MRT: Masked Region Transformer for Layered Image Generation and Editing at Scale

Supplementary Material

Training Caption Length	FID_{merged} ↓	
	Short Cap.	Long Cap.
Short Cap.	17.64	18.56
Long Cap.	17.95	16.15
Mixed (50% short + 50% long)	16.13	15.93

Table 1. Effect of caption length during training. We train models with short captions, long captions, or a mixture of both, and evaluate FID on VC5K test set using short and long captions respectively.

method	Denoise steps	Latency(s)	Speed up	FID_{merged} ↓
Baseline	50	14.4	-	16.02
+ Distill	16	4.5	3.2x	16.21
+ Distill	8	2.3	6.26x	18.58

Table 2. Multi-layer generator distillation with inference time.

#layer numbers	2∼7	8∼11	12∼14	15∼50
$PSNR_{merged}$ ↑	22.51	21.99	21.36	20.65
$SSIM_{merged}$ ↑	0.8932	0.8869	0.8780	0.8610

Table 3. Effect of layer numbers on image-to-layers (I2L) generation quality. We evaluate the model's performance across different ranges of layer numbers in the generated results.

1. Additional ablation experiments

1.1. Mixed Training with Variable Caption Length

Table 1 demonstrates the importance of caption diversity during training. Models trained with mixed caption lengths achieve the best generalization, with FID of 16.13 on short captions and 15.93 on long captions. Training exclusively on one caption type creates a domain gap: short-caption-only training degrades to 18.56 FID on long captions, while long-caption-only training achieves 16.15 FID, showing better robustness but still suboptimal on short captions.

1.2. Effect of Layer Numbers on Image-to-layer

Table 3 demonstrates our method's scalability across different layer counts for the image-to-layers task. The model handles compositions ranging from 2 to 50 layers effectively, maintaining stable performance across this wide range. This flexibility enables decomposition of both simple designs and complex multi-element compositions without architectural modifications.

1.3. Analysis of Distilled Models

To evaluate the real-world efficiency of our approach, we conducted inference speed benchmarks on a single NVIDIA H200 GPU. We compared the standard baseline method (operating at 50 denoising steps) against our distilled MRT model at reduced inference steps (16 and 8 steps). As shown in Table 2, the baseline model requires 14.4 seconds to complete the generation process. In contrast, applying DMD2 distillation significantly accelerates inference. Specifically, our model achieves a $3.2\times$ speed-up (4.5s) at 16 steps with negligible degradation in generation quality (FID increases only slightly from 16.02 to 16.21). Furthermore, reducing the inference budget to just 8 steps yields a massive $6.26\times$ speed-up (2.3s), showing that our method successfully balances high-fidelity generation with interactive-level latency. We also present the generated samples and compare the original and distilled models in Fig. 1.

2. Attention Analysis of Image-to-Layer Model

To validate that our model learns meaningful semantic representations rather than merely memorizing layout priors, we visualize the pixel-wise attention maps generated during the decomposition process. Fig. 3 illustrates the correspondence between the generated transparent layers and their associated attention activations. As observed, the attention mechanism exhibits strong spatial localization capabilities. For each predicted layer, the attention weights (visualized as heatmaps) highly correlate with the semantic boundaries of the target elements. For instance, when reconstructing high-frequency components such as text (*e.g.*, "Bundle of Joy" in the second case, "Love NEVER FELT..." in the third one) or fine-grained graphical elements, the attention is tightly focused on the relevant character strokes and shapes, effectively suppressing background noise. Conversely, for background patterns or larger geometric shapes, the attention acts more broadly to capture the texture and spatial extent of the region. This visualization confirms that the model successfully disentangles the composite image by attending to distinct visual features guided by the layout, ensuring that the resulting RGBA layers possess clean alpha mattes and coherent textures.

3. Efficiency Analysis of Image-to-Layer Model

Latency. As shown in Fig. 2, due to our regional diffusion transformer architecture, we achieve significant speedup and over $100\times$ acceleration compared to Qwen-Image-Layered, which uses the same number of full-resolution tokens to model each transparent layer regardless of the actual area of contents. We achieve similar latency speed-up as the statistics shown in ART [38] and we further applied various advanced cache techniques, model distillation, lower-precision, parallel inference to optimize the latency of our

Figure 1. **Generation quality of distilled models.** We achieve up to 6x speed up without sacrificing the quality and fidelity of images.

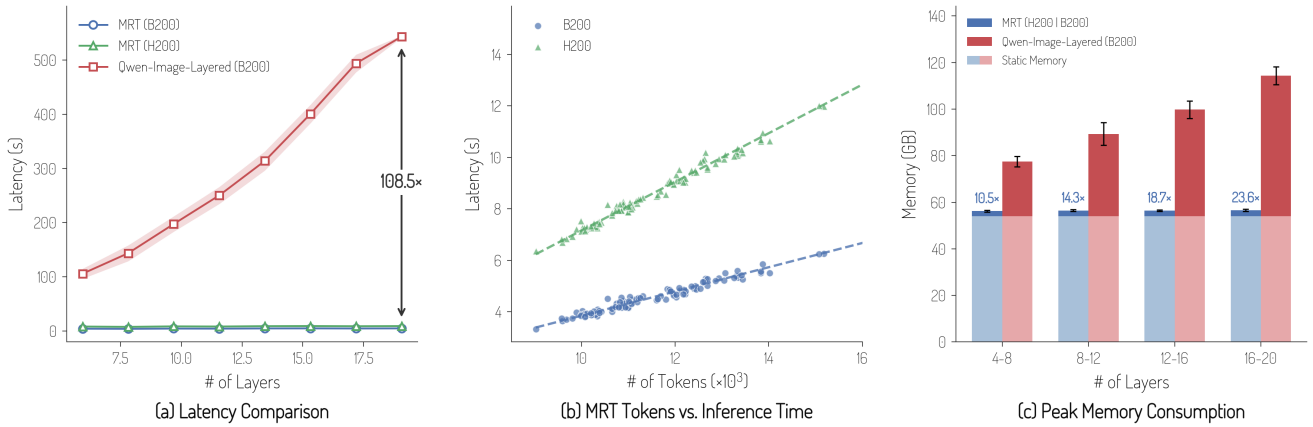


Figure 2. **Inference efficiency comparison between MRT and Qwen-Image-Layered.** (a) Latency scaling with number of layers. MRT maintains near-constant latency (∼5s) while Qwen-Image-Layered scales linearly, resulting in up to 108.5× speedup at ∼20 layers. (b) MRT inference time vs. token count on H200 and B200 GPUs, demonstrating linear scaling behavior. (c) Peak GPU memory consumption across varying layer configurations. The shaded region indicates the baseline memory allocated to model weights. MRT reduces memory consumption by $10.5\times \rightarrow 23.6\times$, with efficiency gains scaling proportionally with layer numbers. *All reported results are conducted over 100 samples on single GPU with identical layer numbers.*

model to within $\sim$ 3 seconds when running with $4\times$ H100 GPUs and $\sim$ 5 seconds on a single H100 GPU when required to decompose a single 1K high-resolution image into nearly 20 transparent layers.

Memory. Unlike Qwen-Image-Layered, which requires significantly more visual tokens to extract more layers from a flat image, our approach is far more memory efficient, requiring far fewer tokens to decompose an image into many transparent layers as shown in Fig. 2. Our method achieves clear advantage in memory usage, which is further strengthened as the number of layer grows.

4. User study details

4.1. User Study on Text-to-Layer Task

To evaluate the generation quality of our models on the `text-to-layer` task, we conducted a user study comparing our method (MRT) with the baseline (ART). We employed a blind, pairwise comparison setup. For each sample, participants were first shown the input text prompt, followed by the corresponding results generated by MRT and ART displayed side-by-side. To eliminate positional bias, the display order (left or right) of these two results was randomized for each evaluation. Participants were asked to cast a three-way forced-choice vote—"Method A is better," "Method B is better," or "Tie"—across four distinct dimensions: (1) elements (layout), (2) visual appeal (aesthetics), (3) correctness of the text (typography), and (4) coherence and quality of each layer (harmonization).

The web-based evaluation interface is shown in Fig. 39, where two generated results are displayed side-by-side with the text caption provided on the right panel.

Figure 3. **Attention map visualizations of image-to-layers task.** We demonstrate the interpretability of our model by visualizing the internal attention weights during the layer generation process. Left: The input composite image and its corresponding layout. Right: The decomposition results. The top row displays the predicted transparent layers, while the bottom row shows the corresponding attention maps overlaid on the input image. Red regions indicate high activation values. The results highlight that the model's attention is semantically selective, accurately aligning with specific visual elements (*e.g.*, text, foreground objects, background patterns) to generate high-quality, disentangled layers.

4.2. User Study on Image-to-Layer Task

For the `image-to-layers` task, we conducted a comprehensive user study by performing three separate pairwise comparisons between our method and three state-of-the-art baselines: (1) Ours vs. LayerD, (2) Ours vs. Lovart, and (3) Ours vs. Roboneo. Each comparison was run as an independent blind test. Participants in each study were presented with a three-image layout: the original input image was displayed as a central reference, while our method's result and the corresponding baseline's result were shown side-by-side. To eliminate positional bias, the display order (left or right) of our result and the baseline's result was fully randomized in every trial. Participants were asked to make a three-way forced-choice vote ("Method A is better," "Method B is better," or "Tie") based on three key metrics: (1) granularity, (2) layer integrity, and (3) layer quality.

The evaluation interface is illustrated in Fig. 40, where the reference input image is shown at the center with decomposition results from two methods displayed on both sides.

5. Limitations

Although our model demonstrates strong performance in the image-to-layer task, it faces challenges when applied to real-world photographs. Specifically, our method often fails to correctly handle shadows, resulting in segmented object layers that exclude shadow regions and leaving the shadows on the background layer, which leads to visual inconsistency. We attribute this limitation primarily to our training data: our model was trained exclusively on design datasets, which are planar and lack physical effects such as shadows, reflections, and refractions that commonly appear in natural scenes. Despite this domain gap, we were pleasantly surprised to find that our method can still generalize reasonably well to real images, even without any supervision on real-world multi-layer data. As shown in our illustrations, most objects are successfully separated, which we believe stems from the strong visual understanding capability inherited from the Qwen-Image backbone, demonstrating the robustness, adaptability, and scalability of our approach. In future work, we plan to extend our method to real-world image scenarios by collecting and training on datasets that include realistic visual effects such as shadows and reflections. We believe such extensions will further enhance the model's ability to produce coherent and physically plausible layer decompositions.

6. Visualizations and Qualitative Analysis

6.1. Diverse Text-to-Layer Generation

We visualize the qualitative results of our Text-to-Layer task in Fig. 4 through Fig. 12. Our Masked Region Transformer demonstrates exceptional versatility in generating high-fidelity multi-layer designs solely from textual descriptions. As shown in Fig. 4 through Fig. 10, the model successfully synthesizes coherent compositions ranging from simple layouts to complex designs with over 25 layers and even more, maintaining strict spatial consistency and stylistic harmony. A key advantage of our approach is the native support for diverse typography; Fig. 11 illustrates our unique overflow generation capability. Unlike prior methods that truncate content at the canvas edge, our model generates complete, full-size RGBA layers that extend beyond the visible background boundary, thereby preserving full editability and reusability for downstream compositional tasks. Furthermore, Fig. 12 highlights the model's capability to render accurate visual text across multiple languages, including Chinese, ensuring practical utility for global design applications.

6.2. Comparative Analysis of Image-to-Layer

In Fig. 13 through Fig. 20, we provide a comprehensive qualitative comparison between our approach and state-of-the-art baselines, including LayerD, Lovart, and RoboNeo. The results consistently demonstrate that our method establishes a new standard for layer decomposition quality. While commercial systems like RoboNeo often introduce visual artifacts or fail to produce clean transparency, and academic baselines like LayerD tend to produce overly grouped layers that limit editing flexibility, our Masked Region Transformer achieves a superior balance. Our method excels in generating precise alpha mattes, maintaining semantic integrity, and achieving appropriate decomposition granularity (e.g., separating distinct visual elements rather than merging them). This is particularly evident in complex overlapping regions, where our model successfully disentangles elements that other methods fail to separate.

6.3. Scalability on Layer Counts in Image-to-Layer

To evaluate the robustness of our framework, we visualize image-to-layers decomposition results across varying degrees of complexity in Fig. 21 through Fig. 25, ranging from 6 layers up to 16 layers. These visualizations confirm that our architecture scales effectively without performance degradation. The model maintains consistent quality in boundary detection and content preservation in cases of a wide range of layer counts. This stability across diverse layer counts validates the efficacy of our masked attention mechanism, proving that the model can handle the structural complexity of professional-grade graphic designs.

6.4. Generalization to AI-generated Images in Image-to-Layer

To validate whether our model can generalize to AI-generated images, we visualize more image-to-layers de-

composition and compare with Qwen-Image-Layered in Fig. 26 through Fig. 29. As shown, our method produces more coherent and well-separated layers, with clearer foreground-background disentanglement and sharper transparency boundaries. In contrast, Qwen-Image-Layered tends to mix unrelated contents into the same layer or repeat one object across multiple layers. This indicates that our model is more robust to the distribution shift of AI-generated data. Overall, the results demonstrate strong generalization ability and reliable layer decomposition on both design style and natural images.

6.5. Context-Aware Layer Addition

Fig. 30 demonstrates the capabilities of our layers-to-layers task, specifically focusing on layer addition. Here, we simulate a user editing workflow where new elements—such as text or decorative objects—are inserted into an existing design based on text prompts and specified bounding boxes. The results show that our model does not merely paste isolated objects; instead, it generates new layers that are contextually aware, matching the lighting, perspective, and artistic style of the existing layers. By conditioning on the full composition, the added layers harmonize seamlessly with the original design, preserving the global aesthetic while fulfilling the user's semantic requirements.

6.6. Layer Restylization and Harmonization

In Fig. 31, we showcase the layer restylization capability, where user-provided assets are transformed to align with a target design's visual identity. Our model effectively transfers style while preserving the geometric structure of the input asset. The visualization demonstrates that our single-pass generation approach ensures cross-layer consistency, successfully adapting the color palette, texture, and artistic rendering of external assets to match the pre-existing composition. This capability is essential for unifying disparate elements into a cohesive graphic design.

6.7. Layout Generalization in Text-to-Layer

Fig. 32 and Fig. 33 present an analysis of the interplay between text prompts and spatial controls. In these experiments, the text prompt contains implicit or explicit descriptions of element positions, while we simultaneously provide varying spatial layouts (bounding boxes) that may conflict with these textual descriptions. Remarkably, the results demonstrate that our model exhibits strong adherence to the user-provided layout, effectively overriding the spatial biases present in the text prompt while retaining the semantic content. This confirms that our framework successfully disentangles semantic generation from spatial arrangement, allowing users to enforce arbitrary layouts—such as moving a title from the top to the bottom—without compromising the generated content's quality or the prompt's semantic fidelity.

