IGD: Instructional Graphic Design with Multimodal Layer Generation

Supplementary Material

1. Standardized Format

A design file consists of a series of multimodal layers, which can be categorized as frame, group, text, image and graphic layers. The frame layer serves as the root layer, forming the foundation of the design file and defining basic canvas properties such as size and color. Its specific attributes include the general attributes listed in Tab. 1. Under the frame layer, any of the other four types of layers can be included. The group layer acts as a virtual logical layer, having no impact on the visual presentation. It serves as a logical constraint, grouping layers with similar functions to enhance the logical structure between them and improve interpretability. This organization supports a more intuitive understanding of layer relationships. Like the frame layer, the group layer includes the general attributes in Tab. 1. The text, image, and graphic layers are entity layers, representing the primary elements of the graphic design file. In addition to general attributes, they each have specific attributes unique to their type, as shown in Tabs. 2 to 4, respectively.

Why use Penpot? 1) Penpot demonstrates exceptional flexibility in rendering, incorporating layer transparency, shadow effects, stroke effects, gradient transitions, *etc*. This functional architecture offers developmental potential for future graphic design methods. 2) Penpot is well-suited for large-scale data management. As the predominant open-source design platform, Penpot can integrate cross-playform and cross-format data with our standardized format. This provides fundamental support for the graphic design community, promotes data unification and sharing, and enables scalability from the data perspective. 3) Penpot provides a user-friendly, interactivate editing platform, which is conducive to data cleaning and annotation, as well as practical use.

Attributes	Description
<position> </position>	The position with *x, y, w, h*, representing the coordinates of the upper left corner and the width and height of the layer, respectively.
<color> </color>	RGB values from 0 to 255.
<opacity> </opacity>	Transparency of the layer as a value from 0 to 1.
<transform> </transform>	Six parameters of affine transformation.
<stroke_style> </stroke_style>	Stroke styles, including solid, dotted, dashed, and mixed.
<stroke_width> </stroke_width>	Width of the stroke.
<stroke_align> </stroke_align>	Alignment of the stroke, including center, inside, and outside.
<shadow_style> </shadow_style>	The style of shadows, including drop and inner shadows.
<shadow_offset> </shadow_offset>	The offset parameters of shadow, including *x, y, spread, blur*.

Table 1. General attributes for all multimodal layers.

Attributes	Description
<text_align> </text_align>	Text alignment, including center, left, right, and justify.
<content> </content>	Text content of the layer.
<font_style> </font_style>	Font styles, including normal and italic.
<font_weight> </font_weight>	Thickness of the font.
<font_family> </font_family>	Name of the font.
<font_size> </font_size>	Font size.
<letter_spacing> </letter_spacing>	Character spacing of text.
<line_height> </line_height>	Line height.

Table 2. Specific attributes for text layers.

Attributes	Description
<image_des> </image_des>	Short tag description of the image, including foreground / background, RGB / RGBA, abstract / cartoon / landscape / object / people.
<image_id> </image_id>	Index ID of the image.

Table 3. Specific attributes for image layers.

2. Datasets and Evaluation Metrics

Fig. 1 provides detailed information about a training sample, including its corresponding instructions and standardized format string. We collected approximately 150k vari-

Attributes	Description
<path> </path>	Describe the path elements in SVG.
<move_to> </move_to>	Move the cursor to a specified point *x, y*.
<line_to> </line_to>	Draw a straight line from the current position to the specified point *x, y*.
<curve_to> </curve_to>	Draw a curve to a point *x, y*, controlled by two control points *c1x, c1y, c2x, c2y*.
<close_path/>	Close the path by drawing a straight line back to the starting point.

Table 4. Specific attributes for graphic layers.

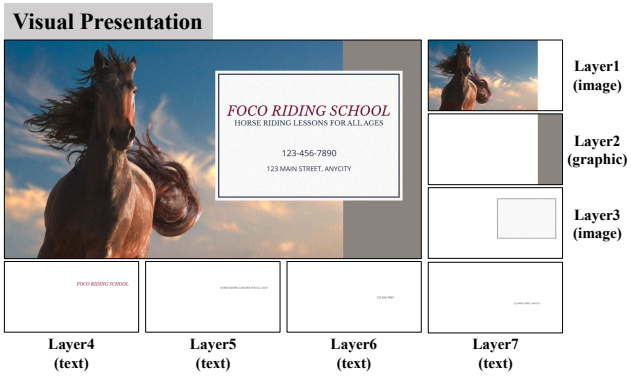

Figure 1. Detailed presentation of a training sample.

ous design files from free data on the Internet and paid data, and after filtering, obtained 90k samples. Specifically, for the images, we filter out QR codes, low-quality images, and images containing excessive text. For the graphic design file, we filter out samples where the tokenized length of the standardized format string exceeded 16k. Empty layers and redundant group layers are simplified or removed. Fig. 5 present some refined design files for training. For each design file, we generated an additional 20 training samples through random augmentation and 5 samples through semantic augmentation.

