ART: Anonymous Region Transformer for Variable Multi-Layer Transparent Image Generation

Supplementary Material

1. Detailed List of Prompts and Anonymous Region Layouts

Tables 4 to 6 illustrate the detailed global prompts and anonymous layouts used in Figure 5 and Figure 6 of the main paper, respectively. In the first two rows of Table 6, we select the global prompts based on the intentions outlined in the DESIGNINTENTION benchmark for fair comparisons.

Table 7 and Table 8 illustrate the detailed instructions used in our user study on the PHOTO-MULTI-LAYER-BENCH benchmark and DESIGN-MULTI-LAYER-BENCH benchmark, respectively.

2. Analyzing the Conflicts within Semantic Layouts

As mentioned in the main paper, we observe lower coherence in the generated multi-layer images with the semantic layout approach. First, we present some typical results in Figure 1, marking the inconsistent regions between the predicted global reference image and the merged global image. Second, we visualize the attention maps between the regional visual tokens (as Query) and the combination of the region-caption text tokens and the global visual tokens from the global reference image (as Key and Value).

We observe that the visual tokens of each region primarily attend to the region-wise prompts while relying less on the predicted reference image, resulting in less coherent outputs. The purpose of predicting the global reference image is to ensure coherence across different layers. We infer that the essential reasons behind the conflict between the global reference image and the region-wise prompts stem from the disparity between the region-wise prompts and the global prompts, as *there exists a non-trivial gap between the global prompt and the region prompt associated with the same regional crop.*

3. Analyzing the Inferred Label Assignments within Anonymous Region Layouts

To measure the distance between the inferred label assignments and the human annotations provided by the anonymous region layout, we calculate the averaged layer-wise CLIP scores. These scores reflect whether the generated transparent layers in each anonymous region match the human-annotated ground-truth region-wise prompts by computing the CLIP scores between the regional visual features and the regional prompt text features.

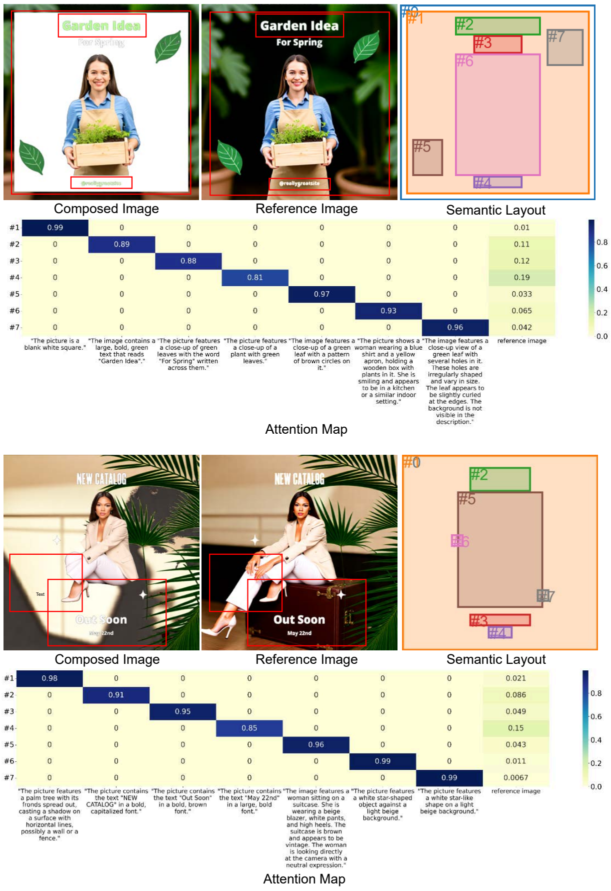


Figure 1. **Conflicts presented in Semantic Layout based Results**: We display the composed entire image in the 1st column, the reference image in the 2nd column, and the semantic layout in the 3rd column. The conflicted regions are marked with red bounding boxes in both the composed entire images and the reference images. We visualize the attention maps between semantic regions, region-wise prompts, and the global reference images.

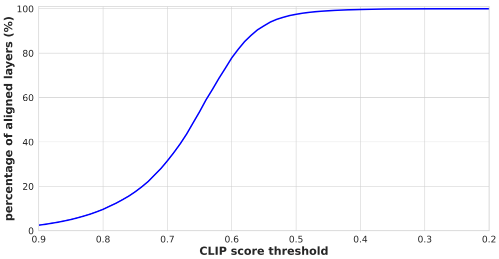


Figure 2. **Percentage of Inferred Label Assignments Matching Human Annotations**

#layer numbers	3∼8	9∼12	13∼15	16∼51
$\text{FID}_{\text{merged}}$	49.83	47.19	44.56	42.40

Table 1. Effect of different layer numbers.

#text tokens	23∼58	59∼83	84∼159	160∼272
$\text{FID}_{\text{merged}}$	27.70	26.98	28.66	28.22

Table 2. Effect of different caption length.

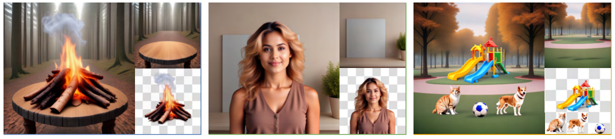


Figure 3. Multi-layer natural image generation results.

Figure 2 plots the curve of the percentage of aligned layers at different CLIP score thresholds, based on statistics from the test set consisting of 5,000 multi-layer transparent images. We attribute the alignment between the inferred label assignments from the generative model and the human annotations to Schema Theory.

4. Effect of number of transparent layers and the complexity of the scenarios described in the text.

We study whether our ART performs robustly across various input complexities by partitioning the test set into different groups according to the number of transparent layers and the number of text tokens (which reflects the complexity of the scenarios) and report the quantitative comparison results on these subsets in Table 1 and Table 2. We can see that our ART achieves even better performance with an increasing number of transparent layers and slightly weaker performance when handling longer text tokens. We attribute this to the distributions of these factors in the training set.

5. Multi-layer Natural Image Generation Results.

Our approach can be directly applied to multi-layer natural image generation without any modifications, given access to a high-quality multi-layer natural image dataset. To this end, we show that our ART achieves strong results even when fine-tuned on only a 20 curated high-quality multi-layer natural images. Figure 3 shows some qualitative results and we believe the results can continue to improve with access to more high-quality multi-layer natural images.

6. Qualitative Multi-Layer Transparent Image Generation Results with around 50 Layers

One key advantage of our approach is its ability to support the generation of tens of high-quality transparent layers from a global prompt and an ultra-dense anonymous region layout. We present the generated multi-layer image results with 40, 45, and 51 layers in Figures 4 to 6, respectively. These results highlight our method's capability to generate an *exceptionally high* number of layers, in contrast to previous works, which are limited to generating only a small number of layers.

7. Implementation of Layout Conditional Multi-Layer 3D RoPE

We present the PyTorch implementation of the proposed layout-conditional multi-layer 3D RoPE in Algorithm 1 and its usage within the Attention Module in Algorithm 2.

8. Layout Variation

One key advantage of our approach is that our Anonymous Region Transformer generalizes to various layouts given a fixed global prompt. The ART model is capable of adaptively assigning semantic concepts to fit diverse anonymous region layouts. We illustrate some qualitative results in Tables 9 to 17.

9. Layer-wise Editing

The purpose of the experiment is to demonstrate the effectiveness of the proposed ART method in enabling layer-wise image editing, specifically the accurate regeneration of contents on specific layers. The layer-wise editing pipeline consists of three steps: modifying the input prompt, regenerating the layers that need to be edited, and freezing the remaining layers. We have provided an editing result in Figure 7. As can be observed, our model can accurately regenerate specific content on the editable layers to meet the requirements from the input prompt. Moreover, the newly generated layer remains harmonious with the rest while keeping other layers unchanged, providing a feasible approach to precisely and independently control the style and contents of each layer.

10. Details of Transparency Encoding

Here, we provide additional details on the transparency encoding introduced in Section 3.1. The overall goal is to transform a 4-channel RGBA image into its 3-channel RGB counterpart, facilitating the reuse of pretrained three-channel image generation models while effectively embedding the alpha channel information into the RGB channels.

For each RGBA image $\mathbf{I}_{\text{fg}}^{i} \in \mathbb{R}^{H_i \times W_i \times 4}$, we first linearly normalize the three RGB channels $\mathbf{I}_{\text{fg,RGB}}^{i} \in \mathbb{R}^{H_i \times W_i \times 3}$ from the range $[0, 255]$ to $[-1, 1]$, following the standard practice in Flux.1 models. Similarly, we linearly transform the alpha channel $\mathbf{I}_{\text{fg},\alpha}^{i} \in \mathbb{R}^{H_i \times W_i \times 1}$ from $[0, 255]$ to $[-1, 1]$, where -1 represents fully transparent pixels and 1 represents fully opaque pixels.

To encode transparency information from the alpha channel into the RGB channels, we apply the following transformation:

$$\hat{\mathbf{I}}^{i}_{\text{fg}} = (0.5\mathbf{I}^{i}_{\text{fg},\alpha} + 0.5) \times \mathbf{I}^{i}_{\text{fg,RGB}}.$$

Here, the coefficient $(0.5\mathbf{I}^{i}_{\text{fg},\alpha} + 0.5)$ linearly maps the alpha channel from $[-1, 1]$ to $[0, 1]$. This ensures that the RGB values of fully opaque pixels remain unchanged, while fully transparent pixels are mapped to pure gray (RGB = $(0, 0, 0)$ in the $[-1, 1]$ range). Semi-transparent pixels undergo a proportional transformation based on their alpha values.