6.8. Layout-Guided Image Decomposition

For the Image-to-Layer task, Fig. 34 through Fig. 36 visualize the input raster images alongside their corresponding layout structures (bounding boxes and Z-order) used during inference. These examples illustrate how the model utilizes layout information—whether derived from automatic detectors or manual annotation—as a structural prior to guide the decomposition process. The visualizations show that the model accurately resolves ambiguities in the raster image by leveraging the provided spatial cues, resulting in semantically meaningful layers that strictly conform to the specified boundaries. This highlights the model's ability to produce controllable and predictable decompositions essential for professional editing workflows.

6.9. Generalization to Natural Scenes

Although our model is trained exclusively on graphic design datasets (posters, flyers, etc.), Fig. 37 demonstrates its zero-shot generalization capability to real-world natural images. The model successfully segments objects from photographs into transparent layers, leveraging the strong visual understanding inherited from the Qwen-Image backbone. However, we observe a specific limitation due to the domain gap: unlike flat graphic designs, real-world scenes contain complex physical lighting effects. Consequently, the model often fails to associate cast shadows with their respective objects, leaving shadows on the background layer rather than the object layer. Despite this limitation regarding physical lighting effects, the structural decomposition remains surprisingly robust for out-of-domain data.

6.10. Failure Cases

Finally, we analyze representative failure cases in Fig. 38 to provide a balanced view of our method's current limitations. We observe a common issue across all four tasks: some transparent backgrounds are decoded into gray instead of remaining transparent. This ambiguity arises because our VAE encoder currently uses a 3-channel input, which compresses transparent layers into a gray representation that the decoder sometimes misinterprets. Future work could address this by adopting a 4-channel encoder or alternative encoding schemes. Additionally, we identify task-specific limitations: 1) for text generation, the model sometimes struggles with rendering very small glyphs accurately; and 2) for layer-to-layer tasks, we occasionally observe failures in identity preservation (IP) and instruction following, particularly when complex style transfer or precise object insertion is required. These cases outline critical directions for future research in multi-layer generative modeling.

Figure 4. **Text-to-layers generation examples.** We visualize diverse text-to-layers generation results from our method, showing the input text prompts and corresponding multi-layer outputs. Each example displays individual transparent RGBA layers along with the merged composition. Our approach generates coherent multi-layer designs that maintain spatial consistency, stylistic harmony, and accurate layer boundaries, demonstrating strong alignment between text prompts and layered compositions.

Figure 5. **Additional text-to-layers generation examples.** More examples demonstrating our method's capability to generate multi-layer designs from text descriptions. These results showcase the diversity of generated layouts, layer compositions, and visual styles.

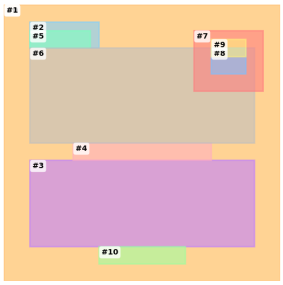
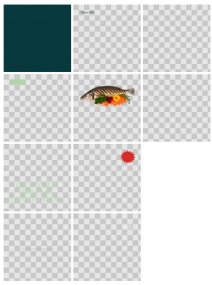
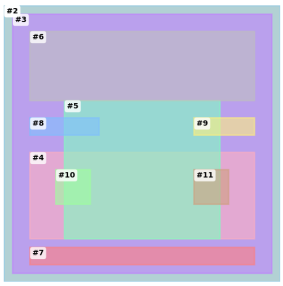
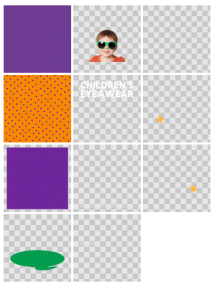
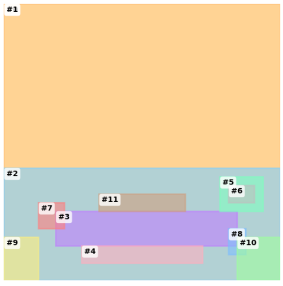
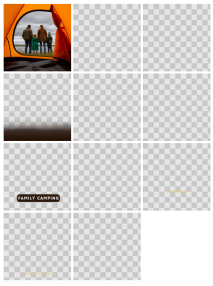
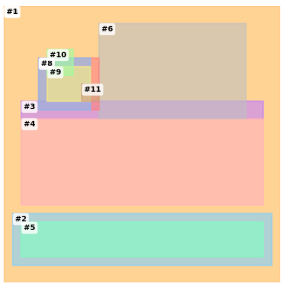
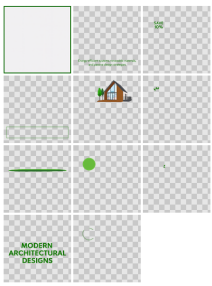
Prompt	Layout	Generated	Layer Images
The image presents a vibrant and enticing advertisement for a seafood restaurant's lunch special. The background is a rich teal color, providing a striking contrast to the other elements in the composition. This deep hue is complemented by a lighter green rectangular banner at the top, which reads "Today's Offer," creating a clear focal point that draws the viewer's attention. At the center of the image is a beautifully plated dish featuring a whole fish, which is the star of the advertisement. The fish is presented with meticulous attention to detail, showcasing its grilled skin with distinct char marks, and garnished with fresh herbs and colorful edible flowers. The dish is artfully arranged, with a bed of vibrant orange garnishes that add a pop of color and texture, enhancing the overall visual appeal. Prominently displayed in bold, modern typography is the phrase "GET YOUR SPECIAL LUNCH MENU," which is positioned just above the fish. The text is in a large, light green font that stands out against the darker background, ensuring readability and impact. Below this, the price "ONLY $5" is highlighted in a playful red starburst shape, further emphasizing the affordability of the offer and enticing potential customers. The overall artistic style of the image is contemporary and promotional, utilizing clean lines and a well-organized layout that effectively communicates the message. The use of high-quality photography for the food adds a realistic touch, making the dish look appetizing and inviting. The theme of the image is promotional and culinary, aimed at attracting customers to the restaurant for a special lunch deal. The mood conveyed is one of excitement and indulgence, encouraging viewers to take advantage of the offer. In terms of symbolism, the fish represents freshness and quality, key attributes for a seafood restaurant. The vibrant colors and appealing presentation suggest a delightful dining experience. This image could be effectively used in various contexts, such as social media posts, flyers, or website banners, to promote the restaurant's lunch specials and attract a wider audience.			
The image presents a vibrant and engaging advertisement for children's eyewear, set against a playful and colorful background. The overall background features a rich purple hue, which serves as a bold canvas for the promotional elements. This purple is framed by a lively orange border adorned with small polka dots, adding a cheerful and inviting touch to the design. At the top of the image, the text "CHILDREN'S EYEWEAR" is prominently displayed in large, white, bold letters, ensuring that the message is clear and eye-catching. Below this, the phrase "GET 40% OFF" is presented in a smaller font, also in white, which contrasts well against the purple background, drawing attention to the promotional offer. The central part of the image includes a green oval shape that contains additional text and graphic elements. Inside this green area, the phrase "CRAFTED FOR COMFORT, STYLED FOR SMILES" is written in a friendly and approachable font, reinforcing the product's appeal to both children and parents. The green color of the oval contrasts nicely with the purple background, creating a sense of balance and harmony. Flanking the green oval are two playful graphic elements resembling puzzle pieces, colored in a bright yellow-orange. These elements add a whimsical touch to the composition, enhancing the overall theme of fun and creativity associated with children's products. The artistic style of the image is modern and graphic, utilizing bold colors and clear typography to convey its message effectively. The use of contrasting colors and playful shapes contributes to a lively and engaging atmosphere, making it suitable for a target audience of parents looking for stylish and comfortable eyewear for their children. Overall, the image conveys a theme of joy and accessibility, with a mood that is both inviting and cheerful. It is well-suited for use in marketing materials, social media promotions, or as part of a website dedicated to children's products, effectively capturing the attention of potential customers and encouraging them to take advantage of the discount offer.			
The image presents a warm and inviting scene that captures the essence of family camping during the autumn season. The background features a soft, cloudy sky with shades of gray, suggesting a cool, overcast day by the water. This natural setting is framed by the vibrant orange fabric of a tent, which creates a cozy and welcoming atmosphere. The orange hue contrasts beautifully with the cooler tones of the sky and water, adding a touch of warmth to the overall composition. In the foreground, a family of four stands together, facing away from the viewer. They are positioned centrally within the tent's opening, creating a sense of intimacy and connection. The family members are dressed in autumn-appropriate clothing, with the adults wearing brown and green jackets, while the children are clad in a mix of colors, including a green coat and a gray beanie. This choice of attire not only reflects the season but also emphasizes the theme of outdoor adventure and togetherness. At the bottom of the image, a dark brown banner spans the width, providing a strong visual anchor. The text within the banner reads "FAMILY CAMPING" in bold, white letters, which stand out prominently against the dark background. Above this, in a smaller, golden font, the words "AUTUMN SEASON" and "JOIN OUR FESTIVAL EVENT" invite viewers to participate the festivities. A circular badge in the top right corner states "LIMITED TICKET!" in a playful font, adding a sense of urgency and excitement to the event. The artistic style of the image is modern and clean, with a focus on realism that captures the genuine expressions and postures of the family. The use of soft edges and natural lighting enhances the inviting mood, making it feel both relatable and aspirational. The overall theme is one of family bonding and outdoor exploration, evoking feelings of nostalgia and warmth associated with autumn activities. Symbolically, the image represents togetherness, adventure, and the joy of spending quality time with loved ones in nature. It could be effectively used for promotional materials related to family camping events, social media posts, or as part of a marketing campaign for outdoor gear and activities. The combination of visual elements and text creates a compelling invitation for families to engage in memorable experiences during the autumn season.			
The image presents a clean and modern design that emphasizes the theme of contemporary architecture with a focus on sustainability. The background is a soft, muted white, which provides a neutral canvas that allows the main elements to stand out prominently. Surrounding the central design is a thin green border, enhancing the overall aesthetic and framing the content effectively. At the center of the image is a stylized illustration of a modern house, characterized by its triangular roof and wooden exterior. The house features large windows that reflect a contemporary architectural style, and it is complemented by a small deck area adorned with a railing. To the left of the house, there are two stylized trees, rendered in a vibrant green, which add a touch of nature and reinforce the theme of eco-friendliness. The house is depicted in a simplified, almost cartoonish style, which contributes to a friendly and approachable feel. Prominently displayed above the house is a circular badge that reads "SAVE 10%" in a playful font, suggesting a promotional offer. This element is designed to catch the viewer's attention and is colored in a bright green that harmonizes with the overall color scheme. Below the house, the text "MODERN ARCHITECTURAL DESIGNS" is presented in bold, uppercase letters, using a darker green that contrasts well with the lighter background. This text is clear and assertive, effectively communicating the primary focus of the image. Further down, a smaller text block provides additional information: "Energy-efficient systems, renewable materials, and passive design strategies." This text is written in a more subdued font, ensuring that it complements the bolder headings without overwhelming them. The overall composition is well-balanced, with the house and trees anchoring the visual elements while the text guides the viewer's eye. The artistic style of the image leans towards a minimalist and modern illustration, utilizing flat design techniques with clean lines and simple shapes. The use of green tones throughout the image not only emphasizes the theme of sustainability but also evokes feelings of growth and harmony with nature. The overall theme of the image is promotional and informative, aimed at attracting potential customers interested in modern architectural solutions that prioritize energy efficiency and sustainability. The mood conveyed is optimistic and forward-thinking, appealing to environmentally conscious individuals or businesses. This image could be effectively used in marketing materials, such as flyers, social media posts, or website banners, to promote architectural services or products that align with modern, eco-friendly design principles.			