The detailed prompts for evaluation graphic design images using GPT-4o are shown below. It conducts a comprehensive evaluation from five dimensions: **Quality** assesses image clarity, relevance to the theme, and visual appeal. **Harmony** evaluates the harmonious use of colors to form a cohesive overall design. **Accuracy** determines whether the information conveyed in the graphic design aligns with the instructions and whether text is clearly legible. **Layout** judges the relative positioning of layers for aesthetic appeal. **Innovation** evaluates the creativity of the graphic design. Each dimension contributes to a comprehensive assessment of both visual and communicative quality in graphic design. Fig. 6 shows the feedback of GPT-4o on several examples.

We use an OCR engine[1] to recognize text in design images and evaluate the accuracy of rendered text using character-level precision, recall, and f-measure. Specifically, a character in the OCR recognition result is defined as a True Positive (TP) if it appears in the annotation; otherwise, it is classified as a False Positive (FP). A False Negative (FN) indicates that a character is only present in the annotation but absent from the OCR recognition result. Accordingly, character-level precision, recall, and f-measure can be formulated as follows:

$$Char\_P = \frac{TP}{TP + FP}, \tag{1}$$

$$Char\_R = \frac{TP}{TP + FN}, \tag{2}$$

$$Char\_F = \frac{2 \times Char\_P \times Char\_R}{Char\_P + Char\_R}. \tag{3}$$

3. Implementation Details

We adopt the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 0. The length of the condition queries $\boldsymbol{c}$ is set to 64, and the batch size per GPU is configured as 1. IGD is trained on 32 NVIDIA H800 GPUs. In stage 1 augmented pre-training, we train IGD for 1 epoch using 1.8M randomly augmented data and 450k semantically augmented data. The learning rate was set to 1e-4. During this

[1] https://github.com/PaddlePaddle/PaddleOCR

Evaluation Prompt

You are an autonomous AI Assistant assisting designers by providing insightful, objective, and constructive feedback on graphic design images. Your goal is to offer a thorough and unbiased assessment based on established design principles and industry standards. You will identify potential areas of improvement and provide actionable feedback to enhance the overall performance of graphic design images. Maintain a consistently high standard of criticism throughout the assessment process. You will be provided with a graphic design image, its expected type, and the information it should communicate. Your task is to grade the design objectively based on the following criteria and provide detailed feedback for each category. The design quality will be assessed on a scale of: bad, poor, fair, good, excellent. Comment each rule in relation to this picture at first. If the output is too long, it will be truncated. Respond in JSON format, no other information.
The following criteria should guide your evaluation:
1. Quality: Assess the clarity, relevance to the theme, and overall visual appeal of the image. High-quality designs should elevate the aesthetic of the project and ensure that the image content is consistent with the intended information.
Excellent: The image is clear, appropriate, and enhances the communication of the intended message with no ambiguity.
Bad: The image is distorted, blurry, or unrecognizable and does not align with the intended message.
2. Harmony: Evaluate the use of color in the design, focusing on how cohesive and complementary the colors are to each other. Consider how color choices support the emotional tone and overall impact of the design.
Excellent: Color choices are harmonious and strengthen the message, with complementary colors creating a unified, visually appealing design.
Bad: The colors are jarring or mismatched, creating a disorganized and uncomfortable visual experience.
3. Accuracy: Review the accuracy of the information conveyed, particularly focusing on text clarity, correctness, and relevance, including both English and Chinese.
Excellent: The design accurately delivers the intended message with no spelling or grammatical errors or fabricated characters, engaging the target audience effectively.
Bad: The text contains significant errors or fabricated characters, making it difficult to understand or recognize. The content is irrelevant to the intended message.
4. Layout: Evaluate the organization and interaction of elements within the design, including how well text, images, and graphics are arranged.
Excellent: The layout is clear and logically organized, guiding the viewer's attention and enhancing readability through a balanced and innovative arrangement.
Bad: The layout is chaotic, lacking clear structure or visual flow, making the design difficult to navigate or comprehend.
5. Innovation: Assess the originality and creativity of the design, considering how it stands out visually and conceptually.
Excellent: The design demonstrates creativity and originality, offering a fresh visual experience and resonating with current trends and the audience's interests.
Bad: The design is uninspired and lacks any notable creative elements, presenting a bland and forgettable visual experience.