11. Evaluation in text generation

Method	$\text{PSNR}^{\text{layer}}_{\text{rgb}}$	$\text{PSNR}^{\text{layer}}_{\text{alpha}}$	PSNR	$\text{FID}_{\text{merged}}$
Single-layer Autoencoder w/ CNN	30.10	20.12	26.88	5.12
Single-layer Autoencoder w/ ViT	33.64	22.47	28.76	3.39
Multi-layer Autoencoder w/ ViT	**34.80**	**24.25**	**31.37**	**2.76**

Table 3. Ablation of autoencoder (all trained with our MLTD data).

Here we provide more evaluation for the advantages of our multi-layer transparent image autoencoder, which has been previously illustrated in Figure 8. The images are generated by encoding and decoding the same ground-truth image, which effectively reflects the quality of the reconstructed multi-layer images. The superior performance in text generation of our method can be attributed to the following key factors: (1) the use of Vision Transformer (ViT) for visual text modeling, which outperforms CNN-based autoencoders by predicting more accurate edges. In contrast, both LayerDiffuse and Flux-RGBA rely on CNN-based autoencoders; (2) the multi-layer autoencoder architecture, which enables explicit interactions across different layers by jointly encoding and decoding them, leading to better performance compared to single-layer methods. Additionally, our results benefit from the multi-layer transparent design dataset (MLTD), which includes a larger number of visual text layers. As shown in Table 3, replacing CNN with ViT and adopting a multi-layer structure both contribute to improved performance.

12. Future works

We believe our work lays a solid foundation for the next generation of generative models that can produce a variable number of coherent transparent layers and support flexible image editing through layer compositing. Looking forward, we identify several promising directions for future research: (i) *Enhancing Visual Aesthetics*: A key challenge is to improve the visual appealing of the generated transparent layers and ensure that the composite images achieve parity with those produced by state-of-the-art text-to-image models such as FLUX. (ii) *Anonymous Region Layouts*: We anticipate that leveraging anonymous region layouts will transform conventional layout-to-image generation tasks. This approach has the potential to eliminate the need for complex regional prompt annotations and to simplify the modeling process by granting models greater control. (iii) *Human Interaction with ART*: We also see great promise in integrating user requirements into the multi-layer image generation system. Future work could explore interactive methods for incorporating real-time user feedback, enabling dynamic refinement of generated layers and more personalized editing workflows.

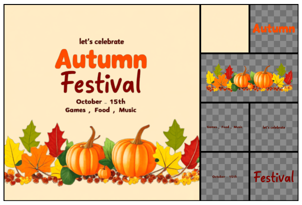
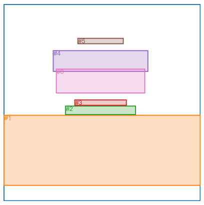
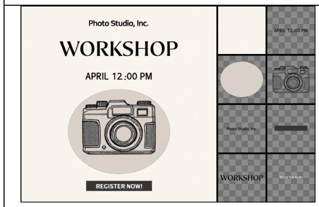
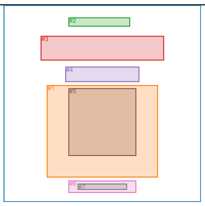
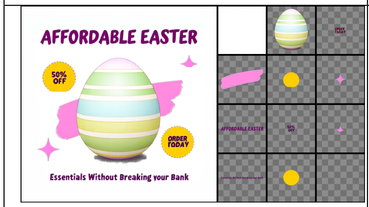
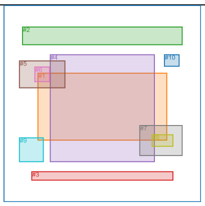
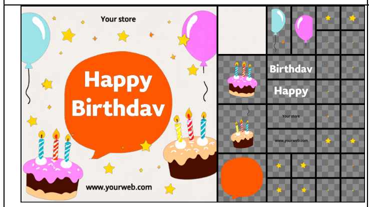
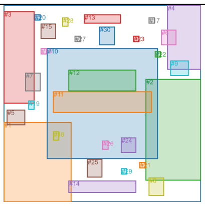
Multi-layer Transparent Image	Anonymous Region Layout	Global Prompt
 let's celebrate Autumn Festival October . 15th Games , Food , Music		The image is a poster for an Autumn Festival. The festival is scheduled to take place from October 15th to October 21st. The poster features a variety of autumn-themed elements, including pumpkins, leaves, and berries. The text on the poster is in a playful, handwritten font, and it reads ”let’s celebrate Autumn Festival”. The poster also includes a list of activities that will be available at the festival, such as games, food, and music. The overall color scheme of the poster is warm, with shades of orange, yellow, and green, which are typical colors associated with autumn. The poster is designed to be eye-catching and inviting, encouraging people to come and enjoy the festival.
 Photo Studio, Inc. WORKSHOP APRIL 12:00 PM REGISTER NOW!		A promotional flyer for a photography workshop hosted by Photo Studio, Inc. on April 12 at 9:00 AM. It features a vintage camera illustration and a ”Register Now!” button at the bottom.
 AFFORDABLE EASTER 50% OFF ORDER TODAY Essentials Without Breaking your Bank		A promotional Easter-themed graphic featuring a large, colorful egg with text ”AFFORDABLE EASTER” at the top. It includes discount badges stating ”50% OFF” and ”ORDER TODAY” on either side of the egg, with the tagline ”Essentials Without Breaking the Bank” at the bottom.
		A festive birthday card design features an orange speech bubble with ”Happy Birthday” text in white, surrounded by balloons, stars, and cakes with candles. The top reads ”Your store” and the bottom displays ”www.yourweb.com”. The background is light with playful elements creating a cheerful vibe.

Table 4. Detailed anonymous region layouts and global prompts for multi-layer image generation in Figure 5 of the main paper.

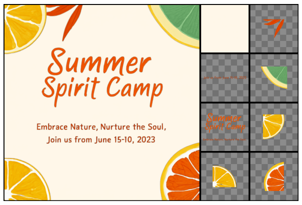
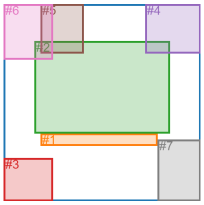
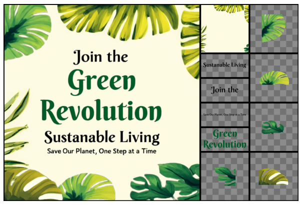
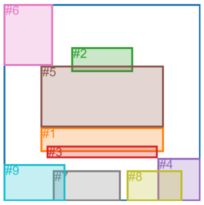
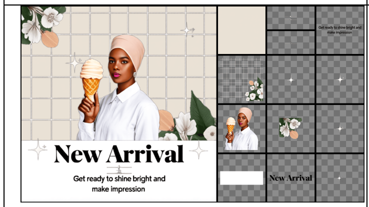
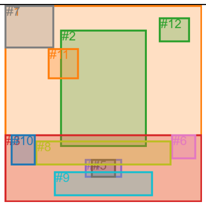
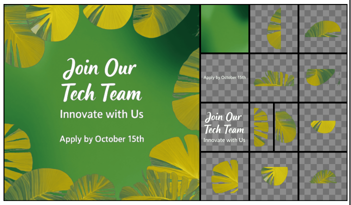
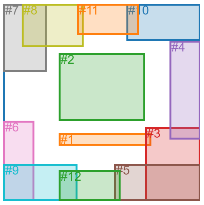
Multi-layer Transparent Image	Anonymous Region Layout	Global Prompt
 Summer Spirit Camp Embrace Nature, Nurture the Soul, Join us from June 15-10, 2023		The image is a romantic and spiritual graphic design, likely intended for a summer camp brochure. The overall design showcases a citrus-inspired palette, featuring vibrant oranges, yellows, and soft greens, which enhances its sophisticated and refreshing atmosphere. Styled in ornamental calligraphy, the design features seamless patterns that evoke a sense of harmony and continuity, appealing to fashion-forward thinkers who appreciate intricate details. The title, ”Summer Spirit Camp,” is written in a Brush font, with a size of 95px. Positioned at the top of the image, it is bold and immediately captures the viewer’s attention, setting a tone of elegance and anticipation. Below the title, a secondary text reads ”Embrace Nature, Nurture the Soul,” sized at 100px. This text complements the main title by highlighting the camp’s core values, inviting viewers to explore a deeper connection with nature and spirituality. At the bottom, another piece of text states ”Join us from June 10-15, 2023,” written in a smaller 100px font. This serves as supporting information, providing essential logistics such as the date, ensuring clarity and accessibility for potential attendees. The text content in this design is specific and directly contributes to the overall purpose of the graphic, effectively conveying the essence of the summer camp experience. This design captures an artistic and creative spirit, making it both visually striking and emotionally resonant. The seamless integration of text and imagery creates a cohesive narrative that resonates with the intended audience, inviting them to embark on a transformative journey.
 Join the Green Revolution Sustanable Living Save Our Planet, One Step at a Time		The image is a romantic and polished graphic design, likely intended for an environmental campaign. The overall design showcases a tropical greens and yellows color scheme that enhances its ornamental atmosphere. Styled in hand-drawn doodles, the design features whimsical hand-lettering, adding to its appeal for luxury consumers. This design captures a thoughtful and balanced aesthetic, making it both visually striking and emotionally resonant. Text elements play a crucial role in conveying the message. The title reads ”Join the Green Revolution,” written in a Condensed serif font, with a size of 88px. Positioned at the top of the image, it is bold and immediately captures the viewer’s attention. Below the title, a secondary text reads ”Sustainable Living,” providing additional information and complementing the main title. This text is sized at 24px, maintaining a harmonious balance with the title. At the bottom, another piece of text states ”Save Our Planet, One Step at a Time,” written in a smaller 18px font. This serves as supporting information, encouraging action and engagement. The text content in this design is specific and directly contributes to the overall purpose of the graphic. The title announces the campaign’s mission, the secondary text highlights the theme, and the footer provides a motivational call to action. The combination of tropical colors, hand-drawn elements, and carefully chosen typography creates a design that is both visually appealing and emotionally impactful, resonating with an audience passionate about environmental sustainability.
 New Arrival Get ready to shine bright and make impression		A promotional graphic features a person in a white outfit and headscarf holding an ice cream cone, with the text ”New Arrival” and ”Get ready to shine bright and make impression” on a stylish, modern background with grid and floral elements.
 Join Our Tech Team Innovate with Us Apply by October 15th	 #7 #8 #11 #10 #2 #4 #6 #1 #3 #9 #12 #5	The image is a painterly and soft graphic design, likely intended for a job recruitment poster. The overall design showcases a tropical greens and yellows color scheme, enhancing its structured yet inviting atmosphere. Styled with motion blur visuals, the design features repeating motifs that add a dynamic appeal, making it particularly attractive for tech companies. This design captures a timeless yet modern aesthetic, making it both visually striking and emotionally resonant. Text elements play a crucial role in conveying the message. The title reads ”Join Our Tech Team,” written in an Italic Serif font, with a size of 20px. Positioned at the top of the image, it is bold and immediately captures the viewer’s attention. Below the title, a secondary text reads ”Innovate with Us,” sized at 72px, providing additional information and complementing the main title. At the bottom, another piece of text states ”Apply by October 15th,” written in a smaller 36px font. This serves as supporting information, specifying logistics like a deadline. The text content in this design is specific and directly contributes to the overall purpose of the graphic. The title announces the recruitment opportunity, the secondary text highlights the company’s mission, and the footer provides essential application details. This cohesive blend of design elements and text creates a compelling invitation for potential candidates, evoking a sense of excitement and opportunity.