Figure 6. **Additional text-to-layers generation examples.**

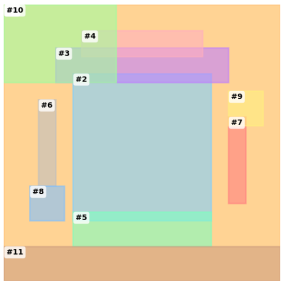
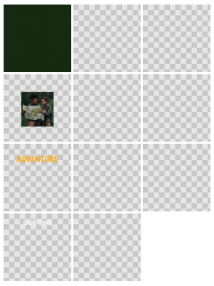
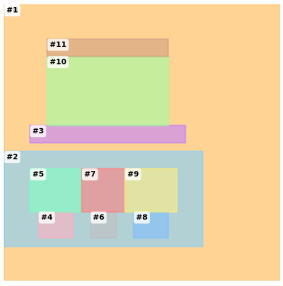
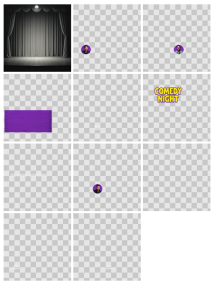
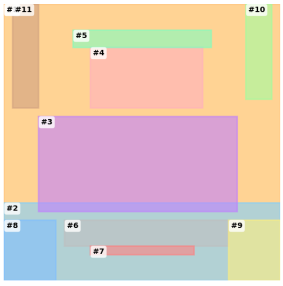
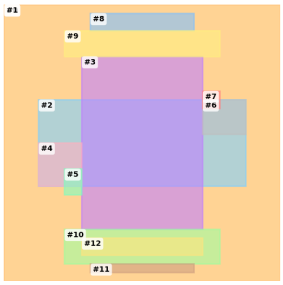
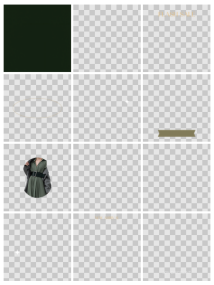
Prompt	Layout	Generated	Layer Images
The image presents a vibrant and engaging design that captures the spirit of adventure. The background features a rich, dark green color that provides a strong contrast to the lighter elements in the foreground. This deep hue evokes a sense of nature and exploration, aligning perfectly with the theme of outdoor adventures. At the center of the image, two individuals are depicted, engaged in a moment of planning their journey. They are holding a map, which serves as a focal point, symbolizing exploration and discovery. The individuals are dressed in casual outdoor attire, suggesting they are ready for an adventure. The textures of their clothing, combined with the natural setting behind them, enhance the authenticity of the scene. The top of the image prominently displays the phrase "OUR TIME ADVENTURE" in a bold, playful font. The word "ADVENTURE" is highlighted in a warm, golden yellow, which stands out against the darker background and draws the viewer's attention. This choice of color adds a sense of excitement and positivity to the overall message. Beneath the main title, there is a motivational tagline: "Embark on your next adventure with peace of mind. Prioritize your body's safety and health every step of the way." This text is presented in a clean, white font that is easy to read and complements the overall design. The message encourages viewers to embrace adventure while also being mindful of their well-being. The artistic style of the image leans towards a modern, graphic design approach, with a mix of photography and digital elements. The use of dashed lines and a paper airplane graphic adds a whimsical touch, suggesting movement and the idea of travel. Overall, the theme of the image is adventurous and motivational, conveying a sense of excitement and readiness for exploration. The mood is uplifting and encouraging, making it suitable for various contexts, such as travel blogs, social media posts, or promotional materials for outdoor activities. This image effectively inspires viewers to embark on their own adventures while keeping their health and safety in mind.			
The image presents a vibrant promotional design for a comedy event titled "Comedy Night." The background features a stage setting with soft, draped curtains that create a sense of anticipation and excitement. The color scheme is predominantly dark, with shades of gray and black, which contrasts effectively with the bright and playful colors used in the text and graphic elements. At the top of the image, the date and location are prominently displayed: "MARCH 15 | CHICAGO, IL." This information is set in a clean, sans-serif font that is easy to read, ensuring that the essential details are immediately noticeable. Below this, the main title "COMEDY NIGHT" is rendered in bold, yellow letters with a playful, cartoonish style. The text is outlined in purple, adding depth and a fun, energetic vibe to the overall design. Beneath the title, there is a tagline that reads "EPIC STANDUP SHOW COMING," which is presented in a smaller, white font. This text is positioned to maintain a balanced composition, drawing the viewer's eye from the title down to the details of the event. The lower section of the image features three circular portraits of the performers, each enclosed in a purple background that matches the outline of the title. The names of the comedians—Jake Davis, Noah Smith, and Leo Jones—are displayed beneath their respective images in a clean, white font, ensuring clarity and visibility. The use of circular frames for the portraits adds a modern touch to the design. The overall artistic style of the image is contemporary and engaging, utilizing a mix of bold typography and vibrant colors to convey a sense of fun and entertainment. The mood is light-hearted and inviting, perfectly suited for a comedy show, suggesting an evening filled with laughter and enjoyment. This image could be effectively used for social media promotions, event flyers, or digital advertisements, appealing to audiences looking for a night of entertainment and humor. The combination of visual elements and text creates a cohesive and attractive promotional piece that captures the essence of a lively comedy night.			
The image presents a vibrant and engaging promotional graphic for a movie night event. The background is a deep gray, providing a modern and sleek canvas that enhances the visibility of the foreground elements. This color choice creates a contrast that draws attention to the text and illustrations. At the top of the image, bold white text reads, "PLEASE JOIN US FOR," which is positioned above the main event title, "MOVIE NIGHT," rendered in large, eye-catching letters. The font is playful yet clear, suggesting a fun and inviting atmosphere. The use of white against the dark background ensures that the text stands out prominently. The central composition features three stylized figures seated in plush red cinema chairs, each holding popcorn. The characters are illustrated in a minimalist style, with simplified features and muted colors that convey a casual and relaxed vibe. The woman on the left is depicted with long, flowing hair and a yellow top, while the man in the middle wears a purple shirt, and the figure on the right is dressed in a light blue shirt. This arrangement creates a sense of camaraderie and enjoyment, typical of a movie night experience. At the bottom of the image, additional details are provided in a clean, sans-serif font. The text includes the date, "SATURDAY, JULY 23 | 8 PM," and the location, "AT MOVIE CINEMA," all of which are clearly legible and well-organized. The inclusion of a website link, "www.COMPANYNAME.com," suggests a professional touch, encouraging attendees to seek more information. The overall artistic style is digital and contemporary, characterized by flat colors and simple shapes, which contribute to a friendly and approachable aesthetic. The use of popcorn graphics in the bottom left corner adds a playful element, reinforcing the theme of a movie night. The mood conveyed by the image is festive and inviting, making it suitable for social gatherings and community events. The combination of engaging visuals and clear information makes this image ideal for use in social media promotions, flyers, or event announcements, effectively capturing the excitement of a movie night.			
The image presents a stylish promotional graphic with a modern and elegant aesthetic. The background is a deep, rich green that provides a sophisticated canvas, enhancing the overall visual appeal. This dark backdrop contrasts beautifully with the lighter elements in the foreground, creating a striking visual hierarchy. At the center of the composition is a figure dressed in a fashionable outfit, which consists of a long, flowing dress paired with a stylish jacket and a wide belt that cinches at the waist. The clothing features a muted color palette, primarily in shades of green and gray, which harmonizes with the background while also allowing the figure to stand out. The outfit's textures and layers suggest a contemporary fashion style, appealing to a modern audience. Surrounding the figure are decorative elements, including a circular shape that frames the model, along with star-like motifs that add a touch of whimsy and sparkle. These elements are rendered in white, which contrasts sharply against the dark background, drawing the viewer's eye toward the central figure and enhancing the overall composition. Prominently displayed at the top of the image is the text "NEW ARRIVAL" in a bold, uppercase font, conveying a sense of urgency and excitement. Below this, the phrase "FLASH SALE" is presented in a slightly larger font, emphasizing the promotional aspect of the image. The text is rendered in a warm gold color, which adds a luxurious feel and complements the overall color scheme. At the bottom of the image, a banner announces a "50% DISCOUNT," using a clean, sans-serif font that is easy to read. This call to action is crucial for engaging potential customers and encourages immediate interest in the sale. The website URL, "www.fashionstyle.com," is also included, providing a direct link for viewers to explore further. The overall artistic style of the image is modern and polished, with a focus on fashion and retail. The combination of elegant typography, a cohesive color palette, and stylish imagery creates a mood that is both inviting and aspirational. This image is well-suited for use in digital marketing campaigns, social media promotions, or as part of an online store's advertising materials, effectively capturing the essence of contemporary fashion retail.			

Figure 7. **Additional text-to-layers generation examples.**

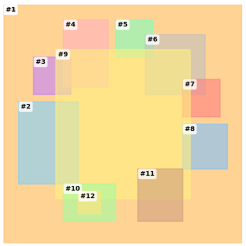
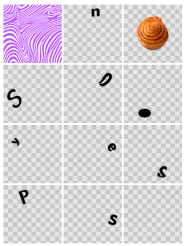
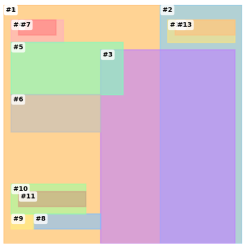
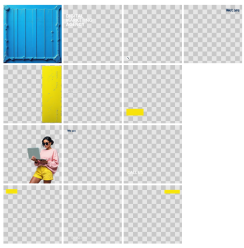
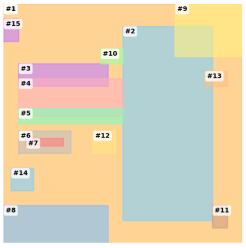
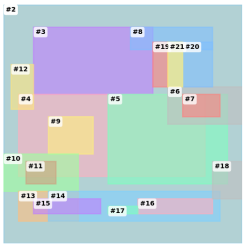
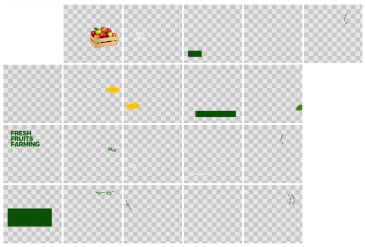
Prompt	Layout	Generated	Layer Images
The image presents a vibrant and playful composition centered around a delicious pastry, likely a cinnamon roll or a similar baked good. The background features a dynamic pattern of swirling stripes in shades of lavender and white, creating a sense of movement and energy that draws the viewer's eye. The alternating colors of the background evoke a lighthearted and whimsical atmosphere, enhancing the overall appeal of the image. At the forefront, the pastry is prominently displayed, showcasing its golden-brown, flaky layers that spiral inward. The texture of the pastry is rich and inviting, with a glossy sheen that suggests a fresh-baked quality. The intricate details of the roll, including the visible swirls and layers, highlight the craftsmanship involved in its creation, making it the focal point of the image. Surrounding the pastry are bold, black letters that spell out "yPnOsS," with the letters varying in size and orientation, adding to the playful and modern aesthetic. The typography is striking, with a mix of uppercase and lowercase letters that create a sense of spontaneity. Below the pastry, a rounded black banner contains the text "SHELLYP TISSERIE," rendered in a clean, sans-serif font that complements the overall design. This text serves to identify the brand or establishment associated with the pastry, reinforcing its artisanal quality. The artistic style of the image leans towards a contemporary graphic design approach, characterized by its use of bold colors, playful typography, and a focus on food aesthetics. The combination of the vibrant background and the detailed pastry creates a visually appealing contrast that is both eye-catching and appetizing. The theme of the image is celebratory and indulgent, evoking feelings of joy and satisfaction associated with enjoying a sweet treat. The mood conveyed is cheerful and inviting, making it an excellent choice for marketing materials related to a bakery or pastry shop. In terms of symbolism, the pastry represents comfort and indulgence, while the playful typography and background suggest a modern, trendy approach to food. This image could be effectively used in various contexts, such as social media promotions, advertisements for a bakery, or as part of a menu design, appealing to a youthful and vibrant audience looking for delightful culinary experiences.			
The image presents a vibrant and modern design, primarily featuring a bright blue background that conveys a sense of professionalism and creativity. The blue is complemented by a contrasting yellow rectangle at the top, which contains the text "We are" in a bold, sans-serif font. This combination of colors creates a visually appealing and energetic atmosphere. At the center of the composition is a figure dressed in a casual yet stylish outfit, consisting of a light pink oversized sweater and yellow shorts. The individual is holding a laptop, which is prominently displayed in front of them, suggesting a focus on digital technology and marketing services. The laptop screen is blank, which may symbolize the potential for creativity and innovation in digital marketing. The text "DIGITAL MARKETING AGENCY" is positioned prominently in white, bold letters, making it the focal point of the image. Below this, a smaller tagline reads, "We are ready to help your business marketing," which reinforces the agency's commitment to supporting clients. The font used here is clean and modern, enhancing the professional tone of the image. In the lower section, there is a yellow box with the call-to-action "CALL US" in white text, accompanied by a phone icon and a contact number. This element is designed to encourage immediate engagement from potential clients, making it clear how to reach the agency. The overall artistic style of the image is contemporary and graphic, utilizing flat design techniques that emphasize clarity and directness. The bright color scheme and modern typography contribute to a mood that is both inviting and professional, ideal for a digital marketing agency aiming to attract new clients. This image could be effectively used in various contexts, such as social media advertisements, website banners, or promotional materials, to convey the agency's services and encourage potential clients to reach out for assistance. The combination of engaging visuals and clear messaging makes it a strong representation of a modern marketing agency.			
The image presents a modern and stylish promotional design for a clothing item, specifically a hoodie. The background features a soft, creamy beige color that creates a warm and inviting atmosphere. This neutral backdrop is complemented by abstract shapes in darker green tones, adding a touch of sophistication and visual interest without overwhelming the main elements. At the center of the composition is a figure seated on a wooden stool, dressed in a trendy outfit that includes a black hoodie with striking green accents. The hoodie is layered over a light-colored top, and the figure is wearing light-colored pants that contribute to a casual yet fashionable look. The choice of footwear, which consists of rugged boots, enhances the overall streetwear aesthetic. The figure's pose is relaxed, suggesting comfort and confidence in the attire. Text elements are strategically placed around the figure. The phrase "New Arrivals" is prominently displayed at the top in a playful, handwritten font, while the word "Hoodie" is featured in a bold, modern typeface that commands attention. Below the main title, a smaller line reads, "Clotes are the spirit of fashion," which appears to be a typographical error for "Clothes." This text is presented in a more understated font, allowing the main message to stand out. Additionally, there is a call-to-action button that says "Order Now," designed in a contrasting dark green color that invites immediate engagement. The overall artistic style of the image is contemporary and clean, with a focus on minimalism that highlights the product. The use of soft shadows and subtle textures adds depth to the design, making it visually appealing. The theme of the image revolves around fashion and lifestyle, conveying a mood that is both trendy and approachable. It suggests a sense of modernity and casual elegance, appealing to a youthful audience interested in stylish clothing. In terms of symbolism, the hoodie represents comfort and versatility in fashion, while the overall design promotes a sense of accessibility to new trends. This image could be effectively used in various contexts, such as social media advertisements, online clothing stores, or promotional materials for fashion brands, aiming to attract customers looking for the latest in casual wear.			
The image presents a vibrant and engaging promotional graphic centered around fresh fruit farming. The background features a clean, white grid pattern that adds a modern and organized feel to the composition, enhancing the overall visual appeal. In the foreground, a wooden crate brimming with fresh apples takes center stage. The apples are a mix of red and yellow hues, showcasing their natural freshness and inviting appearance. Some apples are partially obscured by green leaves, which add a touch of nature and vitality to the scene. The crate itself is positioned slightly to the right, creating a dynamic balance with the text elements on the left. The text "FRESH FRUITS FARMING" is prominently displayed in bold, green typography at the top of the image. The font is modern and sans-serif, contributing to a clean and contemporary aesthetic. Below this, a striking "45% OFF" is highlighted in a large, white font against a solid green background, drawing immediate attention to the promotional offer. This text is further emphasized by a playful, starburst-shaped "SALE!" graphic in yellow, which adds a sense of urgency and excitement. Additional text elements include a call-to-action "SHOP NOW" in a bold, white font on a dark green rectangular button, accompanied by an arrow pointing to the right, guiding the viewer's eye towards the next step. Below this, the address "123 Any St, Any City" is presented in a smaller, more understated font, ensuring that the focus remains on the promotional message. The overall artistic style of the image is modern and graphic, utilizing a combination of bold colors and playful shapes to create an eye-catching advertisement. The use of green tones not only aligns with the theme of freshness and nature but also evokes feelings of health and vitality. The theme of the image is promotional and inviting, aimed at attracting customers to purchase fresh fruits. The mood conveyed is energetic and positive, encouraging viewers to take advantage of the sale. This image could be effectively used in various marketing contexts, such as social media advertisements, flyers, or website banners, particularly for businesses related to agriculture, grocery stores, or health-focused brands. The combination of appealing visuals and clear messaging makes it a versatile tool for engaging potential customers.			