Method	Quality ↑	Harmony ↑	Accuracy ↑	Layout ↑	Innovation ↑	Char-P ↑	Char-R ↑	Char-F ↑	R_{ali} ↓	R_{ove} ↓	R_{com} ↓	FID ↓
w/o augmentation	80.4	81.2	80.2	62.0	67.6	93.96	80.33	86.61	0.0656	0.0290	25.05	63.73
LLM - Vicuna	72.4	83.6	75.0	61.8	62.6	90.31	76.50	82.83	0.0775	0.0389	31.24	68.14
Freezing SD	77.8	81.6	78.2	67.2	64.2	92.85	**83.53**	**87.94**	**0.0509**	**0.0218**	29.26	59.35
IGD	**81.2**	**88.8**	**81.4**	**72.0**	**68.2**	**94.35**	80.41	86.83	0.0541	0.0245	**21.53**	**56.35**

Table 5. Performance of IGD with different configurations.

stage, only the LLM $\mathcal{F}$ and the image understanding projection layer $\mathcal{L}_{com}$ are trained. In stage 2 end-to-end fine-tuning, we use 90k refined source data training IGD for 7 epochs, with a learning rate of 4e-5. In this stage, we conduct end-to-end synergistic training of the condition queries $\boldsymbol{c}$, the SD $\mathcal{H}$, the LLM $\mathcal{F}$, and both projection layers $\mathcal{L}_{com}$ and $\mathcal{L}_{gen}$.

4. Experiments

4.1. Ablation Study

Data Augmentation During the training process, random augmentation and semantic augmentation strategies are employed to effectively expand the dataset while preserving the diversity of text layers. In stage 1, these augmented data are used to enhance the capability in layout design and multimodal layer attribute prediction, aiming to achieve harmonious visual effects. However, random augmentation data lack contextual semantic coherence, and although semantic augmentation data maintain textual consistency within individual design files, they may exhibit semantic discrepancies with other image layers. Therefore, we adopt the multimodal input mode in stage 1 to enable the model to preliminarily complete layout designs based on text length or fundamental semantic associations. As shown in Tab. 5, omitting the pre-training process in stage 1 leads to a decline in the layout design capability and harmonization of the design image. This is primarily attributed to the difficulty of directly training layout design and image generation capabilities in stage 2, as well as the lack of diverse data, which limits the model's generalization ability. In contrast, incorporating stage 1 training allows the model to achieve preliminary layout coordination based on text length and basic semantics. Building on this foundation, stage 2 training fur-

Figure 2. The sample on the left is generated by IGD, and the sample on the right is one that has not been trained with the conversion of simple images into SVG.

ther strengthens the model's refined layout design capability and image generation performance, thereby significantly improving overall performance.

Image Processing In graphic design files, images predominantly composed of solid colors or simple abstract patterns are frequently encountered. The repetitive distribution of such homogeneous visual elements tends to bias the model during training towards generating simple color schemes and pattern structures. Furthermore, employing image generation methodologies for producing these elementary graphics increases model complexity and impairs the training efficiency. Therefore, we represent simple images using SVG and achieve their generation through parametric prediction. This approach reduces training complexity by slightly sacrifices unimportant visual details, while helping the model focus on learning the primary content objectives. As shown in Fig. 2, models trained without converting simple images into SVG tend to generate images without clear semantic meaning, and the main graphical elements display flattened representations lacking hierarchical depth.

Method	Char-P ↑	Char-R ↑	Char-F ↑	R_{ali} ↓	R_{ove} ↓	R_{com} ↓	FID ↓
SD3.5	39.77	28.10	32.93	-	-	-	114.36
DALL·E3	17.12	19.66	18.30	-	-	-	137.70
AnyText	65.82	37.27	47.59	-	-	-	140.14
OpenCOLE	8.85	1.64	2.77	0.0795	0.0439	25.09	163.37
IGD	**95.83**	**84.64**	**89.89**	**0.0186**	**0.0344**	**25.42**	**54.12**
OpenCOLE†	47.41	64.06	54.49	-	-	-	-
IGD†	**99.56**	**99.11**	**99.33**	-	-	-	-

Table 6. More metrics on with Chinese user inputs. † indicates the calculation of text accuracy using predicted text layer raw typographic information, i.e. before rendering into an image.

Influence of LLM Our proposed method relies on the LLM to predict multimodal layer attributes and arrange the relative positional layout of inter-layer components. As shown in Tab. 5, we replace the LLM from Qwen2.5 7B with Vicuna 7B. Experimental results indicate that the capability of LLM is crucial for graphic design tasks. IGD based on Qwen2.5 outperforms Vicuna-based model, especially in layout arrangement. The decline in layout capabilities leads to a deterioration in the overall aesthetics of the design images and the potential layer occlusion that impair the legibility of text, collectively manifesting in diminished text accyracy metrics.