Table 5. Detailed anonymous region layouts and global prompts for multi-layer image generation in Figure 5 of the main paper.

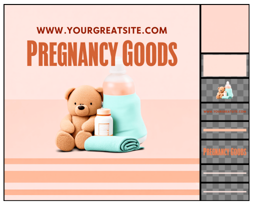
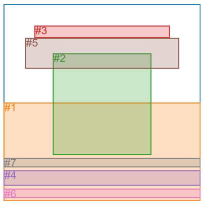
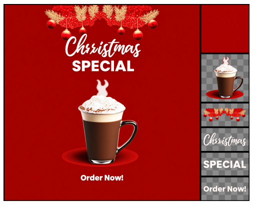
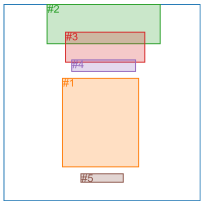
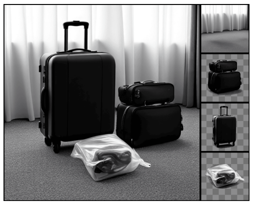
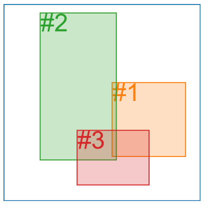
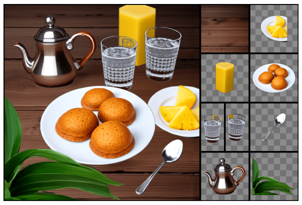
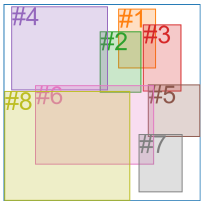
Multi-layer Transparent Image	Anonymous Region Layout	Global Prompt
		The image is designed as a Facebook cover for a website that specializes in selling pregnancy goods. The theme is warm and inviting, geared towards expecting parents. The background features a soft pastel palette, predominantly in shades of baby pink and light blue, which are colors commonly associated with babies and pregnancy. Arranged throughout the image are a selection of baby essentials, which may include items like a plush teddy bear, a set of pastel-colored baby clothes, a small stack of diapers, a baby bottle, and a swaddle blanket. These items are artistically placed to create an impression of luxury and care, suggesting that the website offers a premium selection of products. Prominently displayed within the design is a bold, attractive advertisement for a 15% off discount. This text is strategically positioned to catch the viewer's attention without overshadowing the curated display of goods. The text is written in soft, rounded font to maintain a gentle and friendly aesthetic. In one of the bottom corners, the website URL, 'www.yourgreatsite.com', is included in a clear font for easy readability. The overall effect of the design is comforting and welcoming, aiming to attract expecting parents to explore the website's offerings further. This Facebook cover is effectively tailored to appeal to the needs and desires of its target audience.
		The image is designed as an Instagram story promoting a special Christmas offer for a chocolate drink. The background of the story features a cozy, festive theme with a warm and inviting color scheme, primarily consisting of rich browns and deep reds, reminiscent of hot chocolate and Christmas decor. Centered in the image is a steaming cup of chocolate drink, garnished with a sprinkle of cocoa powder and a cinnamon stick, suggesting warmth and indulgence. To enhance the festive atmosphere, there are elements such as small evergreen branches, a scattering of red berries, and a few decorative golden bells placed around the cup. The text on the story is bold and eye-catching, starting with 'Christmas Special' in elegant white script at the top. Below this, the details of the offer are highlighted in bright red, stating '20% OFF' to capture attention. Further down, the call to action 'Order Now!' is displayed in bold white letters, encouraging viewers to take immediate advantage of the offer. The overall style of the image is cozy and appealing, designed to evoke a sense of the holiday spirit and entice customers to enjoy a delicious chocolate drink during the Christmas season. The aesthetic is suited to engage viewers on social media, making the offer both attractive and memorable.
		The image shows a collection of luggage items on a carpeted floor. There are three main pieces of luggage: a large suitcase, a smaller suitcase, and a duffel bag. The large suitcase is positioned in the center, with the smaller suitcase to its left and the duffel bag to its right. The luggage appears to be packed and ready for travel. In the foreground, there is a plastic bag containing what looks like a pair of shoes. The background features a white curtain, suggesting that the setting might be indoors, possibly a hotel room or a similar temporary accommodation. The image is in black and white, which gives it a timeless or classic feel.
		The image shows a rustic wooden table setting with a variety of items. On the table, there is a plate with six golden-brown, round, hollow pastries, which appear to be madeleines. To the left of the plate, there is a silver teapot with a wooden handle and spout. Next to the teapot, there are three glasses with different designs, filled with a clear liquid, possibly water. To the right of the plate, there is a small white plate with slices of yellow fruit, which could be pineapple. In the foreground, there is a green plant with broad leaves, and a silver spoon is placed on the table. The overall setting suggests a cozy, inviting atmosphere, possibly for a tea or dessert time.