Figure 8. **Additional text-to-layers generation examples.**

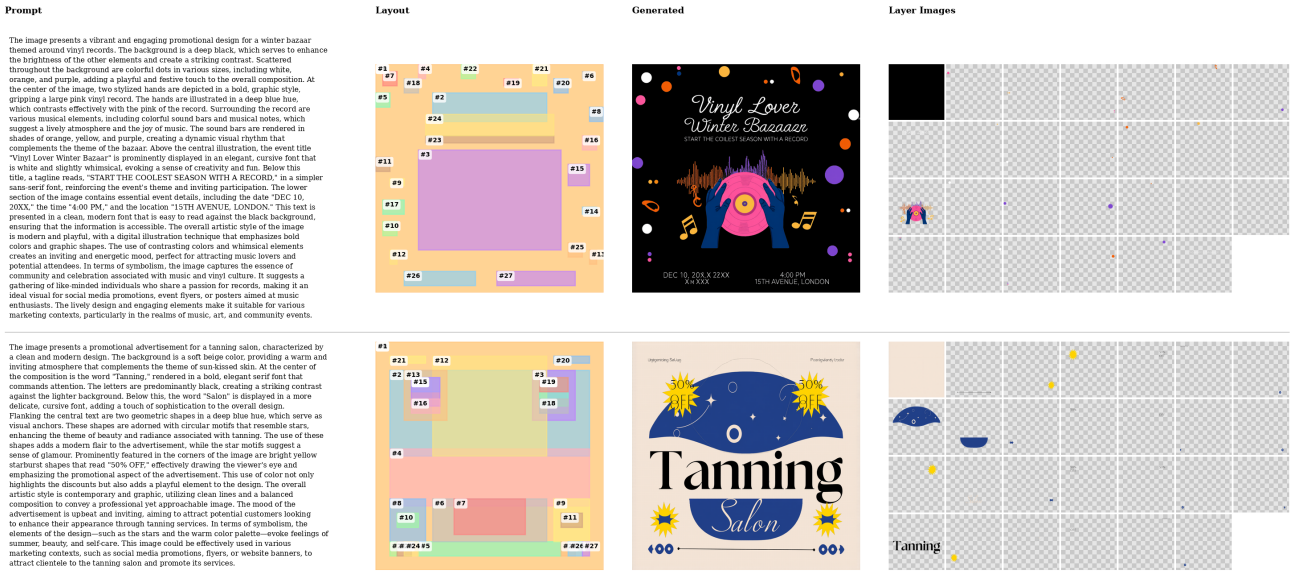


Figure 9. **Additional text-to-layers generation examples.** Our unified framework handles various design complexities, from simple compositions to intricate multi-element designs with over 25 layers, while maintaining generation quality.

Figure 10. **Additional text-to-layers generation examples.** Our model can generate rich images along with high-quality layers beyond flat designs.

Figure 11. **Text-to-layers with overflow layer generation.** Additional examples highlighting our method's unique capability to generate overflow layers that extend beyond the background boundary. As discussed in Section 3 and shown in Fig. 3, over 60% of designs contain overflow layers. Our approach generates complete full-size RGBA layers on the canvas, preserving editability and reusability that previous methods sacrifice by truncating pixels at background boundaries.

Figure 12. **Text-to-layers with multilingual support.** Examples demonstrating our model's capability to generate designs with multilingual text layers. Our dataset includes diverse languages (as shown in Fig. 2), enabling generation of visually rendered text in multiple languages including Chinese. This showcases the model's ability to handle typography across different writing systems while maintaining design quality.

Figure 13. **Qualitative comparison on image-to-layers task.** We compare our method with LayerD, Lovart, and RoboNeo on decomposing a graphic design into transparent layers. Each panel shows: the input image (top-left), followed by our result and baseline results with their decomposed layers. Our method produces cleaner layer boundaries, better granularity, and more complete RGBA layers compared to the baselines.

Figure 14. **Additional qualitative comparison on image-to-layers task.** Our method demonstrates superior layer decomposition quality with better semantic correctness and transparency handling. The decomposed layers from our method maintain higher integrity and can faithfully reconstruct the input image, while baselines show issues with layer artifacts, improper grouping, or incomplete decomposition.

Figure 15. **Additional qualitative comparison on image-to-layers task.** This example further demonstrates our method's advantages in layer quality, integrity, and appropriate granularity. Our approach successfully decomposes complex compositions while avoiding the overly grouped layers produced by LayerD or the artifacts present in commercial system outputs.

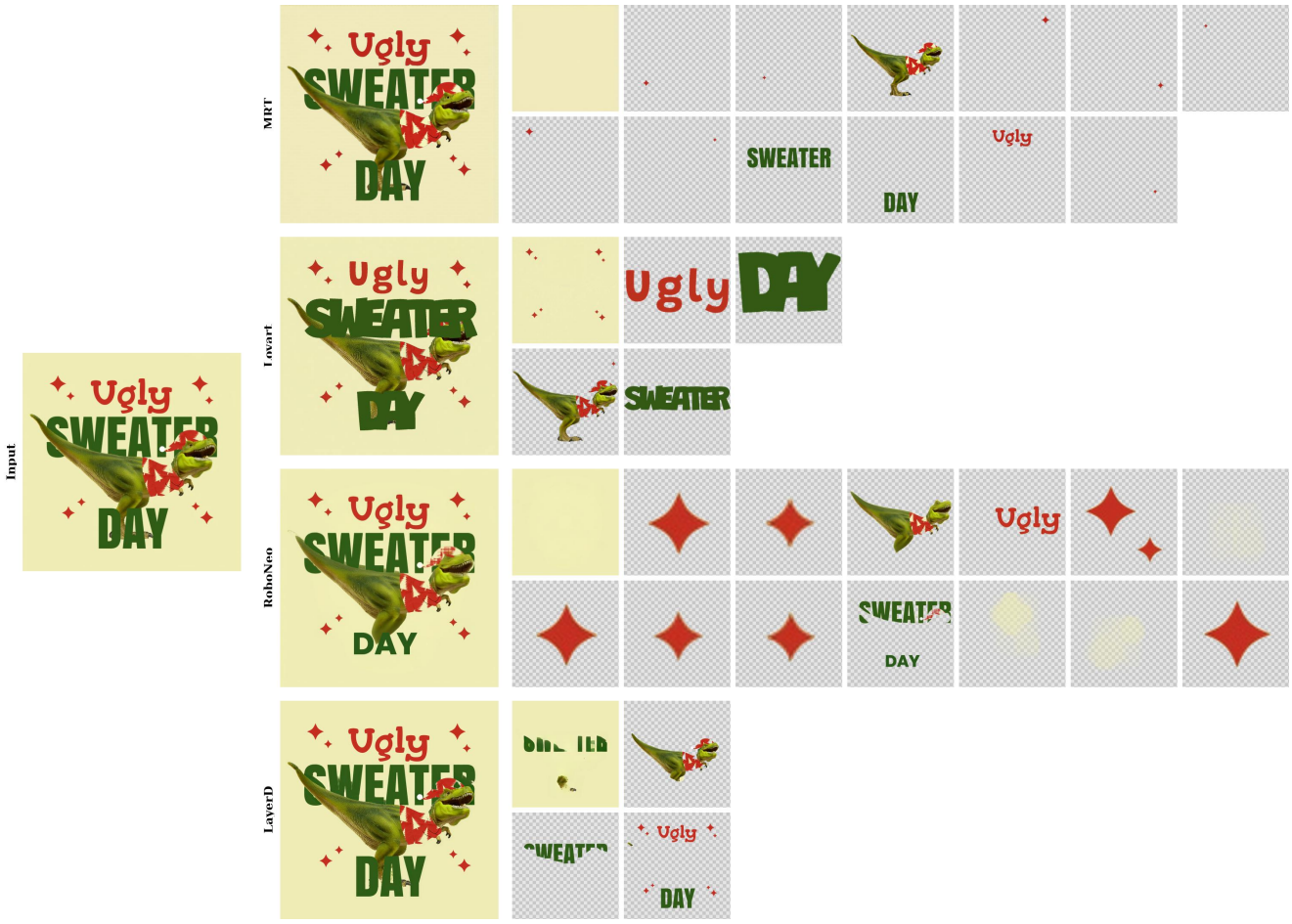


Figure 16. **Additional qualitative comparison on image-to-layers task.** Our method consistently outperforms baselines across different design styles and complexities. The visualization shows that our approach produces high-quality transparent layers with accurate alpha channels and proper semantic decomposition, essential for downstream editing tasks.

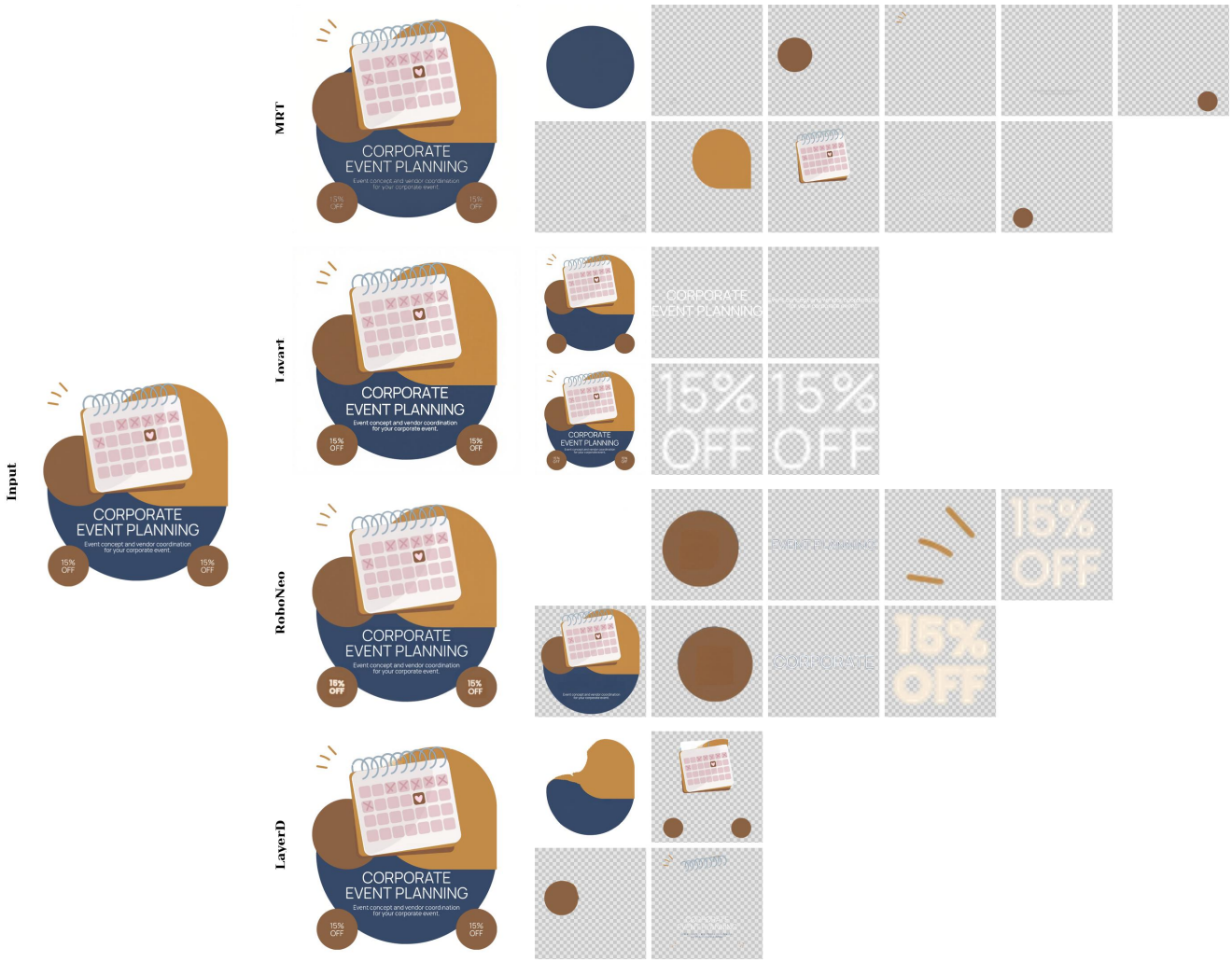


Figure 17. **Additional qualitative comparison on image-to-layers task.** This case highlights our method's ability to handle complex multi-element designs. While commercial systems like RoboNeo suffer from severe artifacts and LayerD produces overly grouped layers that compromise fine-grained editing flexibility, our method maintains both quality and appropriate decomposition granularity.

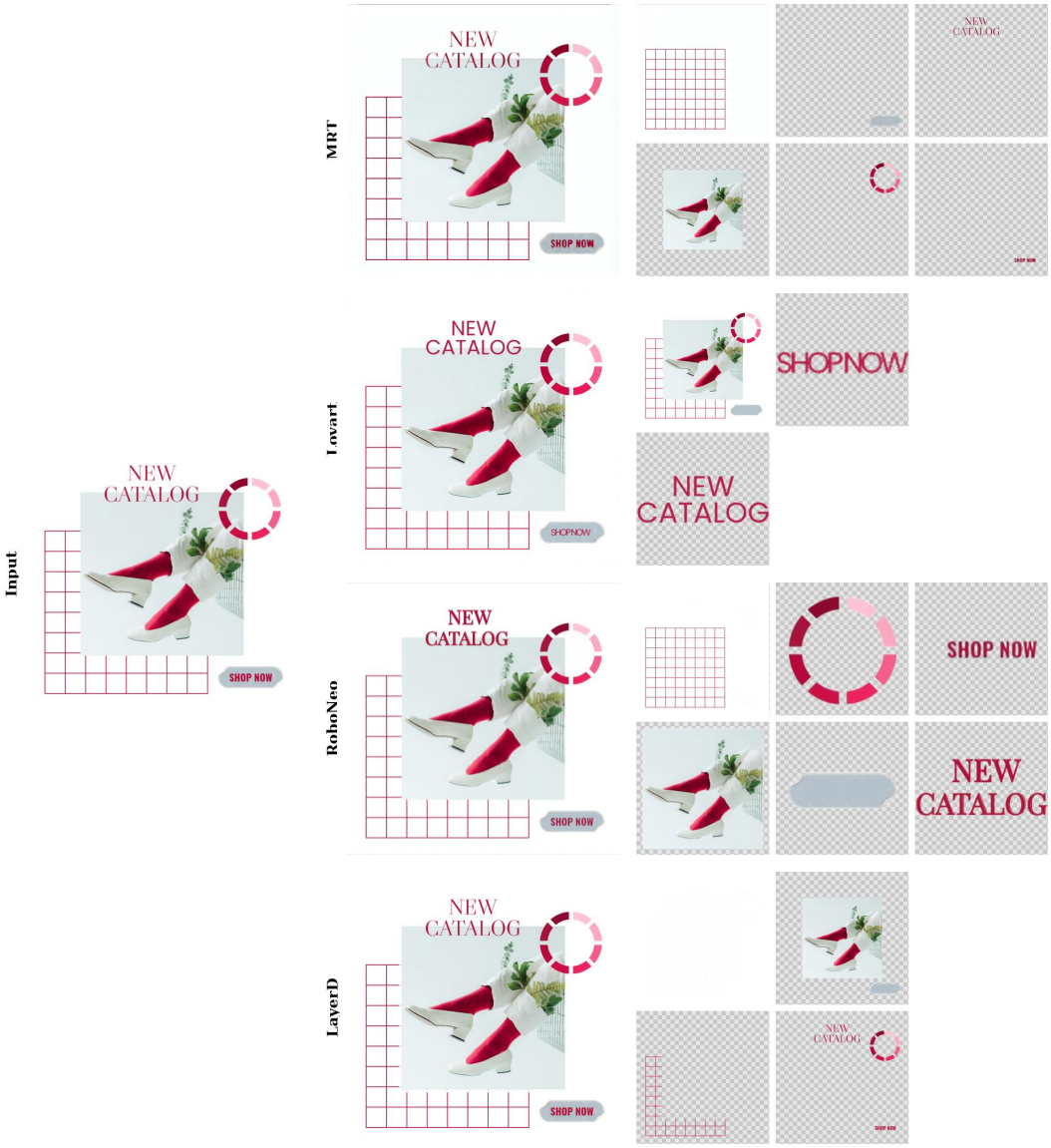


Figure 18. **Additional qualitative comparison on image-to-layers task.** Our method excels at decomposing designs with overlapping elements and complex visual hierarchies. The comparison demonstrates superior performance across all three evaluation dimensions: quality (semantic correctness and transparency), integrity (faithful reconstruction), and granularity (appropriate decomposition level).