End-to-End Training We adopt an end-to-end training process combining LLM and SD to enhance the overall harmony of graphic design images. In this training paradigm, the model leverages existing multimodal layer information as context when generating new layers, ensuring better spatial relationships and stylistic consistency. In contrast, relying solely on natural language descriptions as input to SD preserves semantic accuracy but struggles to capture fine-grained visual details and structural coherence across layers. This limitation makes it difficult to maintain visual consistency between different elements. As shown in Fig. 3, the generated text descriptions exhibit noticeable disharmony when presented as an integrated design. The descriptions for images in Fig. 3(b) are: *"In this image, a blue - toned sky stretches overhead. A vast golden wheat field spreads out beneath, its wheat moving softly in the wind, and a few white clouds floating across the sky project soft shadows onto the wheat."*, *"The image depicts a blank, white rectangular frame with a golden border."*, *"The image shows a person wearing a hat, sitting on a grassy field, holding a bunch of wheat ears, with a smile on their face, and there are trees in the background."*. Additionally, we freeze the SD parameters during the end-to-end training. Experimental results in Tab. 5 indicate that freezing SD has only a slight impact on overall performance. For the graphic metrics R_{ali} and R_{ove}, it results in marginally superior performance, which demonstrates an advantage in layer layout composition.

Language	Method	Quality ↑	Harmony ↑	Accuracy ↑	Layout ↑	Innovation ↑	Char-P ↑	Char-R ↑	Char-F ↑	R_{ali} ↓	R_{ove} ↓	R_{com} ↓	FID ↓
Chinese	OpenCOLE	55.4	75.0	35.6	52.8	52.8	31.40	20.00	24.43	0.1140	0.0534	30.72	115.42
	IGD	83.4	86.8	91.0	75.4	65.8	96.02	86.85	91.20	0.0745	0.0341	19.12	29.67
	IGD†	-	-	-	-	-	99.67	99.36	99.51	-	-	-	-
English	OpenCOLE	71.0	82.2	64.0	61.4	63.2	65.07	55.00	59.61	0.2063	0.0389	33.84	73.35
	IGD	81.8	85.6	82.0	73.2	68.4	93.85	78.80	85.67	0.0671	0.0273	22.81	31.16
	IGD†	-	-	-	-	-	98.23	96.65	97.43	-	-	-	-

Table 7. Comparison in image input mode. For OpenCOLE, since it only supports single image input, we only input the first image and discard subsequent images. † indicates the calculation of text accuracy using predicted text layer raw typographic information, i.e. before rendering into an image.

Figure 3. The generated samples by (a) IGD; (b) LLM generates the natural language description of images, and an additional SD is used to generate images.

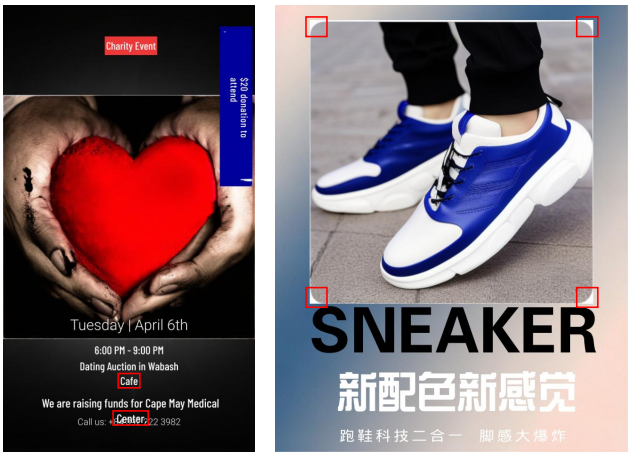

Figure 4. Bad cases. The case on the left shows that there is a problem with hanging individual characters or words. The case on the right demonstrates that IGD is unable to produce a transparent image.

4.2. Comparison with Existing Methods

Tab. 6 presents various metrics on the Chinese benchmark. Methods other than AnyText do not support Chinese, so the text accuracy of Chinese characters is relatively low. The failure of OpenCOLE to successfully render Chinese results in either the absence of any Chinese character in the image or the appearance of placeholder squares, which leads to its low text accuracy. IGD supports both Chinese and English and demonstrates strong performance in both text accuracy and graphic metrics.

Tab. 7 shows the model performance when taking images as input as well. Due to structural limitations of the model framework of OpenCOLE, it only supports the input of one background image, significantly restricting its practical applicability. IGD outperforms OpenCOLE in both GPT evaluations and text accuracy assessments. Additionally, as the images from the benchmark are used directly as input, the rendered results distribution remains close to the benchmark, resulting in low FID scores.

More samples generated by IGD in single modal input mode and multi-modal input mode are shown in Fig. 7 and Fig. 8.