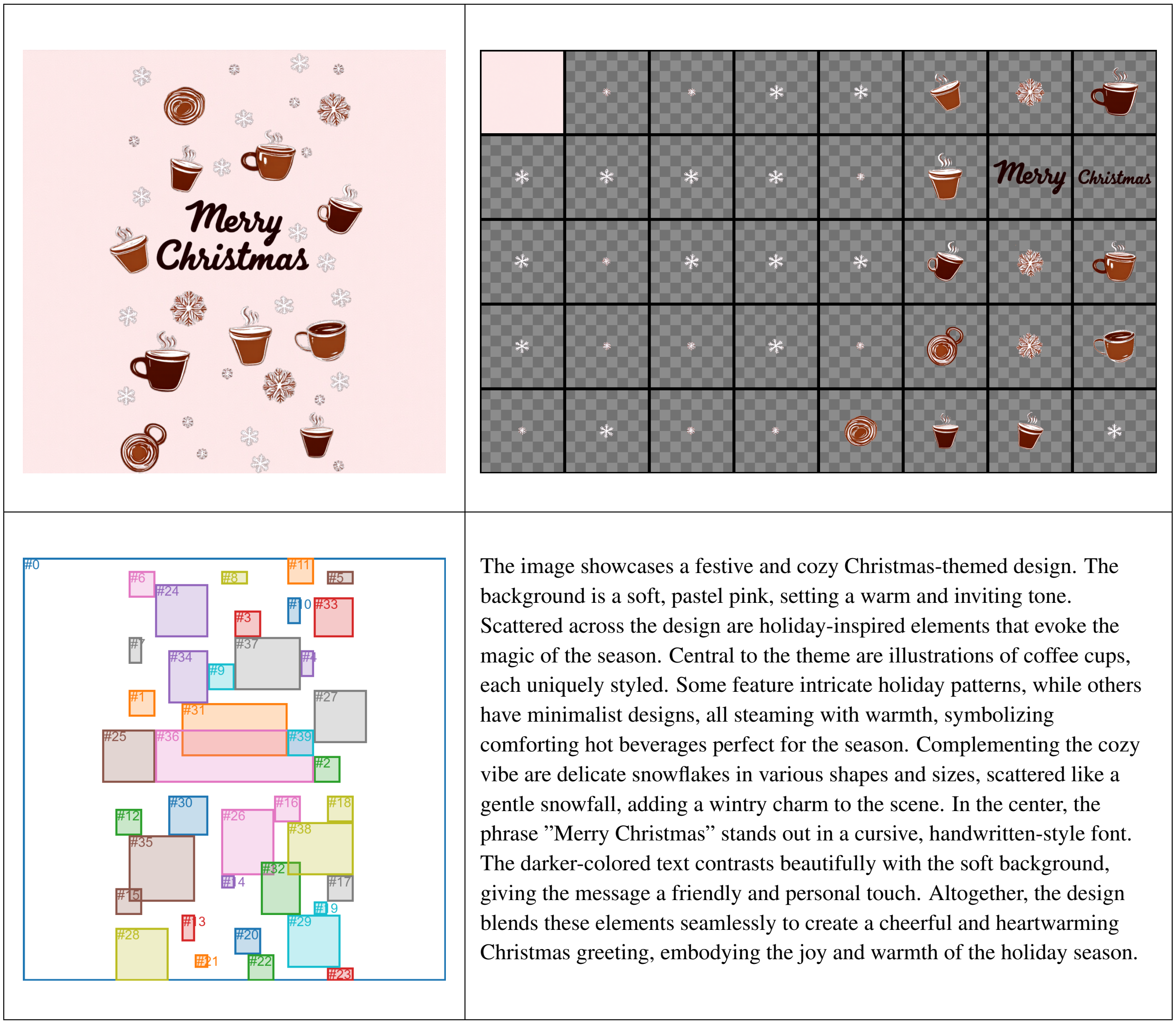
Table 6. Detailed anonymous region layouts and global prompts for multi-layer image generation in Figure 6 of the main paper.

Metrics	Detailed Instruction
Aesthetics	Please evaluate the overall visual appeal of the images. Consider which method produces more visually pleasing and attractive results. Focus on the artistic quality, color harmony, and whether the style matches design aesthetics.
Typography	Please assess the text quality in the generated images. Check if the text is clear, readable and accurately rendered without distortions. Evaluate whether the font style, size and spacing are appropriate, and if the text matches the intended content.
Harmonization	Please examine the harmony of layers around the merged image. Consider whether the transitions between layers are smooth and natural, and if the layer effects enhance the overall visual quality without looking artificial.
Layout	Please evaluate the overall composition and arrangement. Check if text and graphic elements are well-balanced and properly aligned. Consider whether the spacing is appropriate, elements are organized logically, and if there are any awkward overlaps or conflicts between components.

Table 7. Detailed Instructions for the User Study on the DESIGN-MULTI-LAYER-BENCH

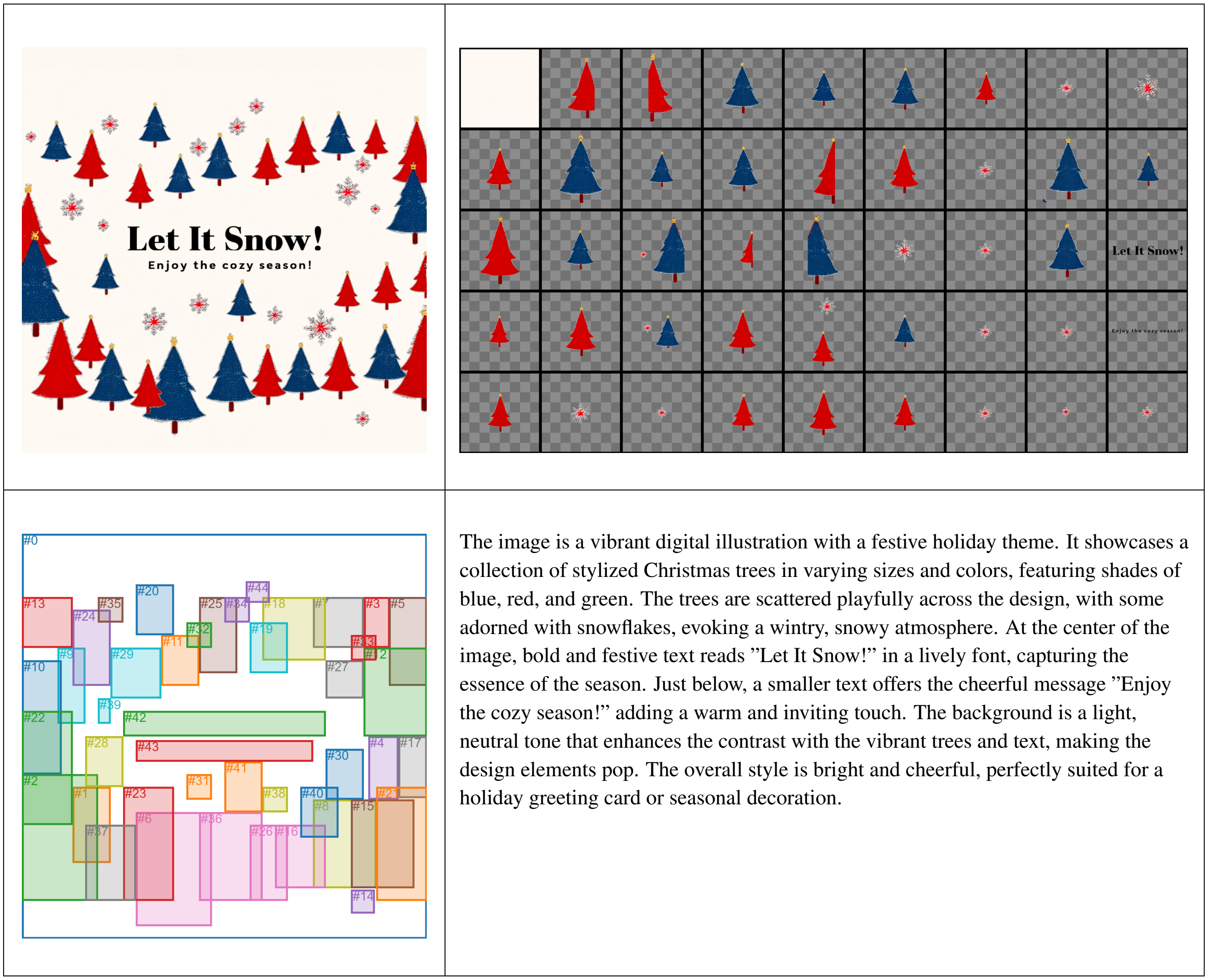
Metrics	Description
Aesthetics	Please evaluate the visual appeal of the generated images. Consider which result looks more visually pleasing and artistically satisfying. Focus on the overall aesthetic quality and visual attractiveness of the designs.
PromptFollow	Please assess how well each generated image matches the given text prompt. Compare the results and determine which method better captures and reflects the requirements specified in the prompt text.
Harmonization	Please examine the visual consistency and smoothness between different layers, particularly focusing on the transitions at the right and bottom edges. Consider whether the layer blending appears natural and well-integrated.

