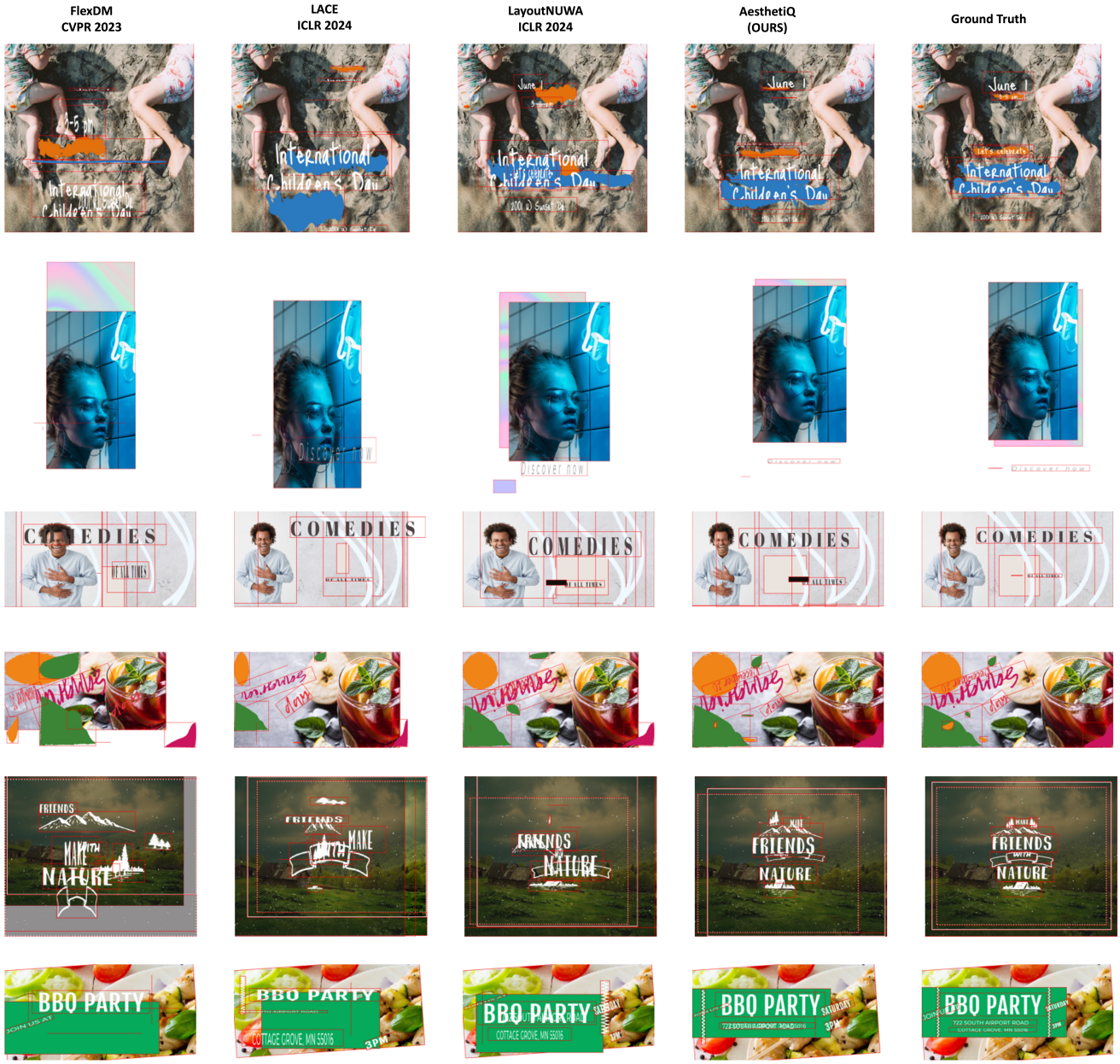

Figure 2. Qualitative comparison of various baselines for layout prediction

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, pages 53728–53741. Curran Associates, Inc., 2023. 3

[5] Zecheng Tang, Chenfei Wu, Juntao Li, and Nan Duan. LayoutNUWA: Revealing the hidden layout expertise of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[6] Haohan Weng, Danqing Huang, Yu Qiao, Zheng Hu, Chin-Yew Lin, Tong Zhang, and C. L. Philip Chen. Desigen: A Pipeline for Controllable Design Template Generation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12721–12732, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2