5. Limitations

As shown Fig. 4, the graphic design images generated by IGD exhibit unexpected hanging issues with individual characters or words. This problem primarily arises from a discrepancy between the space occupied by the rendered

text layers (based on predicted attributes such as font, font size, and letter spacing) and the width of the predicted text layers, leading to suboptimal visual text presentation. We attribute this issue to the insufficient scale and diversity of the training data. This problem is more likely to occur in paragraph text, as this type of data constitutes a relatively small proportion of the overall dataset, resulting in limited robustness of the model. With further scaling up the training data and enriching its diversity, the model's performance in this regard is expected to improve significantly. Moreover, IGD currently cannot generate transparent images with RGBA channels. This limitation stems from the use of the Stable Diffusion model, which only supports the generation of images with RGB color channels. To overcome this constraint, the model's ability for transparent image generation can be improved in the future by replacing the image generation module with a more advanced method that supports RGBA channels.

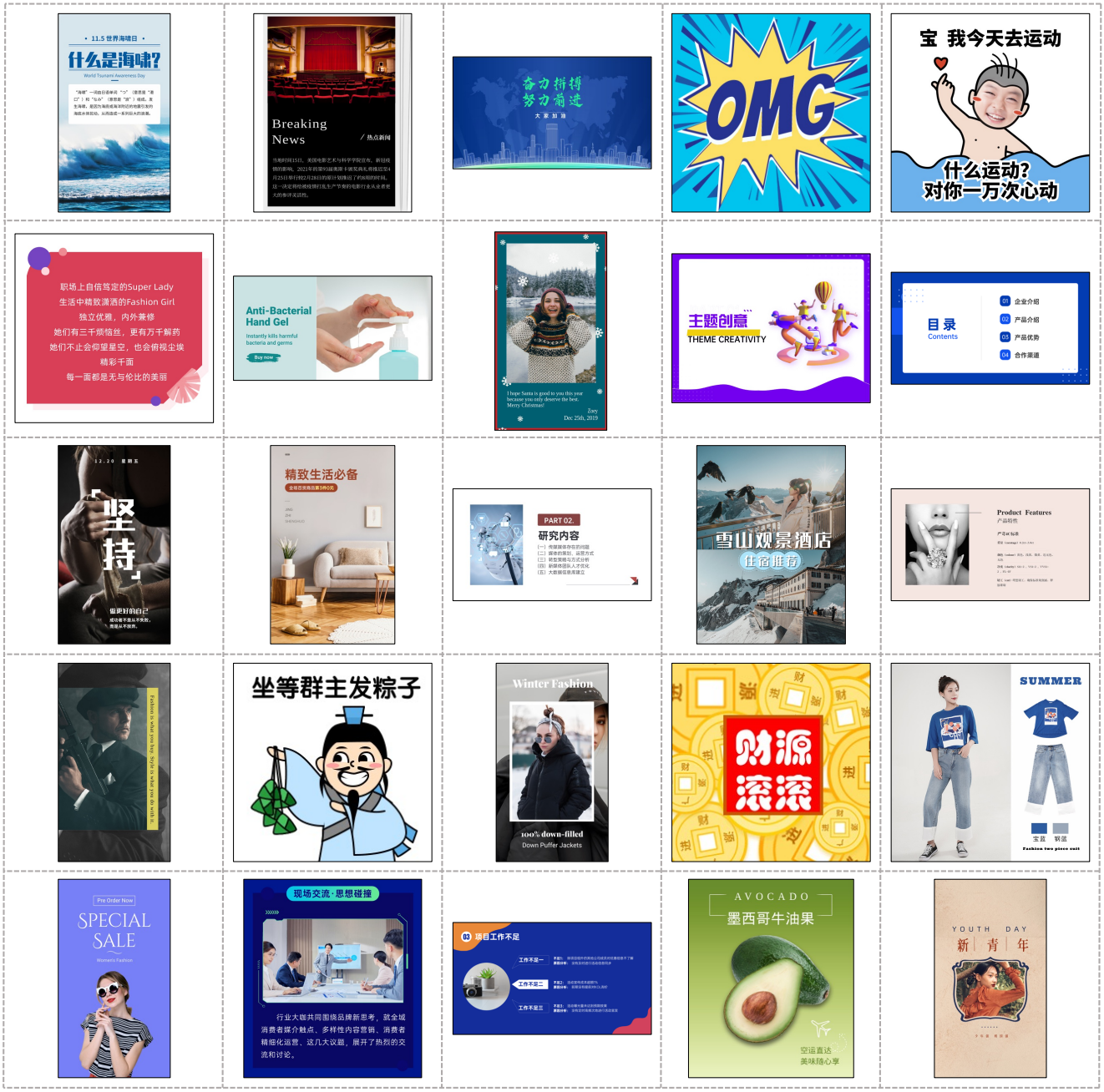

Figure 5. More samples in refined source dataset.