Table 8. Detailed Instructions for the User Study on the PHOTO-MULTI-LAYER-BENCH



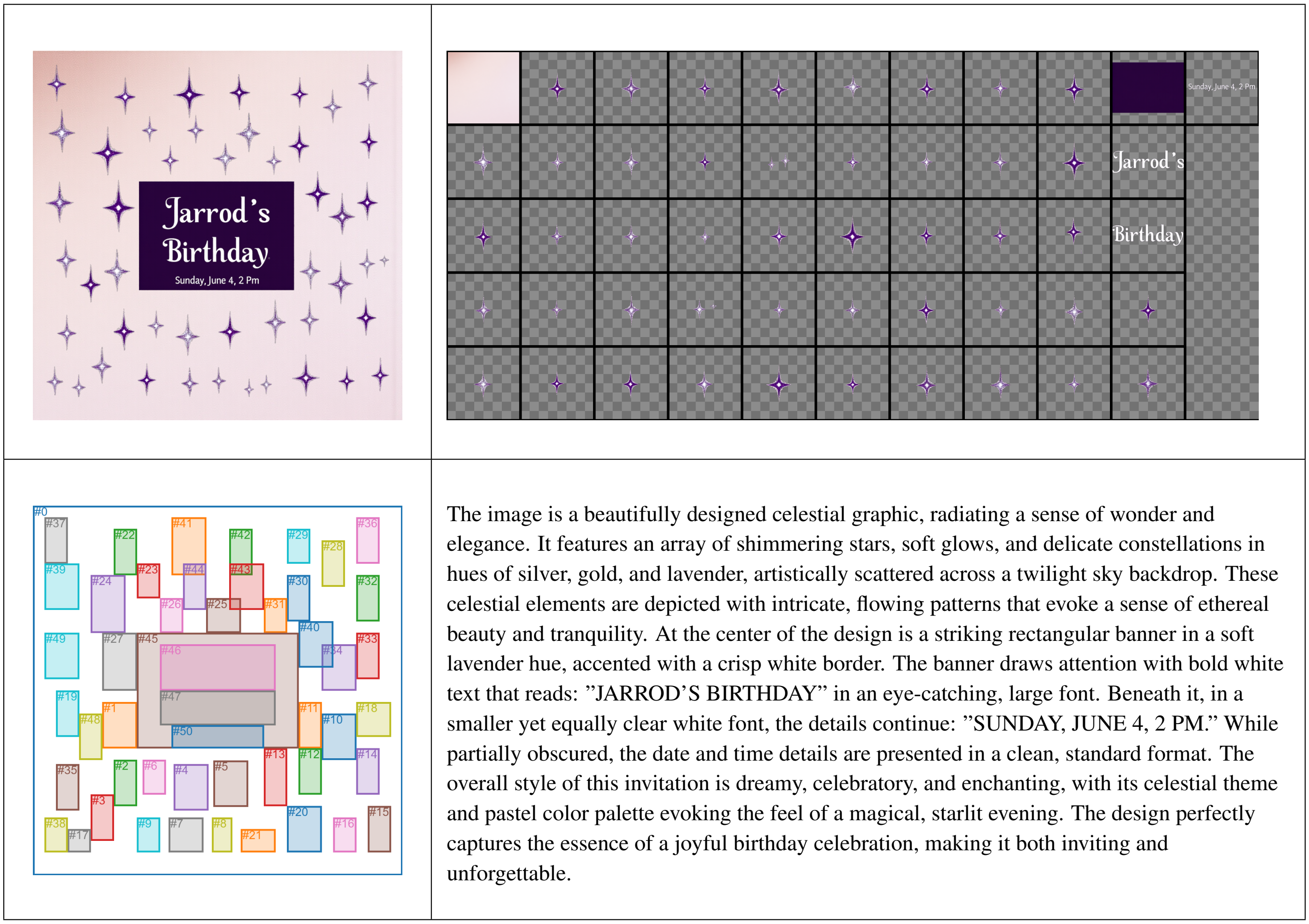
The image showcases a festive and cozy Christmas-themed design. The background is a soft, pastel pink, setting a warm and inviting tone. Scattered across the design are holiday-inspired elements that evoke the magic of the season. Central to the theme are illustrations of coffee cups, each uniquely styled. Some feature intricate holiday patterns, while others have minimalist designs, all steaming with warmth, symbolizing comforting hot beverages perfect for the season. Complementing the cozy vibe are delicate snowflakes in various shapes and sizes, scattered like a gentle snowfall, adding a wintry charm to the scene. In the center, the phrase ”Merry Christmas” stands out in a cursive, handwritten-style font. The darker-colored text contrasts beautifully with the soft background, giving the message a friendly and personal touch. Altogether, the design blends these elements seamlessly to create a cheerful and heartwarming Christmas greeting, embodying the joy and warmth of the holiday season.

Figure 4. Generated Result with 40 transparent image layers. Top-left: Generated Merged Image; Top-Right: Generated Transparent Layers; Bottom-left: Anonymous Region Layout; Bottom-right: Global Prompt.



The image is a vibrant digital illustration with a festive holiday theme. It showcases a collection of stylized Christmas trees in varying sizes and colors, featuring shades of blue, red, and green. The trees are scattered playfully across the design, with some adorned with snowflakes, evoking a wintry, snowy atmosphere. At the center of the image, bold and festive text reads ”Let It Snow!” in a lively font, capturing the essence of the season. Just below, a smaller text offers the cheerful message ”Enjoy the cozy season!” adding a warm and inviting touch. The background is a light, neutral tone that enhances the contrast with the vibrant trees and text, making the design elements pop. The overall style is bright and cheerful, perfectly suited for a holiday greeting card or seasonal decoration.

Figure 5. Generated Result with 45 transparent image layers. Top-left: Generated Merged Image; Top-Right: Generated Transparent Layers; Bottom-left: Anonymous Region Layout; Bottom-right: Global Prompt.



The image is a beautifully designed celestial graphic, radiating a sense of wonder and elegance. It features an array of shimmering stars, soft glows, and delicate constellations in hues of silver, gold, and lavender, artistically scattered across a twilight sky backdrop. These celestial elements are depicted with intricate, flowing patterns that evoke a sense of ethereal beauty and tranquility. At the center of the design is a striking rectangular banner in a soft lavender hue, accented with a crisp white border. The banner draws attention with bold white text that reads: "JARROD'S BIRTHDAY" in an eye-catching, large font. Beneath it, in a smaller yet equally clear white font, the details continue: "SUNDAY, JUNE 4, 2 PM." While partially obscured, the date and time details are presented in a clean, standard format. The overall style of this invitation is dreamy, celebratory, and enchanting, with its celestial theme and pastel color palette evoking the feel of a magical, starlit evening. The design perfectly captures the essence of a joyful birthday celebration, making it both inviting and unforgettable.

Figure 6. Generated Result with 51 transparent image layers. Top-left: Generated Merged Image; Top-Right: Generated Transparent Layers; Bottom-left: Anonymous Region Layout; Bottom-right: Global Prompt.

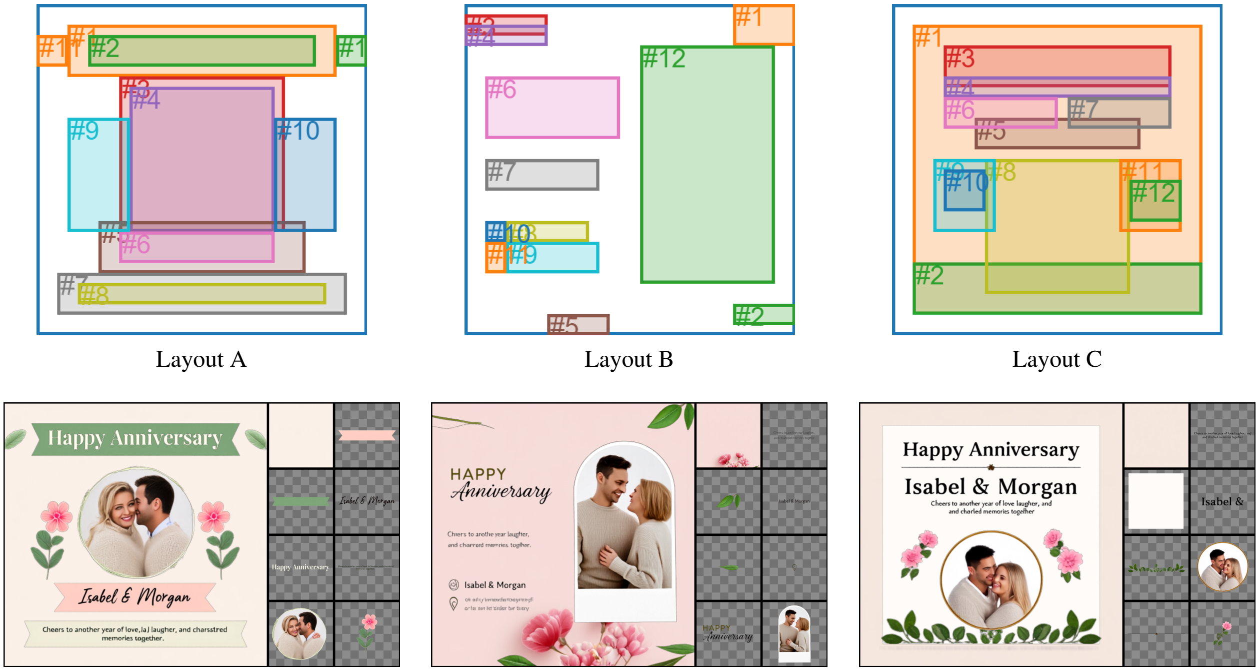
Algorithm 1: Layout Conditional Multi-Layer 3D-RoPE

```
import torch

def get_1d_rotary_pos_embed(dim, pos, theta=10000.0):
    # dim: Dimension of the frequency tensor.
    # pos: Position indices for the frequency tensor. Shape: [S]
    # theta: Scaling factor for frequency computation.

    freqs = 1.0 / (theta ** (torch.arange(0, dim, 2)[:(dim // 2)] / dim))
    freqs = torch.outer(pos, freqs)
    freqs_cos = freqs.cos().repeat_interleave(2, dim=1)
      freqs_sin = freqs.sin().repeat_interleave(2, dim=1)

    return freqs_cos, freqs_sin

def get_3d_rotary_pos_embed(ids, axes_dim = (16, 56, 56)):
    # ids: 3D position indices of visual tokens. Shape: [S, 3]
    # axes_dim: RoPE dimensions for each axis.

    cos_out = []
    sin_out = []
    for i in range(3):
        cos, sin = get_1d_rotary_pos_embed(axes_dim[i], ids[:, i])
        cos_out.append(cos)
        sin_out.append(sin)
    freqs_cos = torch.cat(cos_out, dim=-1)
    freqs_sin = torch.cat(sin_out, dim=-1)

    return freqs_cos, freqs_sin

def prepare_latent_image_ids(height, width, list_layer_box):
    # height: The height of the image latent.
    # width: The width of the image latent.
    # list_layer_box: List of bounding boxes in each layer.

    ids_list = []
    for layer_idx, layer_box in enumerate(list_layer_box):
        ids = torch.zeros(height//2, width//2, 3)
        ids[..., 0] = layer_idx # use the first axis to distinguish layers
        ids[..., 1] = ids[..., 1] + torch.arange(height//2)[:, None]
        ids[..., 2] = ids[..., 2] + torch.arange(width//2)[None, :]

        x1, y1, x2, y2 = layer_box
        ids = ids[y1:y2, x1:x2, :]
        ids = ids.reshape(-1, ids.shape[-1])
        ids_list.append(ids)
    latent_image_ids = torch.cat(ids_list, dim=0)

    return flatent_image_ids
```