Figure 19. **Additional qualitative comparison on image-to-layers task.** This example showcases our method's robustness across different design categories. Our decomposed layers maintain sharp boundaries, clean transparency, and semantic coherence, enabling practical editing workflows that commercial and academic baselines struggle to support.

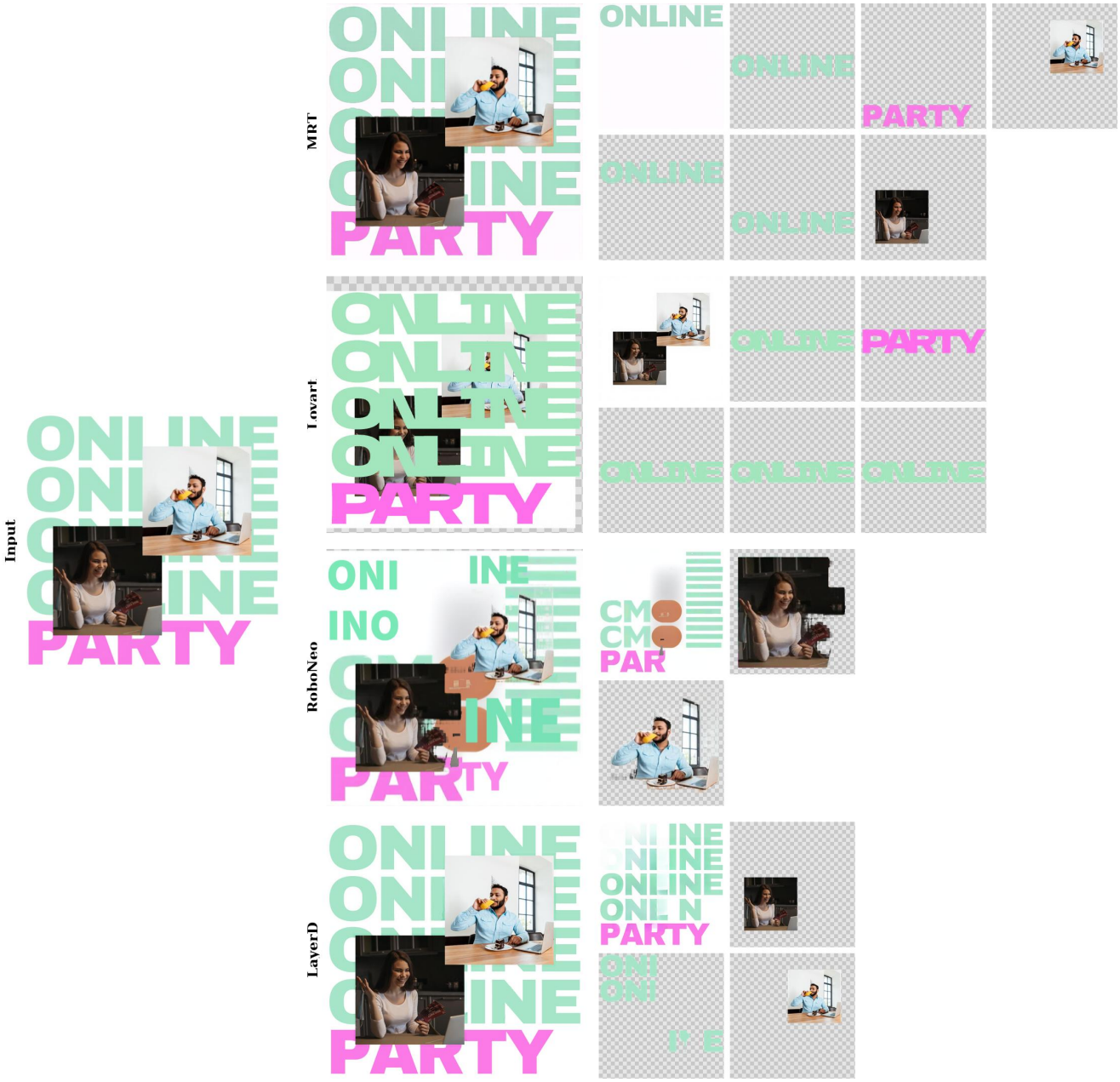


Figure 20. **Additional qualitative comparison on image-to-layers task.** Final comparison case demonstrating consistent quality advantages of our method. The decomposition preserves layer reusability and editability while maintaining visual fidelity, confirming the effectiveness of our masked region transformer framework for the image-to-layers task.

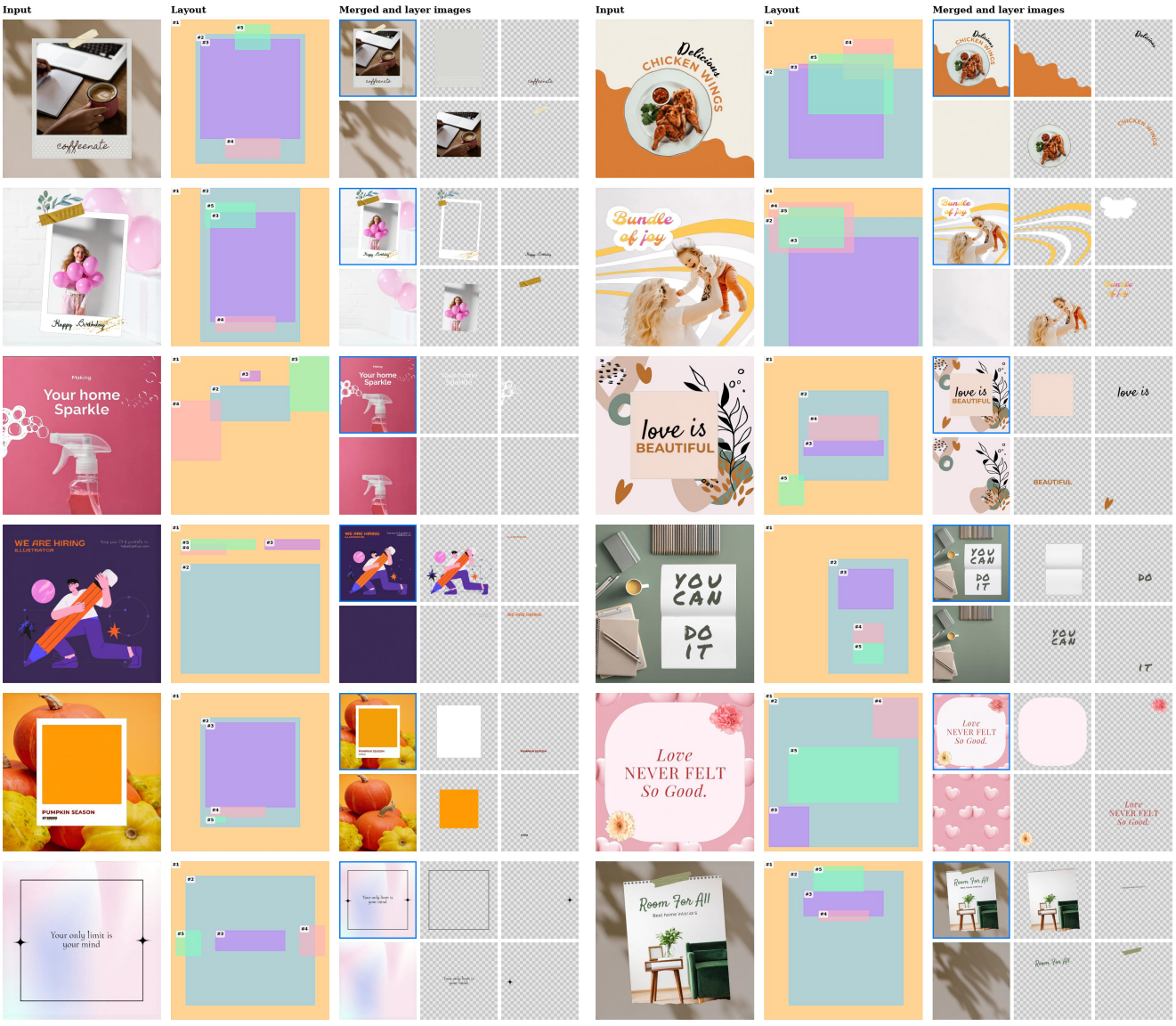


Figure 21. **Image-to-layers visualization with 6 layers.** We visualize the layer-by-layer decomposition process showing individual RGBA layers with transparency. Each layer is displayed separately along with its alpha mask, and the merged composition demonstrates faithful reconstruction of the input design. This visualization demonstrates our method's ability to generate clean, reusable layers with accurate spatial boundaries and alpha channels.

Figure 22. **Image-to-layers visualization with 8 layers.** Decomposition result showing increased layer complexity with 8 distinct transparent layers. Our method successfully handles more complex compositions, maintaining layer quality and proper decomposition granularity across the extended layer hierarchy. Each layer preserves semantic meaning and can be independently edited.

Figure 23. **Image-to-layers visualization with 10 layers.** Further demonstrating scalability to compositions with 10 transparent layers. Our masked region transformer maintains stable performance across different layer counts, producing coherent decompositions without architectural modifications. The visualization shows consistent layer quality from background to foreground elements.

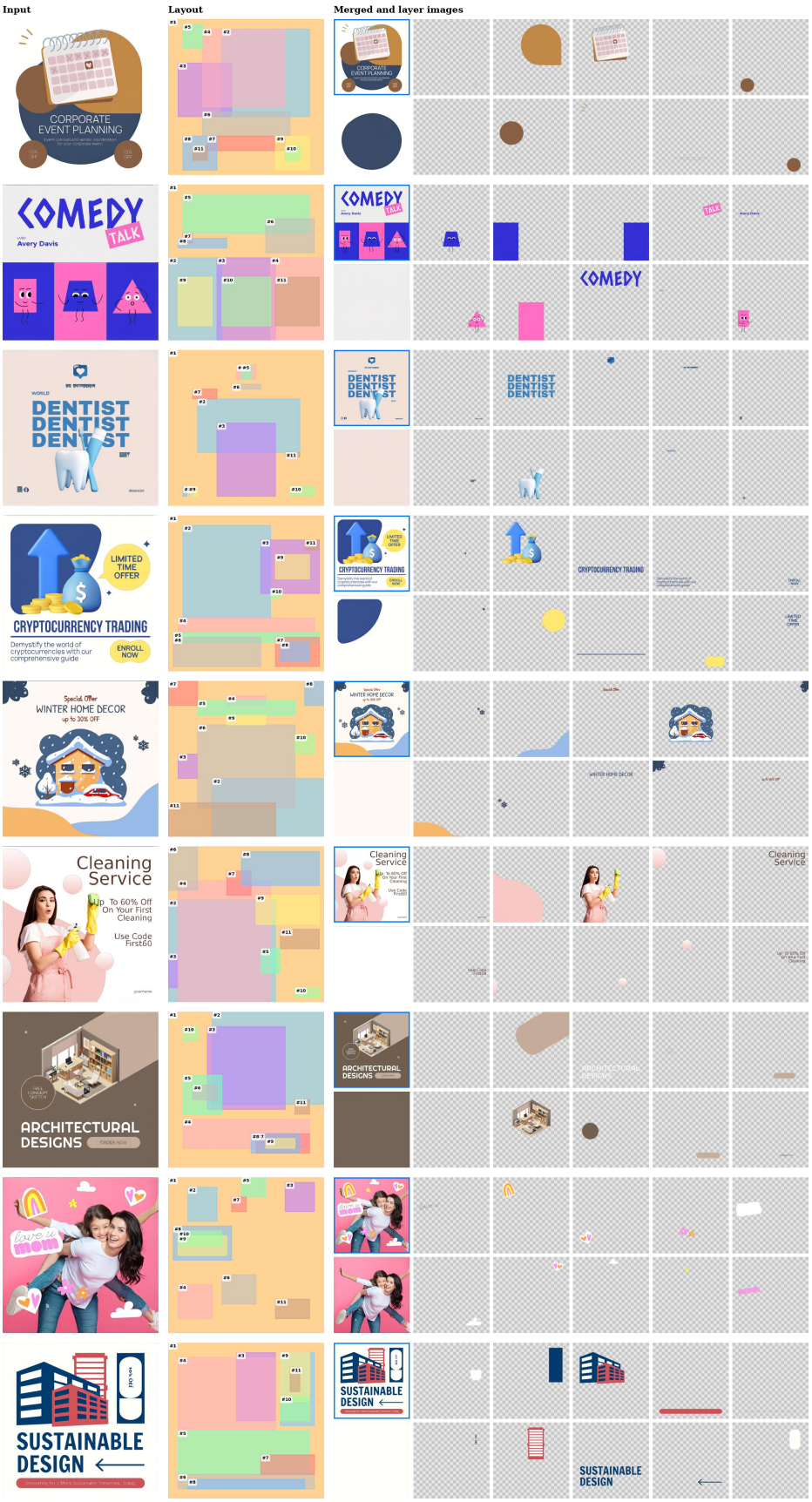


Figure 24. **Image-to-layers visualization with 12 layers.** Decomposition of a complex design into 12 transparent layers, demonstrating our method's capability to handle high layer counts while maintaining decomposition quality. Each layer retains sharp boundaries and proper alpha masks, essential for professional editing workflows.