IGD

GPT Feedback


```
{
 "Quality": {
  "grade": "good",
  "feedback": "The image is visually clear, with the main elements – the text, fireworks, and lantern – standing out well against the background. The theme of celebrating the New Year is conveyed effectively, although the clarity of the text could be improved in certain areas due to the color contrast between the text and background. Enhancing the text readability would increase overall impact."
 },
 "Harmony": {
  "grade": "good",
  "feedback": "The colors are generally harmonious and supportive of a celebratory, warm atmosphere. The red and yellow of the lanterns contrast well with the dark sky, enhancing the festive tone. However, slight adjustments in the brightness or contrast might improve text visibility."
 },
 "Accuracy": {
  "grade": "excellent",
  "feedback": "The text accurately conveys the intended information, including '南来北往,' '奔向生活幸福的方向,' and 'Happy New Year.' All text is free of errors and aligns well with the theme of the event."
 },
 "Layout": {
  "grade": "excellent",
  "feedback": "The layout is well-organized, with the lantern drawing the viewer's attention downward to the main text and then to the 'Happy New Year' message. This flow supports the reading order and guides the viewer's eye naturally through the design. The spacing between elements is balanced, providing a clear and structured composition."
 },
 "Innovation": {
  "grade": "good",
  "feedback": "The design shows creativity in its use of traditional symbols like the lanterns and fireworks, which are well-integrated to convey the New Year theme. While the design is visually appealing, additional unique or modern elements could enhance its memorability."
 }
}
```


OpenCOLE

GPT Feedback


```
{
 "Quality": {
  "grade": "good",
  "feedback": "The image is visually appealing and vibrant, with a clear focus on the festive theme. However, the text 'Happy New Year' is slightly cluttered against the busy background, which reduces overall clarity. Additionally, the year '2015' is outdated and may confuse viewers if the design is intended for a current New Year celebration."
 },
 "Harmony": {
  "grade": "excellent",
  "feedback": "The colors are harmonious and vibrant, effectively conveying a festive and celebratory tone. The combination of warm reds, yellows, and cool blues creates a balanced and visually appealing design. The color choices enhance the energy and mood of the piece."
 },
 "Accuracy": {
  "grade": "fair",
  "feedback": "The English text 'Happy New Year' is accurate, but the Chinese text appears to have placeholder or illegible characters, which detracts from the intended message. This makes it difficult for Chinese-speaking audiences to understand the design's full meaning."
 },
 "Layout": {
  "grade": "fair",
  "feedback": "While the central elements are visually engaging, the layout feels overcrowded, with text and decorative elements competing for attention. The lack of clear visual hierarchy makes it harder to guide the viewer's eye through the design effectively."
 },
 "Innovation": {
  "grade": "good",
  "feedback": "The design is creative and visually rich, with intricate patterns and festive motifs that stand out. However, it lacks a truly unique or innovative element that would make it memorable compared to other New Year-themed designs."
 }
}
```


DALL·E3

GPT Feedback


```
{
 "Quality": {
  "grade": "good",
  "feedback": "The design is visually appealing and well-crafted, with clear central elements that convey a festive, celebratory mood. However, the absence or lack of emphasis on the messages '南来北往' and '奔向生活幸福的方向' detracts from fully meeting the intended information. Overall, the visual quality is high, but there is room for improvement in information clarity and completeness."
 },
 "Harmony": {
  "grade": "excellent",
  "feedback": "The color scheme is harmonious, using traditional red and gold hues that enhance the festive New Year theme. These colors create a unified and visually appealing design that aligns well with Chinese cultural aesthetics, contributing to a warm, joyful atmosphere."
 },
 "Accuracy": {
  "grade": "fair",
  "feedback": "The design includes 'Happy New Year' and some decorative elements associated with the New Year, but it lacks clear representation of the phrases '南来北往' and '奔向生活幸福的方向.' This reduces the accuracy and completeness of the intended message, as two key phrases are missing or underemphasized, impacting the overall communication."
 },
 "Layout": {
  "grade": "good",
  "feedback": "The layout is symmetrical and balanced, with a clear central focus that enhances readability for the main message. However, the design is visually dense, and the absence of a clear layout area for '南来北往' and '奔向生活幸福的方向' means the intended information is not fully communicated. The arrangement could be improved by including these phrases in a visible way."
 },
 "Innovation": {
  "grade": "excellent",
  "feedback": "The design demonstrates a high level of creativity, with intricate patterns, traditional Chinese elements, and a fresh interpretation of New Year visuals. This innovative approach adds to the design's appeal, even though it could be further enhanced by incorporating all the required elements."
 }
}
```