Algorithm 2: Layout Conditional Multi-Layer 3D-RoPE within Attention Module

```python
import torch
import torch.nn as nn
import torch.nn.functional as F

def apply_rotary_pos_embed(x, freqs_cis):
    # x: Query or key tensor to apply rotary embeddings. Shape: [B, H, S, Dh]
    # freqs_cis: Precomputed frequency tensor for complex exponentials. Shape: [S, Dh]

    cos, sin = freqs_cis
    cos = cos[None, None]
    sin = sin[None, None]
    x_real, x_imag = x.reshape(*x.shape[:-1], -1, 2).unbind(-1)
    x_rotated = torch.stack([-x_imag, x_real], dim=-1).flatten(3)
    out = x.float() * cos + x_rotated.float() * sin

    return out

class AttentionProcessor(nn.module):
    to_q: nn.Linear
    to_k: nn.Linear
    to_v: nn.Linear
    to_out: nn.Linear
    def __call__(self, hidden_states, image_rotary_emb):
        # hidden_states: Input hidden states of the block. # [B, S, D]
        # image_rotary_emb: Precomputed 3D-RoPE frequency tensor. # [S, Dh]

        query = self.to_q(hidden_states)
        key = self.to_k(hidden_states)
        value = self.to_v(hidden_states)

        ...

        query = apply_rotary_pos_embed(query, image_rotary_emb)
        key = apply_rotary_pos_embed(key, image_rotary_emb)
        hidden_states = F.scaled_dot_product_attention(query, key, value, is_causal=False)
        hidden_state = self.to_out(hidden_states)

        ...

        return hidden_states
```

Prompt: The image is a graphic design with a celebratory theme. At the top, there is a banner with the text ”Happy Anniversary” in a bold, sans-serif font. Below this banner, there is a circular frame containing a photograph of a couple. The man has short, dark hair and is wearing a light-colored sweater, while the woman has long blonde hair and is also wearing a light-colored sweater. They are both smiling and appear to be embracing each other. Surrounding the circular frame are decorative elements such as pink flowers and green leaves, which add a festive touch to the design. Below the circular frame, there is a text that reads ”Isabel & Morgan” in a cursive, elegant font, suggesting that the couple’s names are Isabel and Morgan. At the bottom of the image, there is a banner with a message that says ”Happy Anniversary! Cheers to another year of love, laughter, and cherished memories together.” This text is in a smaller, sans-serif font and is placed against a solid background, providing a clear message of celebration and well-wishes for the couple. The overall style of the image is warm and celebratory, with a color scheme that includes shades of pink, green, and white, which contribute to a joyful and romantic atmosphere.


#1
#2
#3
#4
#5
#6
#7
#8
#9
#10
#11
#12
Happy Anniversary
Isabel & Morgan
Cheers to another year of love,la) laughter, and charsstred memories together.
HAPPY
Anniversary
Isabel & Morgan
Happy Anniversary
Isabel & Morgan
Cheers to another year of love laughter, and and charled memories together

Layout A

Layout B

Layout C

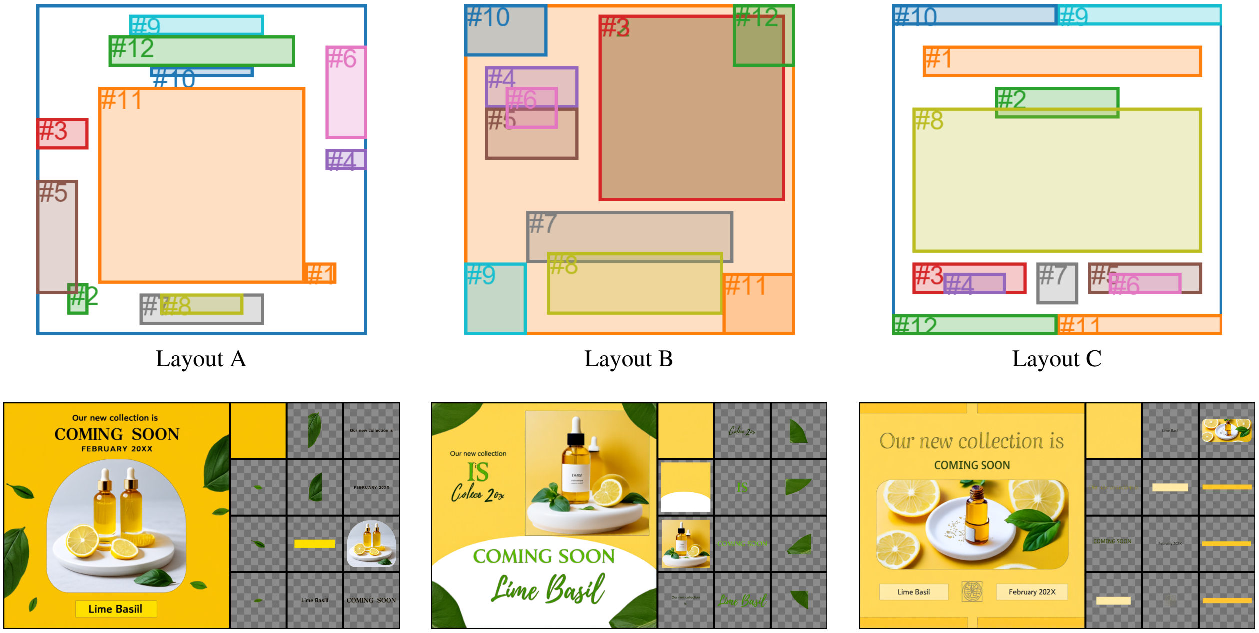
Generated A

Generated B

Generated C

Table 9. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index ‘#0’ is omitted) at the second row and the generated results at the last row. (Case 1)

Prompt: The image is a promotional graphic for a new collection that is coming soon in February 20xx. The central focus of the image is a collection of items that suggest a theme of natural beauty and freshness. There are two bottles of what appears to be a yellow-colored liquid, possibly a fragrance or essential oil, given their shape and the context. The bottles are placed on a white, oval-shaped surface that resembles a soap or a decorative plate. Surrounding the bottles are slices of lemon, which are scattered around the surface, adding a citrus element to the composition. There are also green leaves, possibly basil, which are placed near the lemon slices, contributing to the natural and fresh theme. The background is a solid, warm yellow color that complements the overall color scheme of the image. At the top of the image, there is text that reads ”Our new collection is COMING SOON FEBRUARY 20xx,” indicating the time frame for the release of the new collection. At the bottom, the text ”Lime Basil” is visible, which likely refers to the scent or flavor of the items in the collection. The overall style of the image is clean, modern, and designe5d to evoke a sense of anticipation for the new collection.



Layout A

Layout B

Layout C

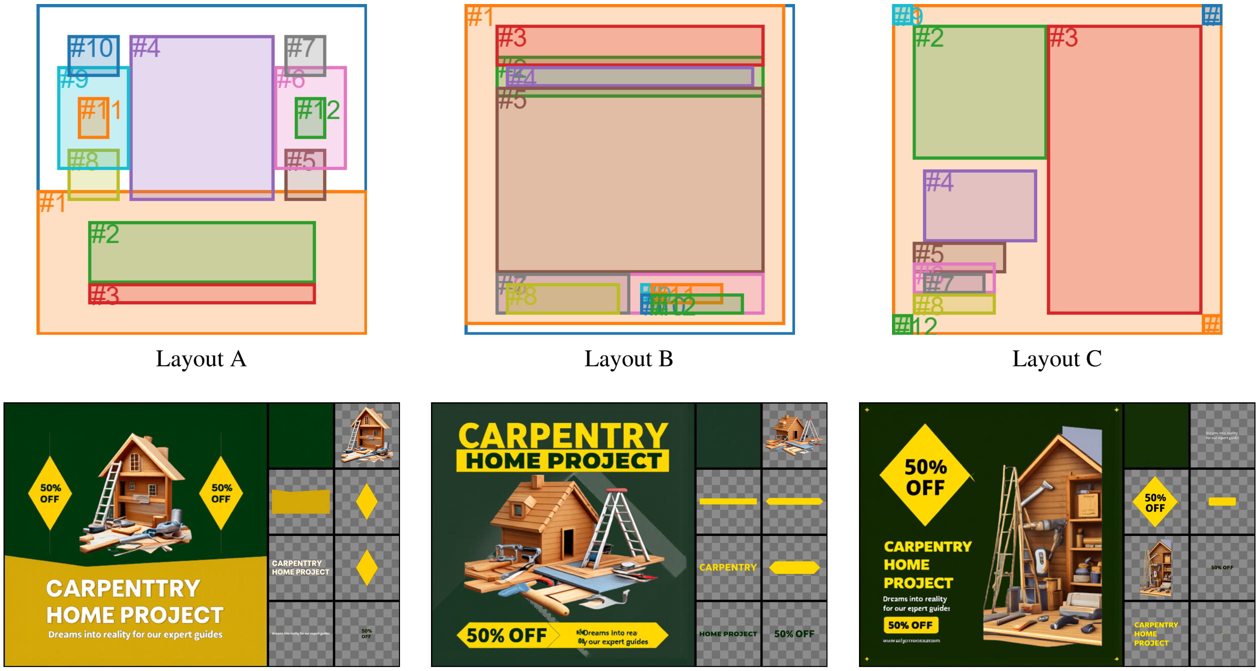
Generated A

Generated B

Generated C

Table 10. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index ‘#0’ is omitted) at the second row and the generated results at the last row. (Case 2)