Figure 25. **Image-to-layers visualization with 14 and 16 layers.** Two examples showcasing our method's scalability to very high layer counts (14 and 16 layers respectively). As shown in Table 3, our approach maintains stable performance across a wide range of layer numbers from 2 to 50 layers, demonstrating flexibility in handling both simple and complex multi-element compositions.

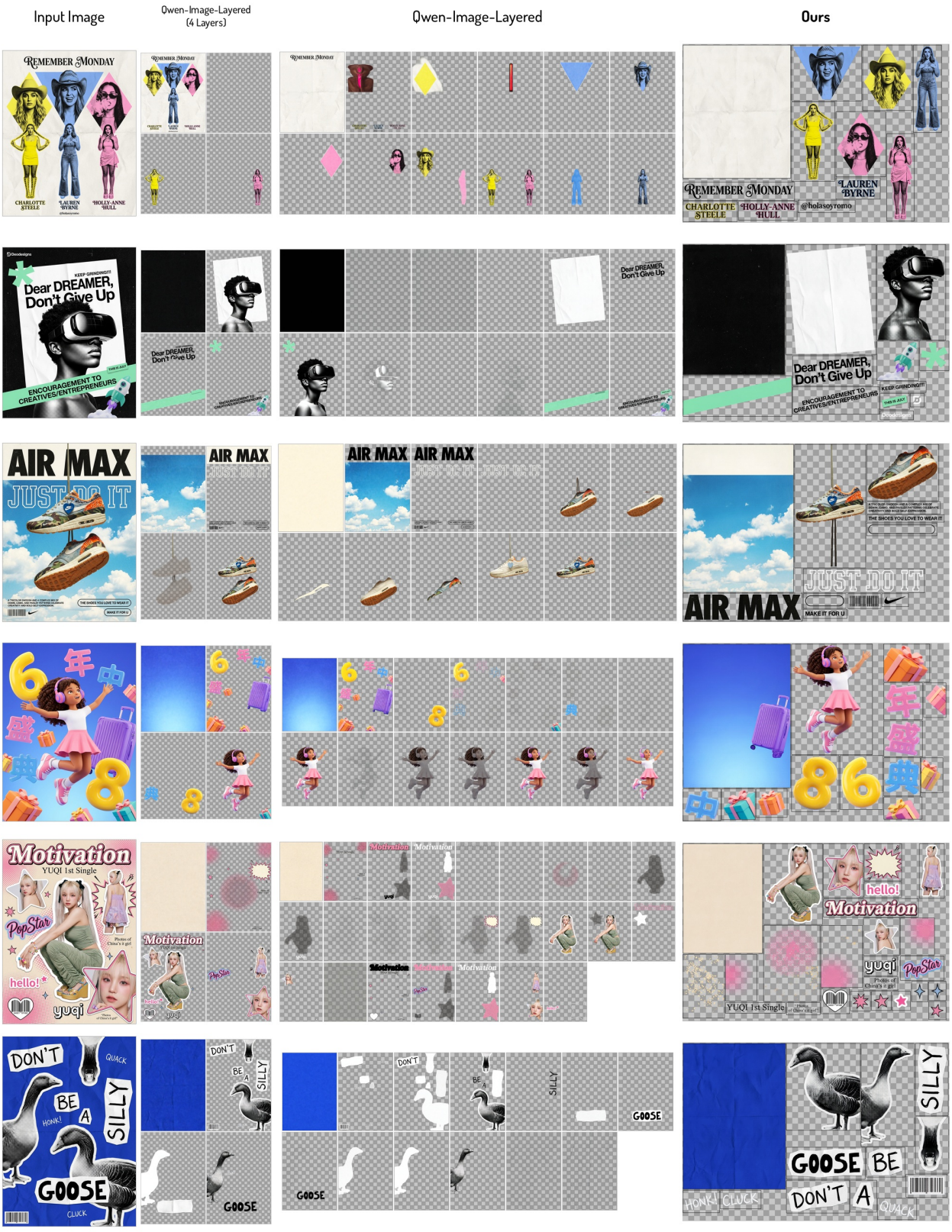


Figure 26. **Image-to-layers comparison with Qwen-Image-Layered**

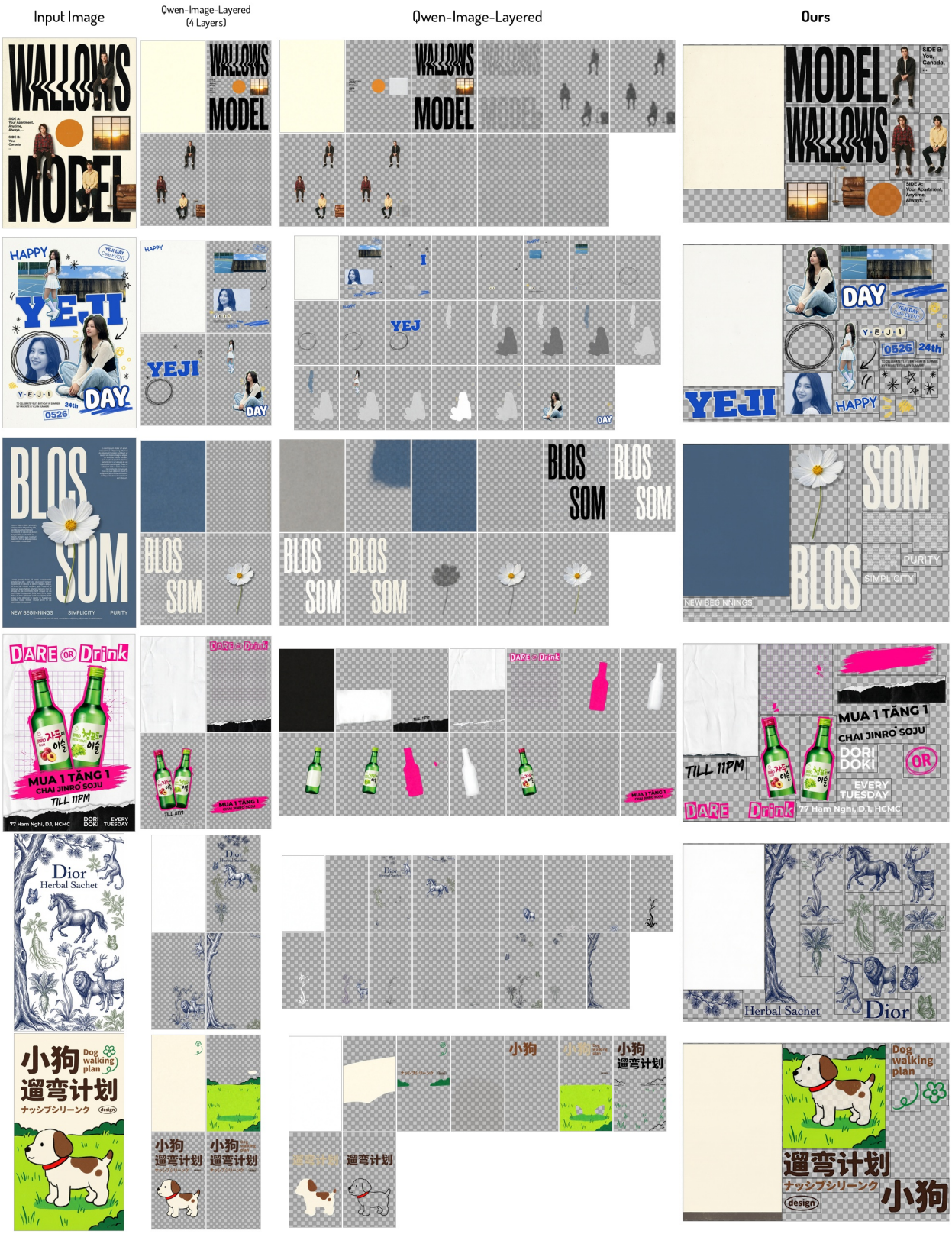


Figure 27. **Image-to-layers comparison with Qwen-Image-Layered**

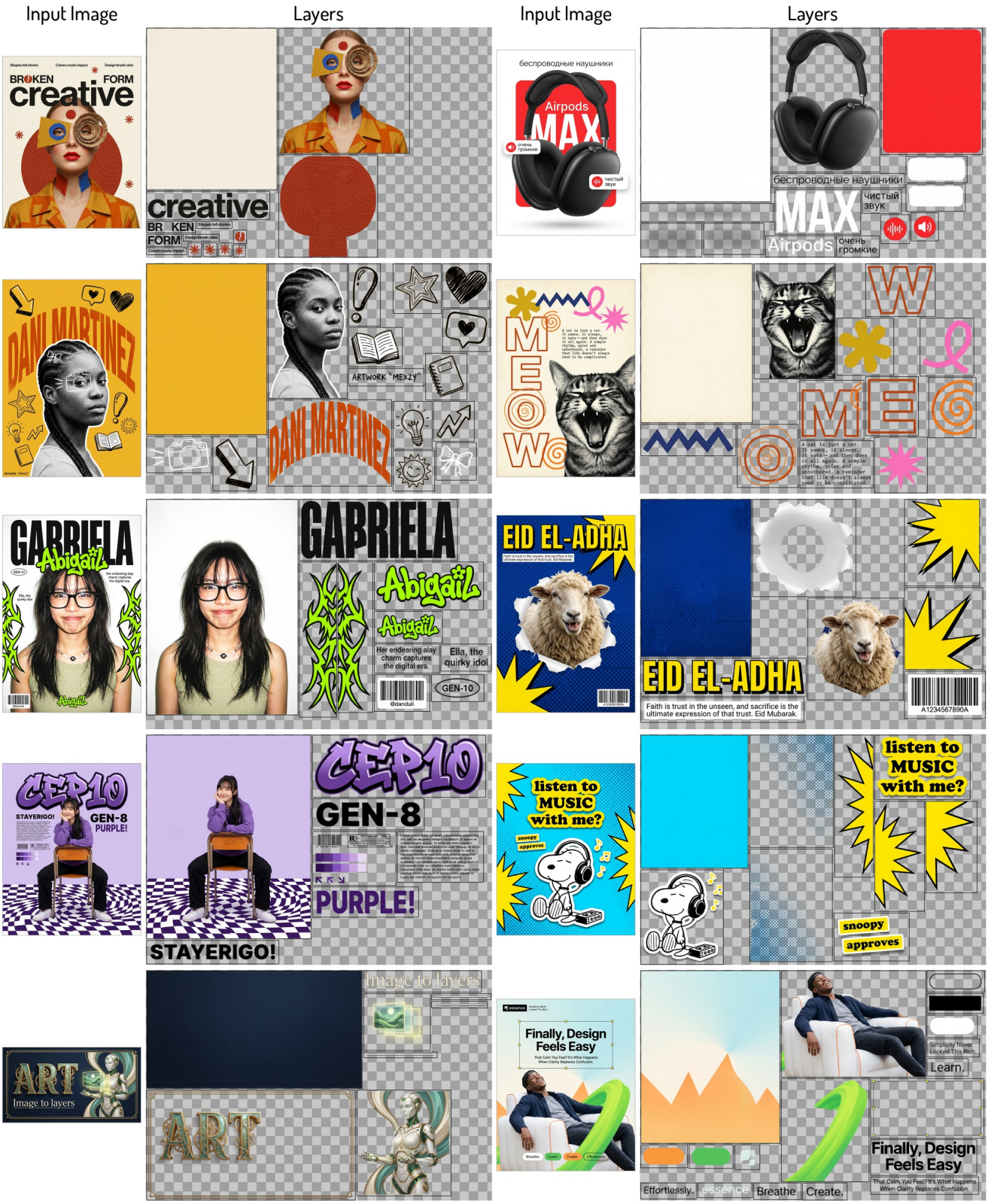


Figure 28. More image-to-layers results on AI-generated images.

Figure 29. More image-to-layers results on AI-generated images.

Figure 30. **Additional examples for layer addition task.** We demonstrate the layers-to-layers capability by adding new layers to existing compositions based on text prompts. Our method generates new layers that maintain cross-layer consistency and harmonize with the existing design's spatial layout and visual style. By generating multiple layers in a single pass and conditioning on all existing layers, our approach better captures inter-layer relationships and produces coherent insertions that preserve global composition.

Figure 31. **Additional examples for layer restylization task.** We visualize the transformation of user-provided assets into stylistically harmonized layers that match the overall composition. Our method performs this restylization in a single pass for all target layers, preserving geometric structure while adapting appearance to align with the existing design's visual identity. The results demonstrate effective style transfer while maintaining layer coherence and compositional harmony.