AnyText

GPT Feedback


```
{
 "Quality": {
  "grade": "good",
  "feedback": "The image quality is clear and visually appealing, with well-defined elements such as the lanterns and gold decorations that create a festive atmosphere fitting for a New Year event. However, the Chinese characters appear somewhat stylistically exaggerated, which may slightly detract from readability."
 },
 "Harmony": {
  "grade": "excellent",
  "feedback": "The red and gold color scheme is harmonious and strongly aligned with traditional New Year themes, enhancing the celebratory tone of the design. The colors work well together to create a luxurious and festive feel, which is ideal for the occasion."
 },
 "Accuracy": {
  "grade": "poor",
  "feedback": "The image does not accurately convey the expected message. The Chinese characters on display do not appear to be '南来北往' or '奔向生活幸福的方向,' which may cause confusion. Additionally, 'Happy New Year – [Canton(?)]' at the bottom has unclear meaning, possibly due to translation or formatting issues."
 },
 "Layout": {
  "grade": "good",
  "feedback": "The layout is well-organized, with a clear structure that draws attention to the primary message at the center and secondary details below. The use of space around the main text and decorative elements is balanced, creating a visually pleasing composition."
 },
 "Innovation": {
  "grade": "fair",
  "feedback": "The design relies on traditional New Year motifs like lanterns and a red-and-gold color scheme. While effective, it lacks unique or creative elements that could make it more memorable or stand out among similar designs."
 }
}
```


Figure 6. Detailed comments by GPT4 on the results generated by the different models.