Prompt: The image features a stylized graphic of a carpentry home project. At the center, there is a three-dimensional illustration of a wooden house with a visible interior. The house is filled with various carpentry tools and materials, such as a ladder, a hammer, a saw, a measuring tape, a paint roller, and a paint tray. These items are arranged to suggest that they are being used for a home renovation or construction project. The background of the image is a dark green color, and there are two yellow diamonds on either side of the house, each containing the text "50% OFF." This suggests that there is a discount offer associated with the carpentry home project. At the bottom of the image, there is a bold text that reads "CARPENTRY HOME PROJECT" in capital letters, indicating the theme of the image. Below this main title, there is a tagline that says "Dreams into reality with our expert guides," which implies that the image is likely an advertisement or promotional material for a service or product related to carpentry and home projects. The overall style of the image is clean and modern, with a clear focus on the carpentry theme and the promotional offer. The use of bright colors and bold text is designed to attract attention and convey the message of the advertisement effectively.



Layout A

Layout B

Layout C

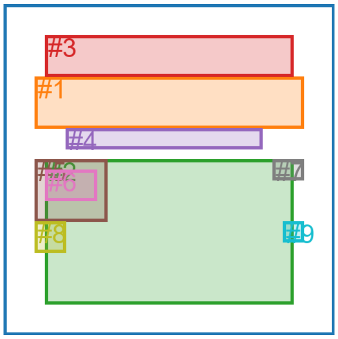
Generated A

Generated B

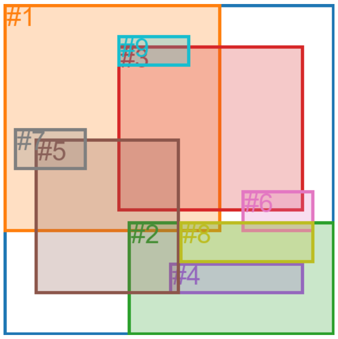
Generated C

Table 11. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index '#0' is omitted) at the second row and the generated results at the last row. (Case 3)

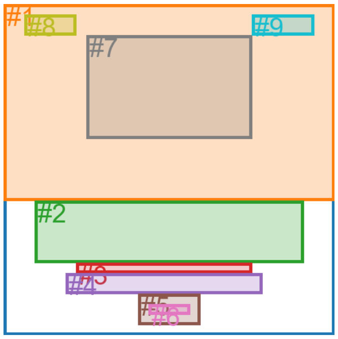
Prompt: The image features a graphic design with a stylized illustration of an urban landscape. The illustration includes various buildings of different shapes and sizes, some with red roofs, and a few trees. The buildings are depicted in a simplified manner, with flat colors and minimal detail, giving the image a modern and clean aesthetic. At the top of the image, there is text that reads ”Urban Vision Architects” in bold, capital letters. Below this, in a smaller font, it says ”Innovative architectural solutions.” To the right of the text, there is a graphic element resembling a star or a sun with rays emanating from it. In the lower left corner, there is a discount offer indicated by the text ”15% OFF” in a bold, sans-serif font. The overall style of the image suggests it could be an advertisement or promotional material for an architectural firm. The color palette is limited, with a dominant beige background that contrasts with the red and black elements of the illustration and text.



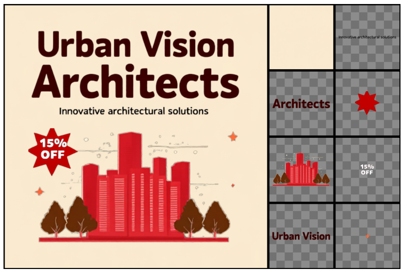
Layout A



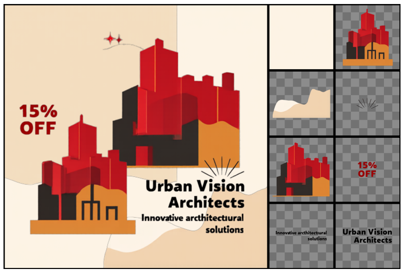
Layout B



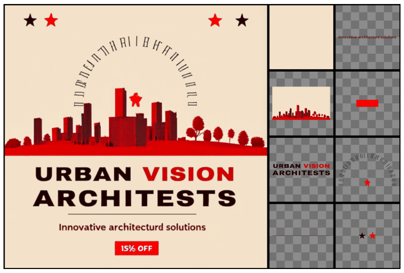
Layout C



Generated A



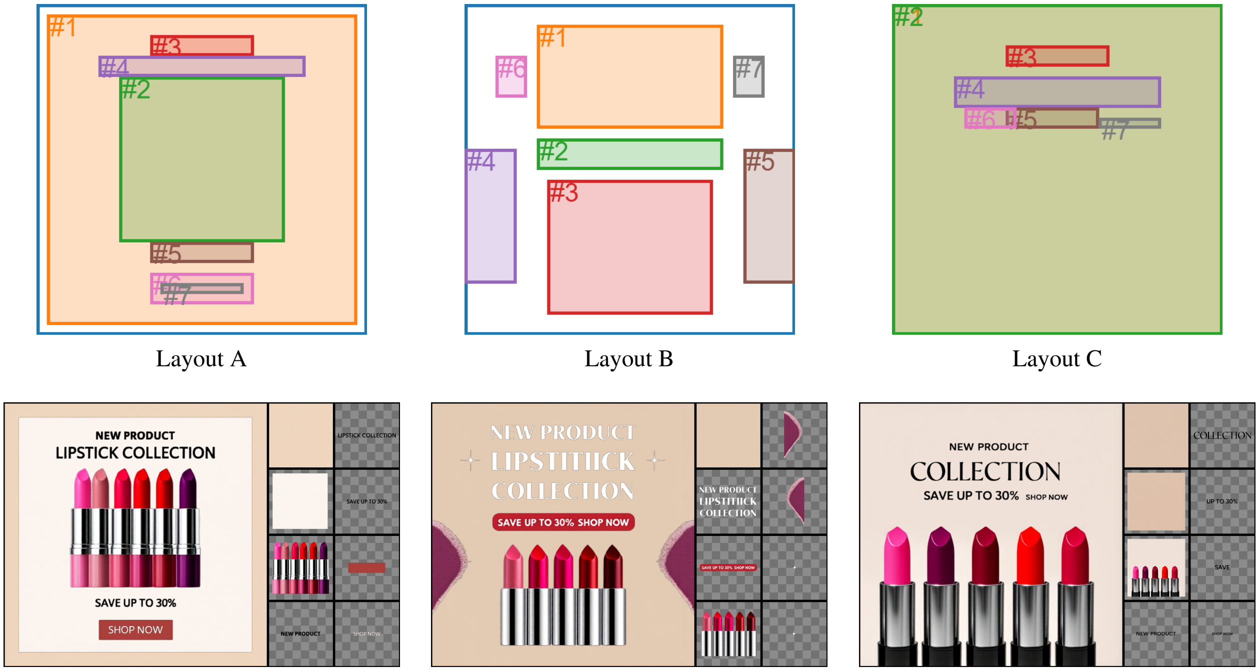
Generated B



Generated C

Table 12. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index ‘#0’ is omitted) at the second row and the generated results at the last row. (Case 4)

Prompt: The image features a collection of lipsticks. There are five lipsticks in total, each with a different color. From left to right, the first lipstick is a light pink, the second is a darker pink, the third is a bright red, the fourth is a deep red, and the fifth is a deep purple. Each lipstick is encased in a silver tube with a clear cap, allowing the color to be visible. The lipsticks are arranged in a straight line, and the background is a neutral beige color. At the top of the image, there is text that reads "NEW PRODUCT LIPSTICK COLLECTION," and at the bottom, there is a promotional message that says "SAVE UP TO 30% SHOP NOW." The overall style of the image is promotional and designed to attract customers to the new lipstick collection.