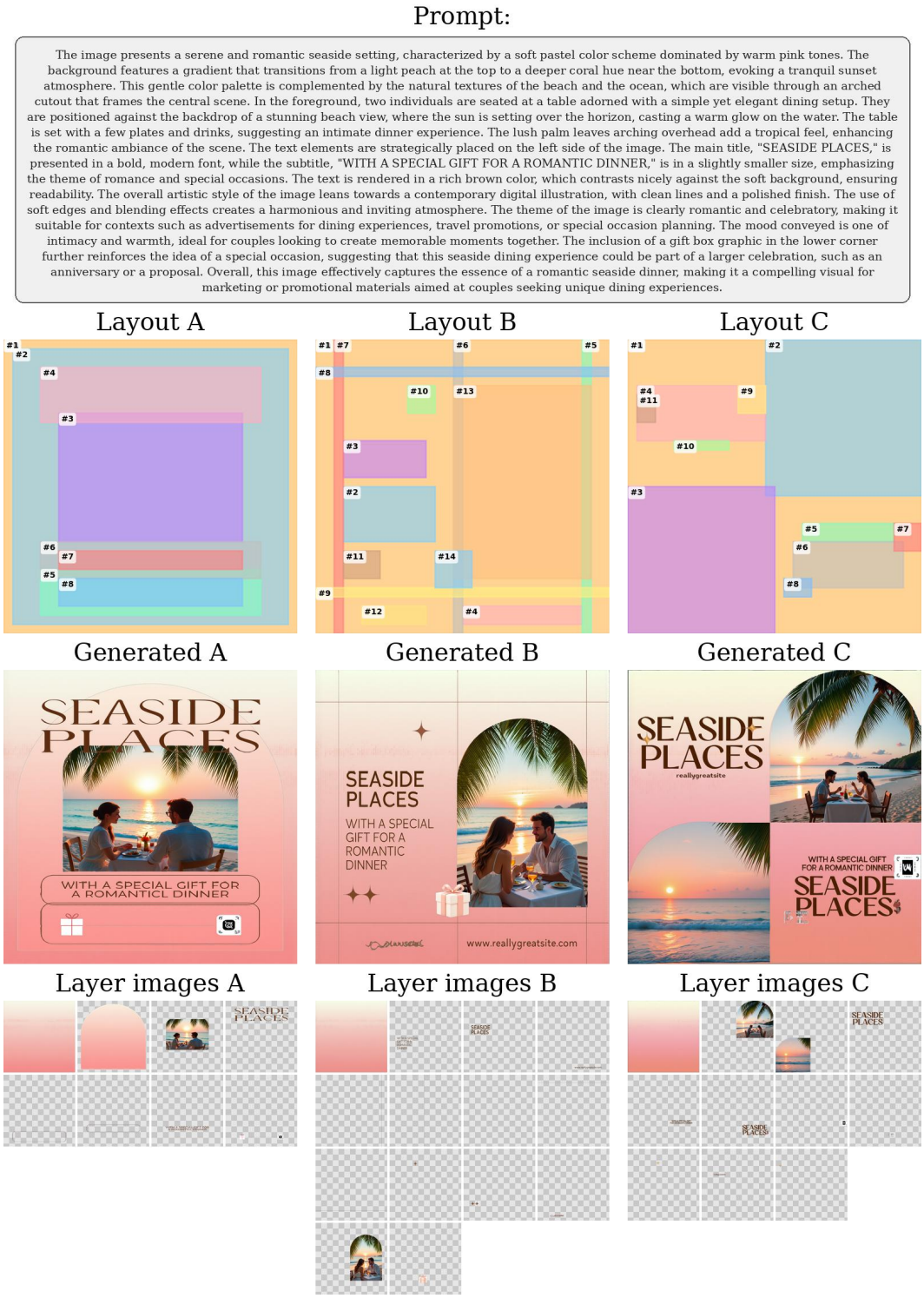


Figure 32. **Text-to-layers: Merged image vs. layout visualization.** Additional example demonstrating our model's ability to generate well-composed multi-layer designs from text prompts. The side-by-side comparison shows how textual descriptions are translated into visual compositions (left) with structured layer hierarchies (right), highlighting the model's capability to learn both aesthetic and structural design principles.

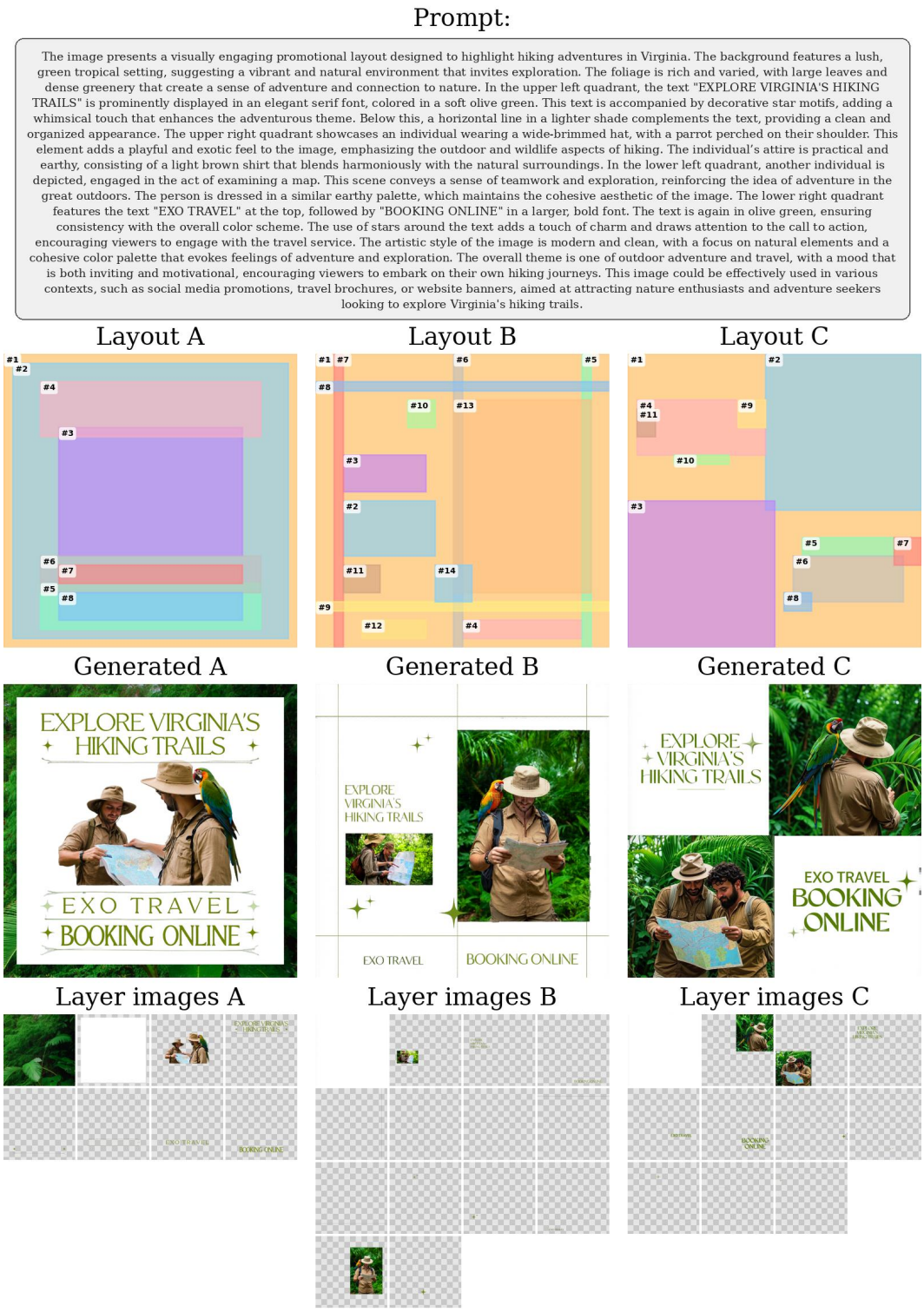


Figure 33. **Text-to-layers: Merged image vs. layout visualization.** Another example showing the relationship between the generated merged design and its underlying layer layout structure. The layout visualization reveals how our model organizes multiple layers with appropriate spatial relationships, z-ordering, and compositional balance to create aesthetically pleasing designs from text descriptions.

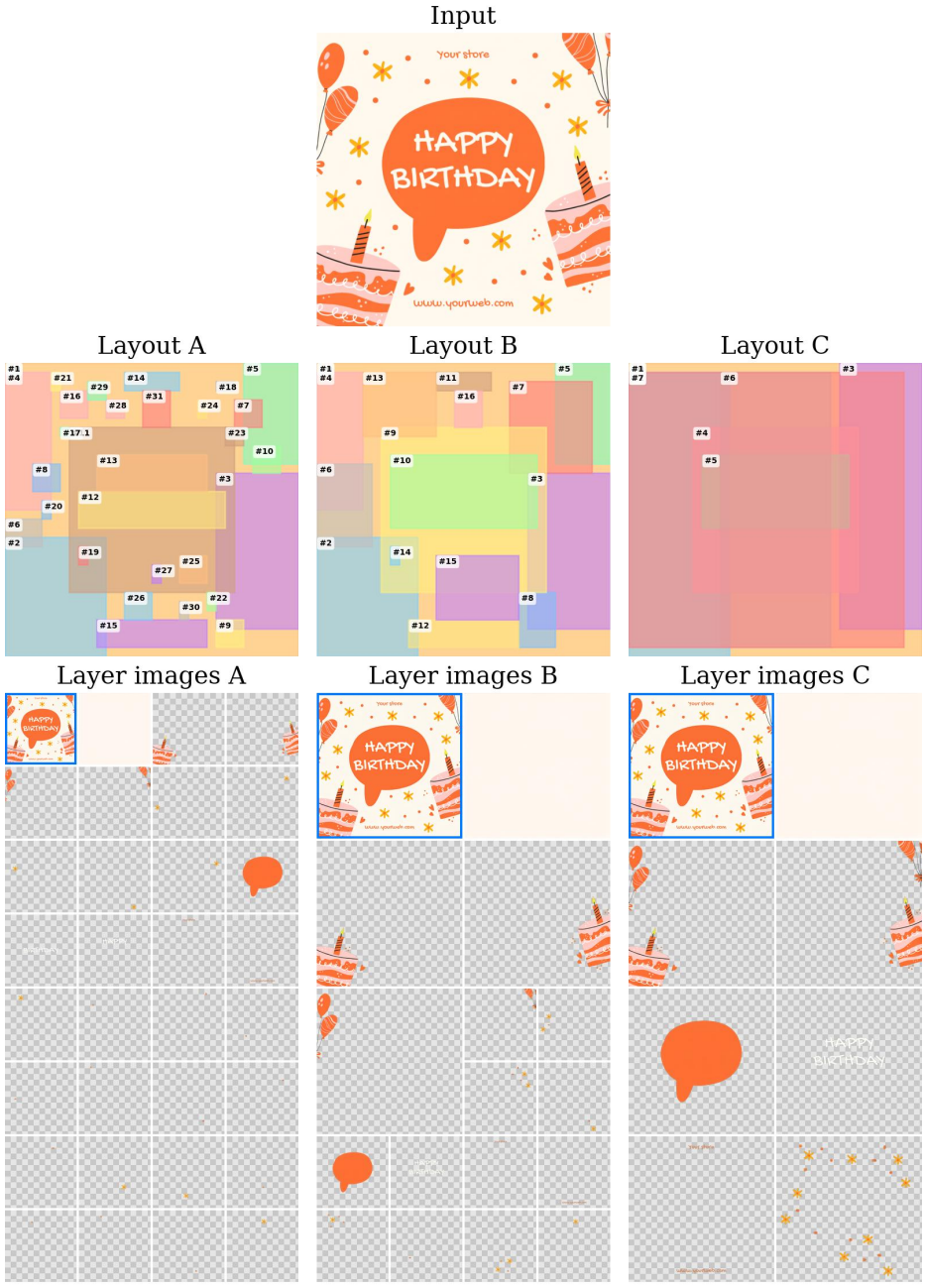


Figure 34. **Image-to-layers: Merged image vs. layout visualization.** We visualize the input image alongside the extracted layer layout structure for the image-to-layers decomposition task. This demonstrates how our method decomposes raster images into semantically meaningful layers with well-defined spatial boundaries. The layout representation shows bounding boxes and z-order that guide the decomposition process.

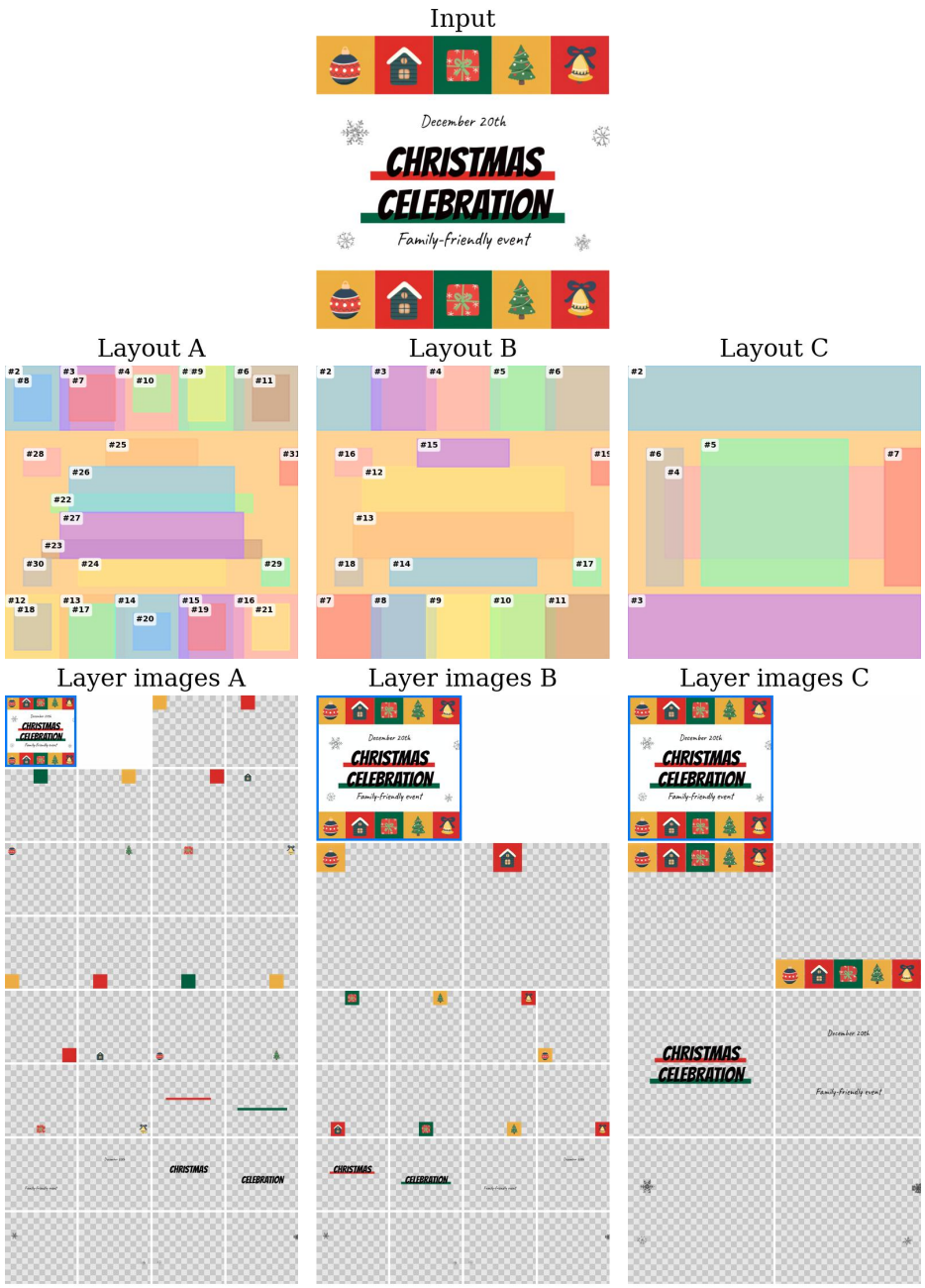


Figure 35. **Image-to-layers: Merged image vs. layout visualization.** Another example illustrating the correspondence between input raster images and their layer layouts. Our method leverages layout information (either from automatic detectors or manual annotations) to perform accurate layer decomposition. The layer grouping augmentation strategy helps improve robustness to noisy or ambiguous layout specifications.