[7] Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Posterllava: Constructing a unified multi-modal layout generator with llm, 2024. 2

Figure 3. Qualitative comparison of various baselines for layout prediction

Figure 4. Qualitative comparison of various baselines for layout prediction

Figure 5. Qualitative comparison of various baselines for layout prediction

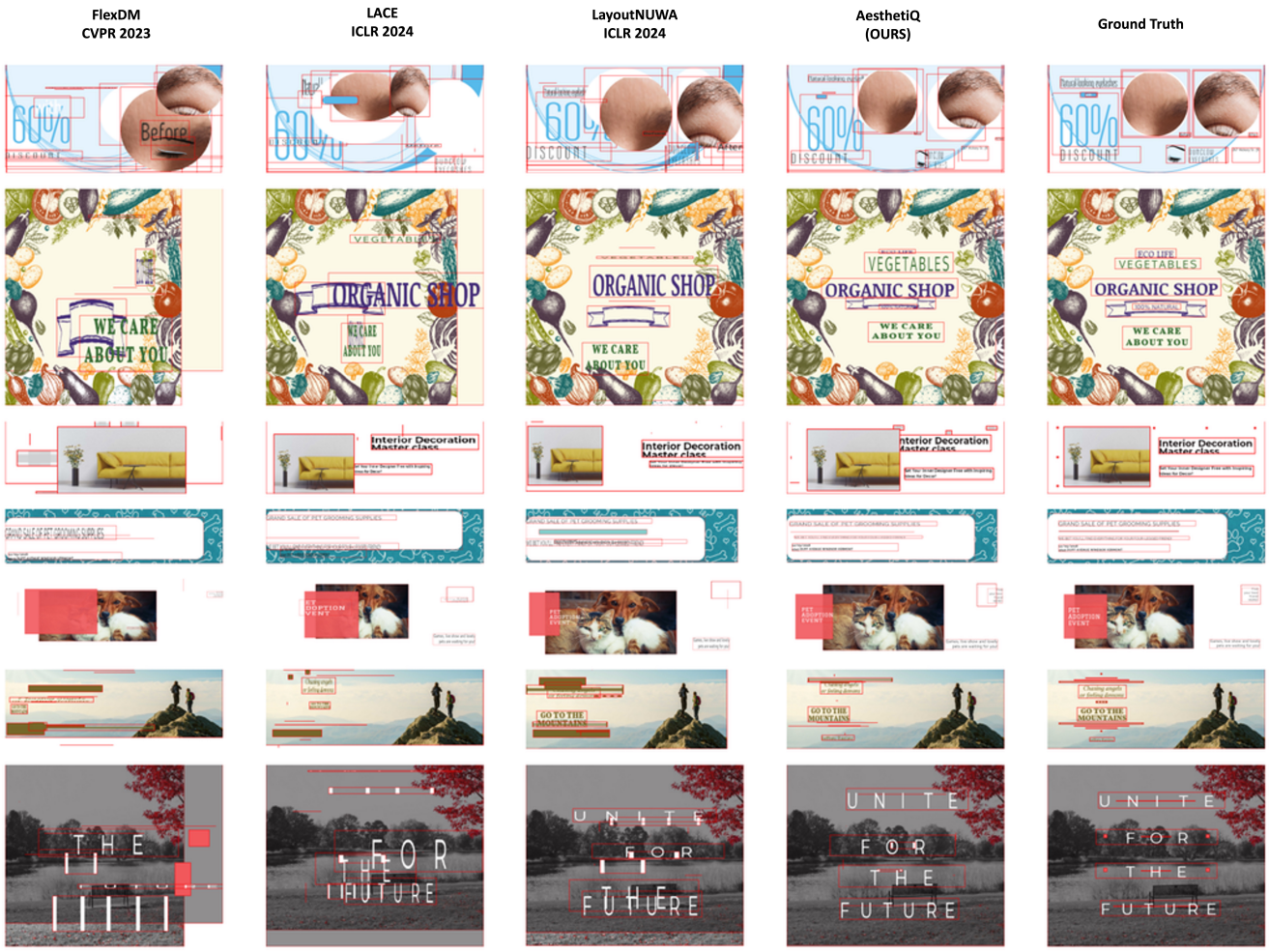

Figure 6. Qualitative comparison of various baselines for layout prediction

Figure 7. Qualitative comparison of various baselines for layout prediction

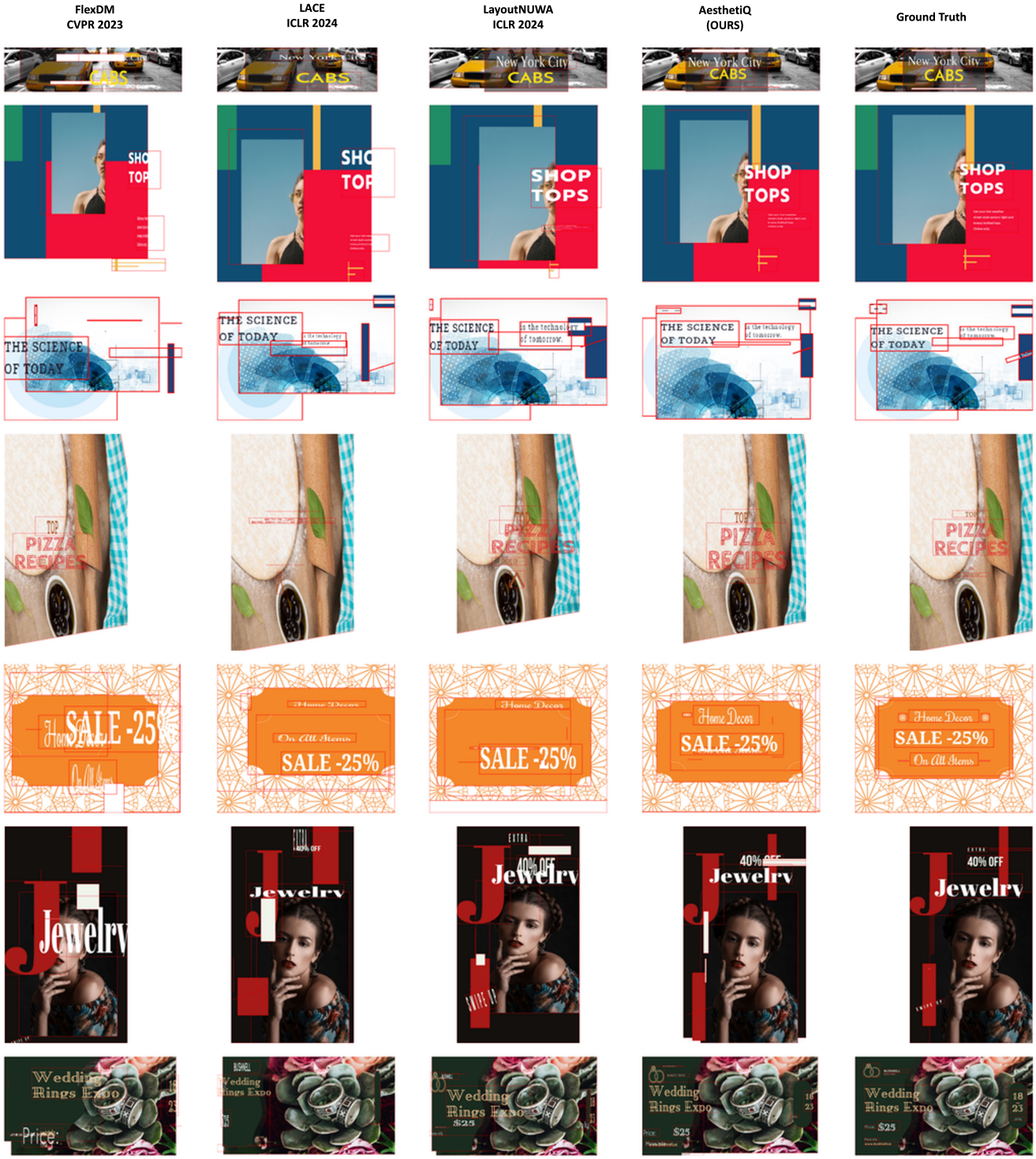

Figure 8. Qualitative comparison of various baselines for layout prediction