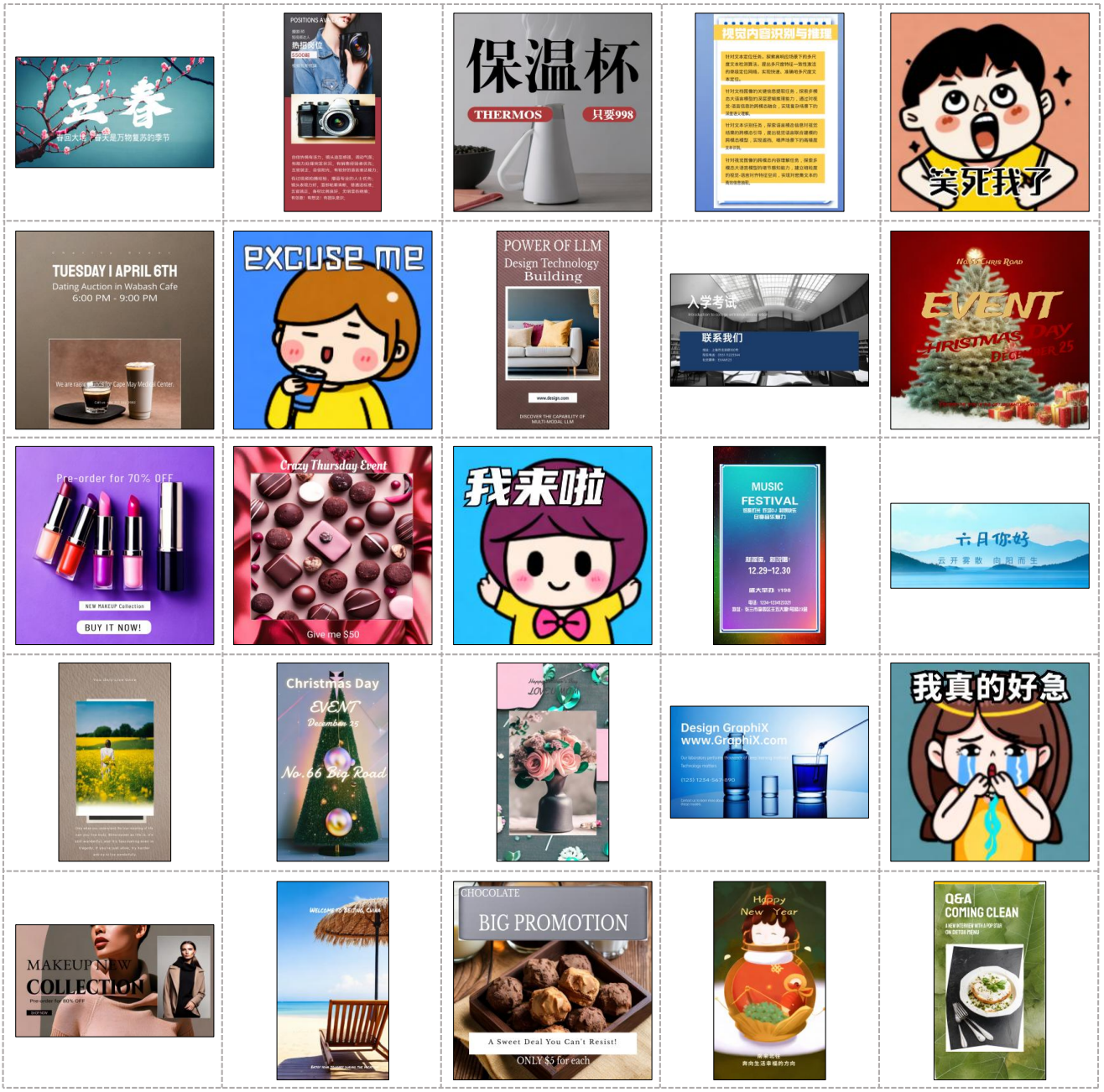

Figure 7. More samples generated by IGD in single modal input mode.

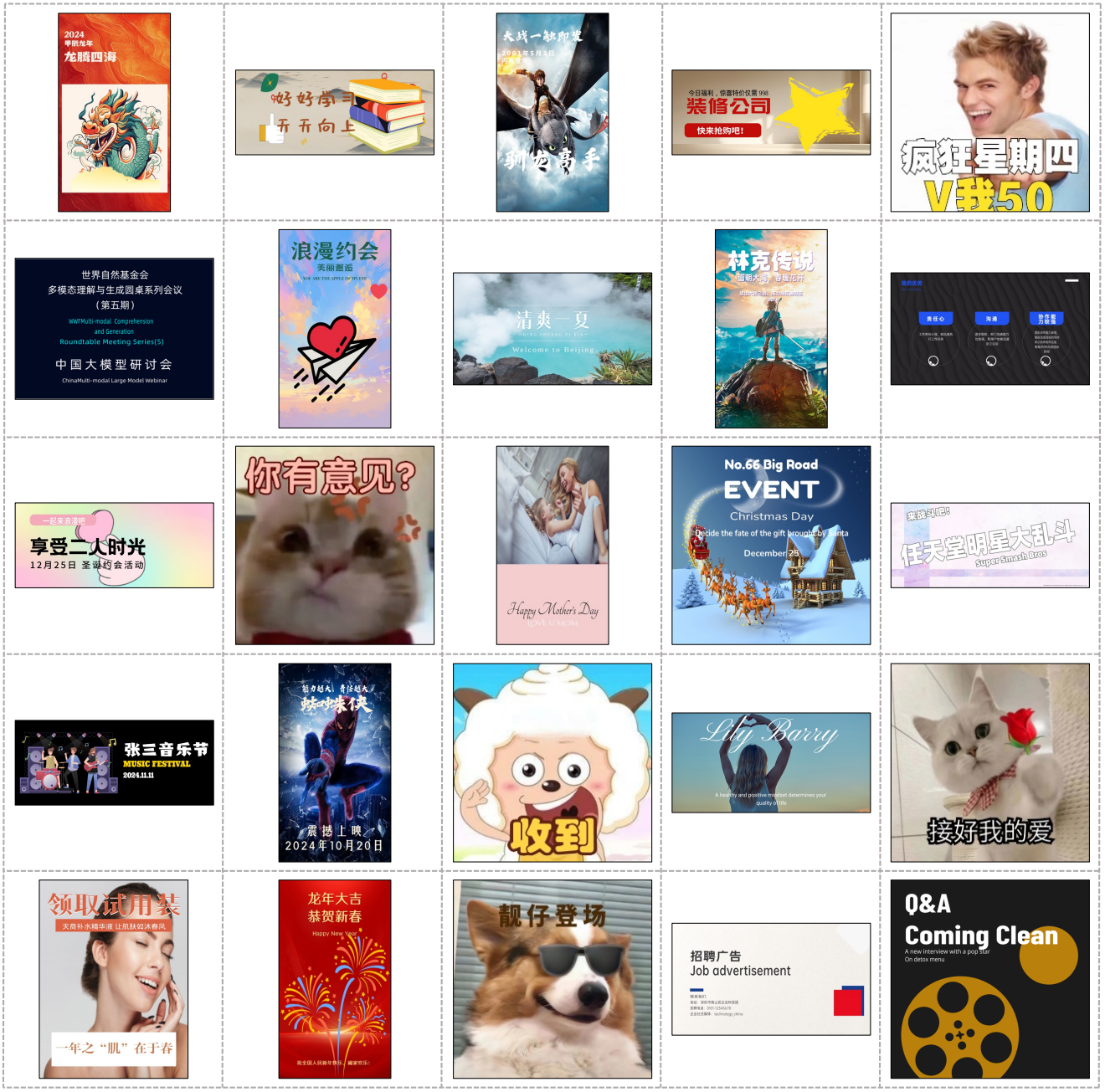

Figure 8. More samples generated by IGD in multi-modal input mode.