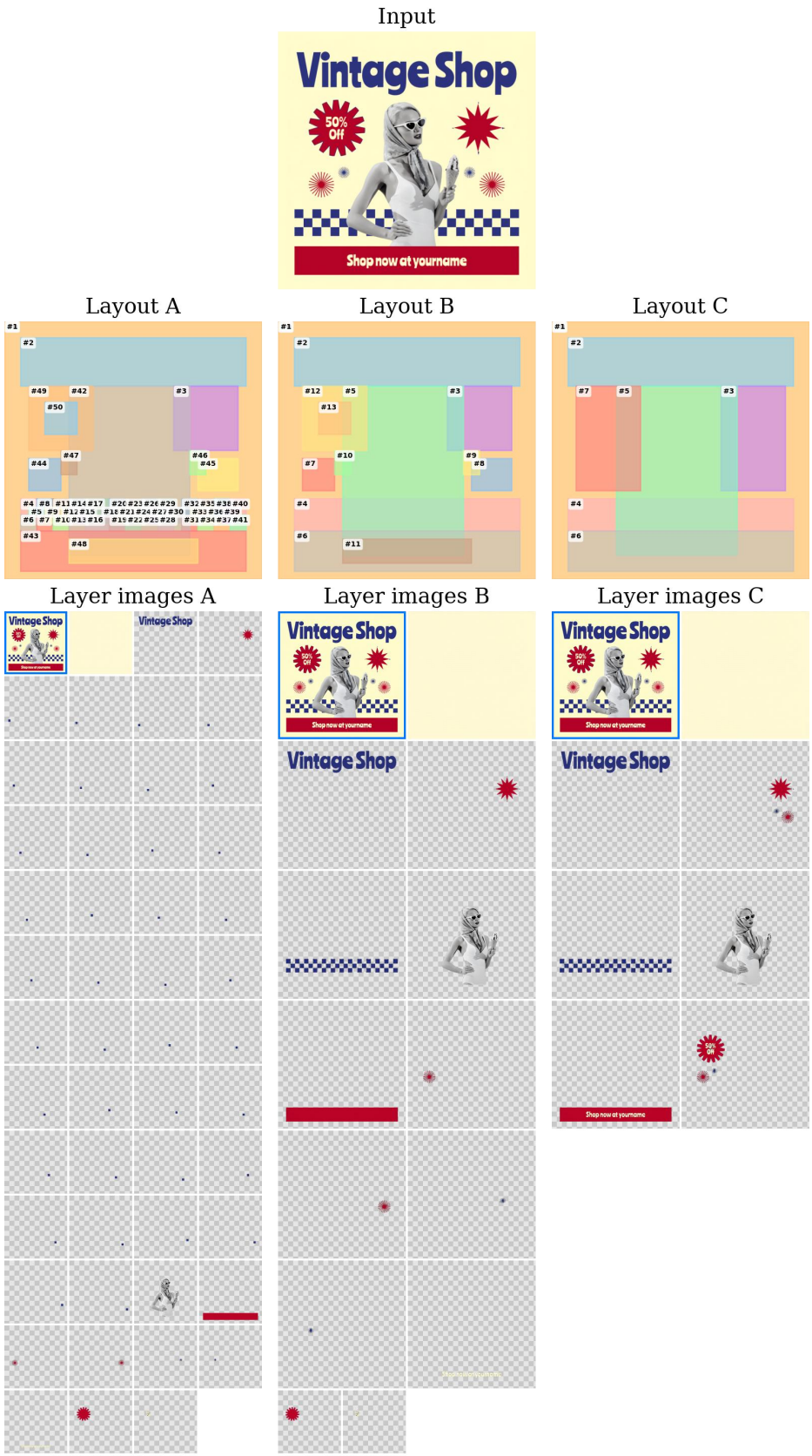


Figure 36. **Image-to-layers: Merged image vs. layout visualization.** Final example showing the input-layout relationship in image-to-layers decomposition. This visualization confirms our method's ability to handle diverse design categories and layout complexities, producing high-quality transparent layers that can be independently edited while maintaining faithful reconstruction of the original composition.

Figure 37. **Image-to-layers on real-world photographs: Limitation analysis.** We demonstrate our method's generalization to out-of-domain natural images. Despite being trained exclusively on design datasets, our model can decompose real photographs into layers. However, as discussed in the Limitations section, the model faces challenges with physical effects like shadows—often excluding shadow regions from object layers and leaving them on the background. This limitation stems from the domain gap between planar designs and real-world scenes with lighting effects. Nevertheless, the strong visual understanding from the Qwen-Image backbone enables reasonable generalization, with most objects successfully separated.

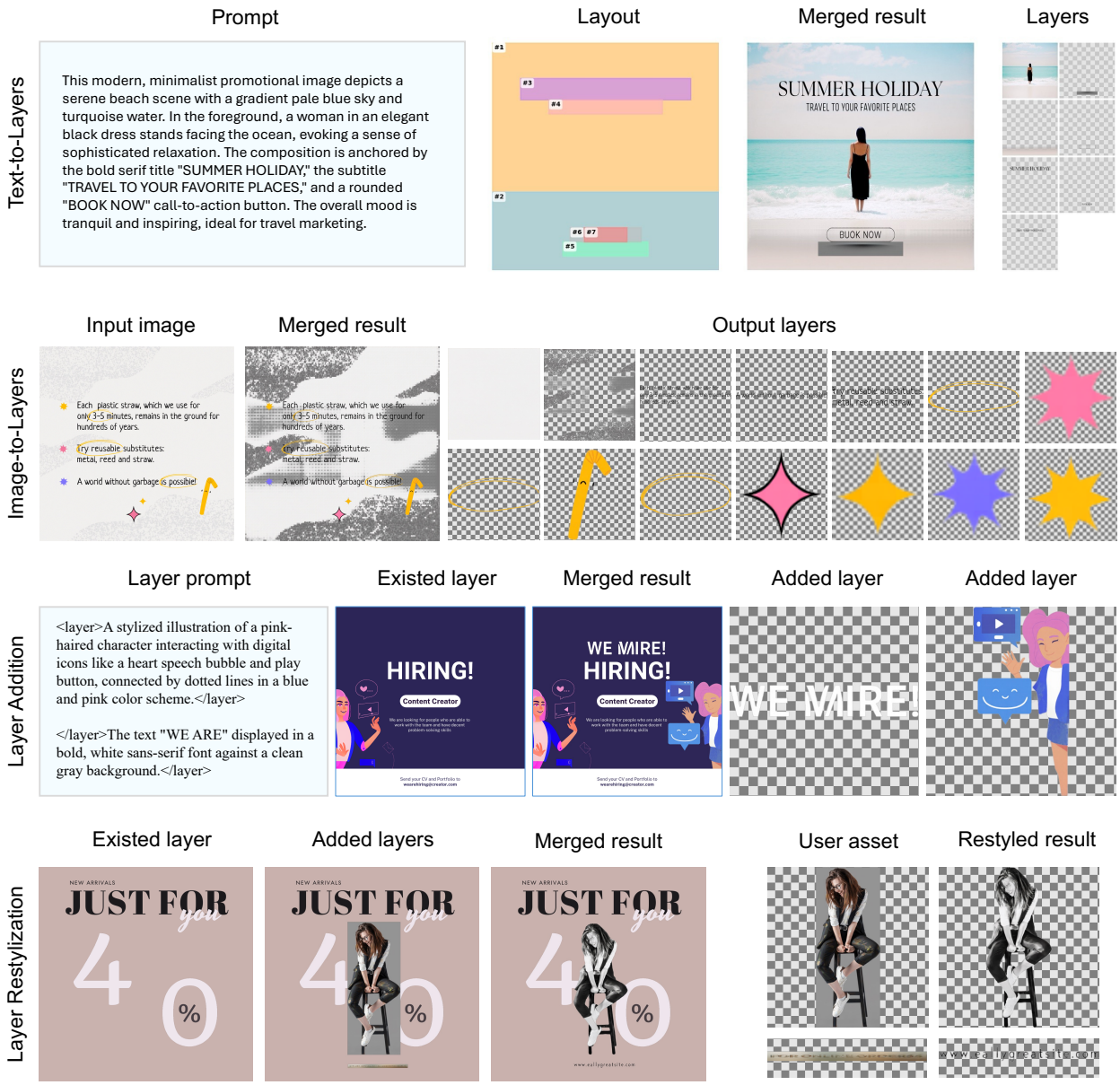


Figure 38. **Failure cases and limitations.** We present representative failure cases across our tasks. A common issue (top right) is the ”gray background” artifact, where transparent areas are decoded as gray due to the ambiguity of 3-channel VAE encoding. Other limitations include (bottom left) malformed glyphs when generating very small text, and (bottom right) occasional failures in identity preservation and instruction following during layer-to-layer editing.

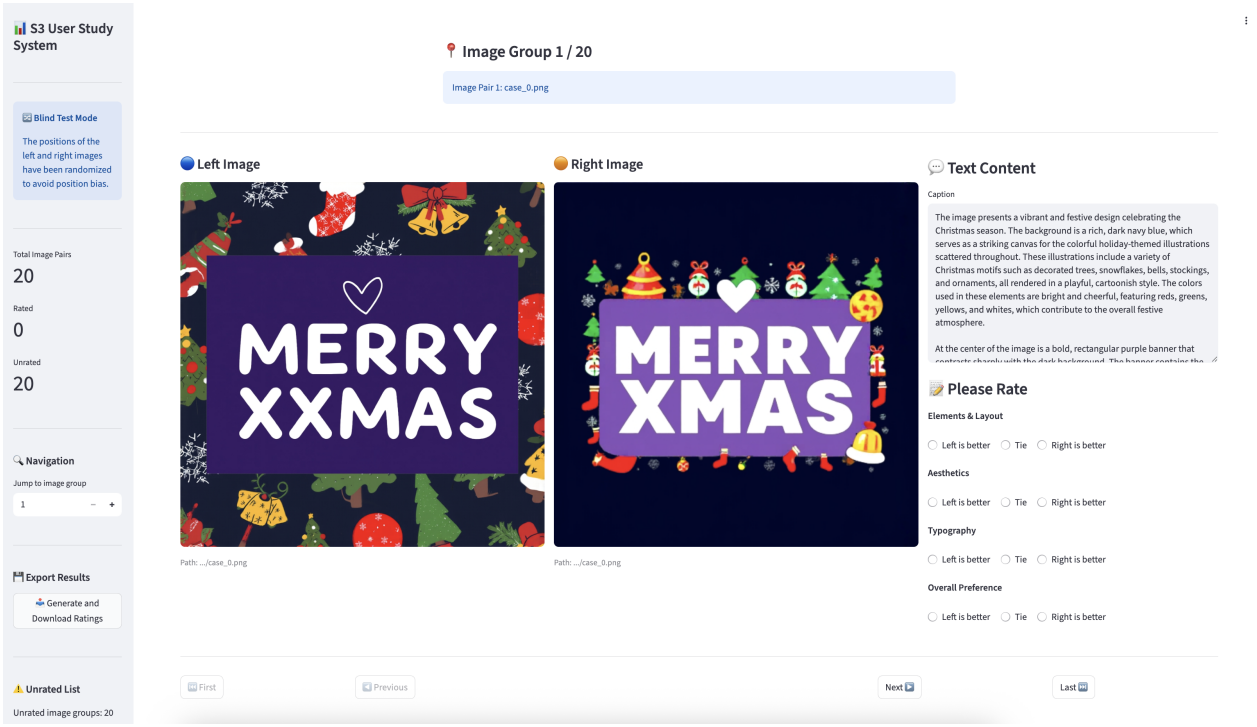


Figure 39. **User study interface for text-to-layers evaluation.** Two generated results are displayed side-by-side with the text caption shown on the right. Participants vote across four dimensions: elements (layout), aesthetics, typography, and overall preference.

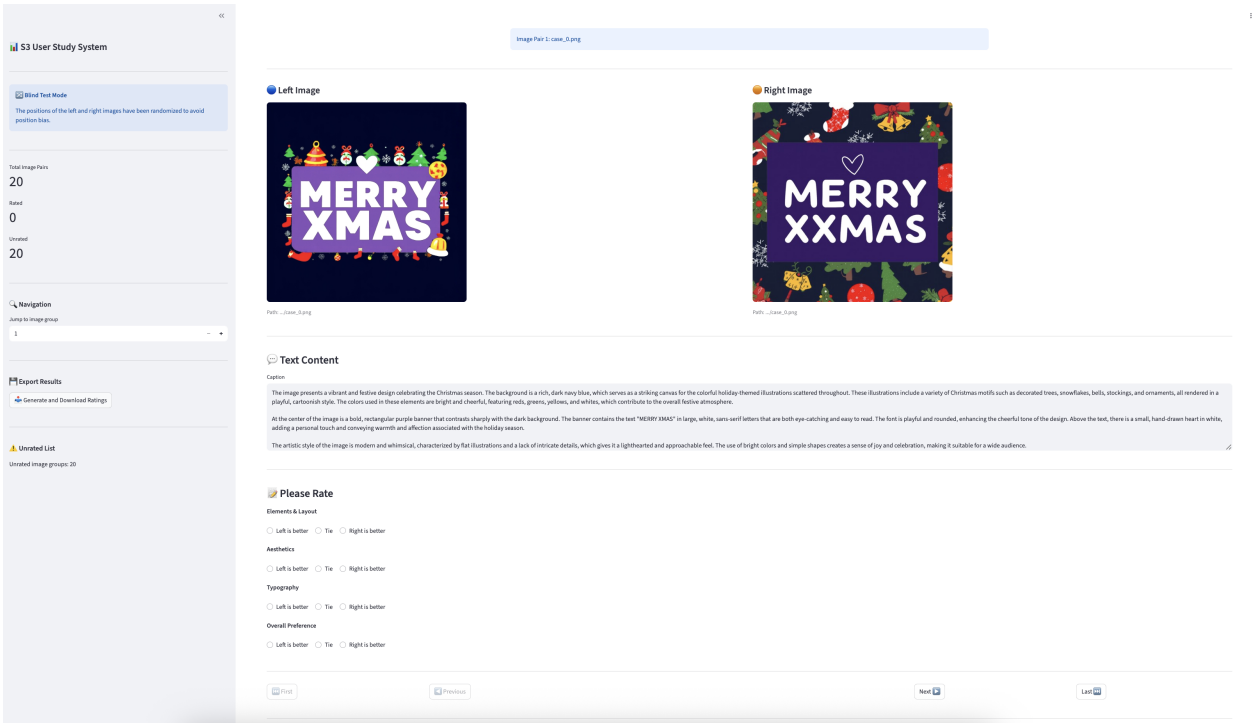


Figure 40. **User study interface for image-to-layers evaluation.** The reference input image is displayed at the center with decomposition results from two methods shown on both sides. Participants evaluate based on three metrics: granularity, layer integrity, and layer quality.