Layout A

Layout B

Layout C

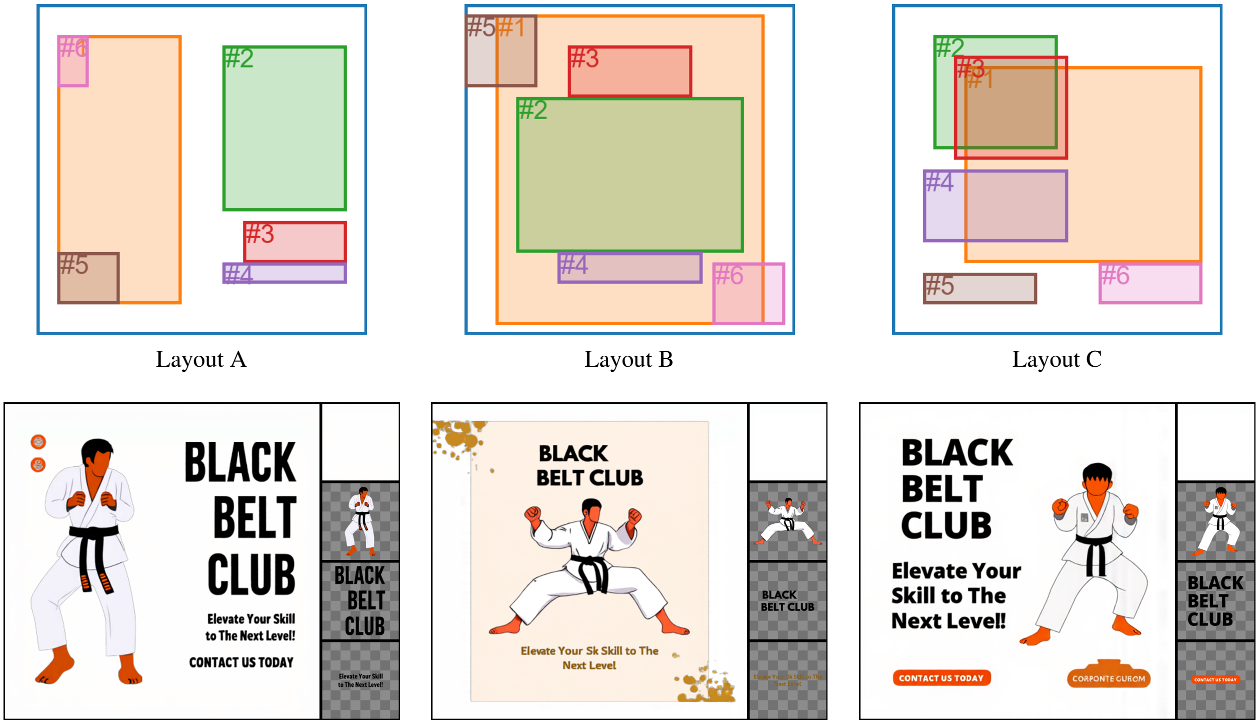
Generated A

Generated B

Generated C

Table 13. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index '#0' is omitted) at the second row and the generated results at the last row. (Case 5)

Prompt: The image features a stylized illustration of a person in a martial arts pose. The individual is depicted in a dynamic stance with one leg extended straight out to the side, while the other leg is bent at the knee, supporting the body. The person is wearing a white martial arts uniform, commonly known as a gi, and a black belt, which signifies a high level of proficiency in the martial art. The belt is tied around the waist, and the person's hands are clenched into fists, suggesting a state of readiness or combat. Above the illustration, there is text that reads "BLACK BELT CLUB" in bold, capital letters, indicating the name of the organization or program being advertised. Below this, there is a slogan that says "Elevate Your Skill to The Next Level!" which is a motivational statement encouraging individuals to improve their martial arts abilities. At the bottom of the image, there is a call to action that says "CONTACT US TODAY," suggesting that interested individuals should reach out to the club for more information or to join. The overall style of the image is clean and modern, with a limited color palette that focuses on the martial arts theme. The illustration is likely intended for promotional purposes, aiming to attract potential members to the Black Belt Club.



Layout A

Layout B

Layout C

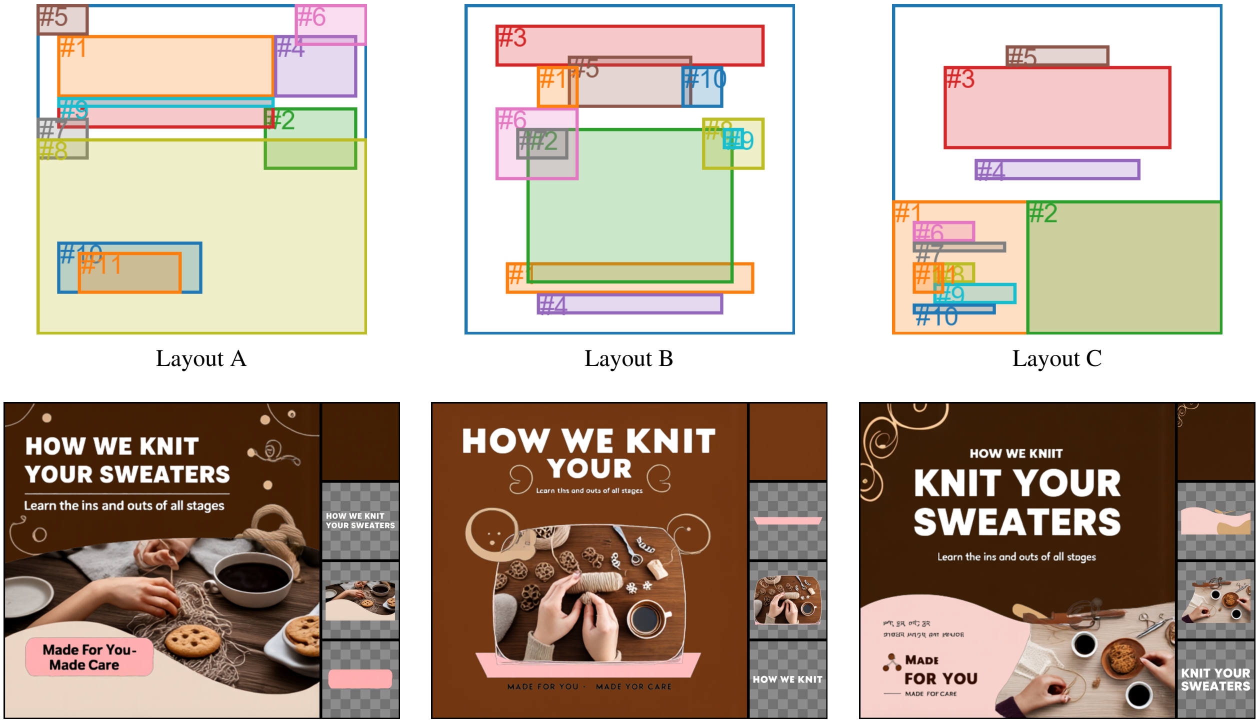
Generated A

Generated B

Generated C

Table 14. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index '#0' is omitted) at the second row and the generated results at the last row. (Case 6)

Prompt: The image is a promotional graphic for a knitting service. It features a warm, inviting design with a wooden table as the central focus. On the table, there are various knitting tools and materials, including a pair of hands actively knitting with yarn, a pair of scissors, a cup of coffee, and a bowl of cookies. The background is a rich, dark brown, and there are decorative elements such as swirls and dots in lighter shades of brown and beige. At the top of the image, in large, bold white letters, the text reads ”HOW WE KNIT YOUR SWEATERS.” Below this, in smaller white font, it says ”Learn the ins and outs of all stages.” At the bottom of the image, there’s a pink banner with white text that states ”MADE FOR YOU - MADE WITH CARE.” The overall style of the image is cozy and crafty, designed to appeal to those interested in handmade knitwear.

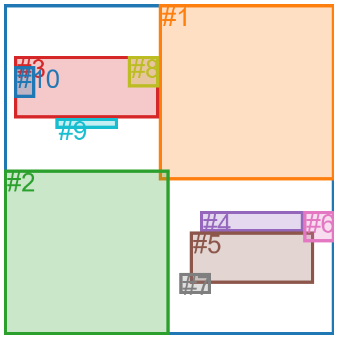

HOW WE KNIT YOUR SWEATERS
Learn the ins and outs of all stages
Made For You- Made Care
HOW WE KNIT YOUR SWEATERS
HOW WE KNIT YOUR
Learn ths and outs of all stages
MADE FOR YOU - MADE YOR CARE
HOW WE KNIT
HOW WE KNIIT
KNIT YOUR SWEATERS
Learn the ins and outs of all stages
MADE FOR YOU
MADE FOR CARE
KNIT YOUR SWEATERS

Layout A Layout B Layout C

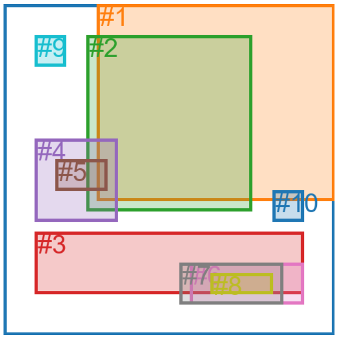
Generated A Generated B Generated C

Table 15. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index ‘#0’ is omitted) at the second row and the generated results at the last row. (Case 7)

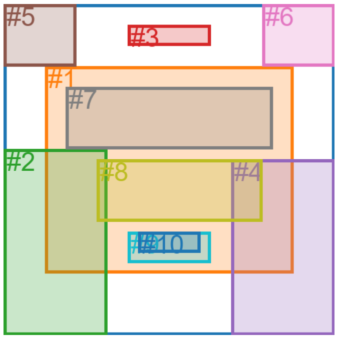
Prompt: The image is a collage of three separate photographs, each depicting a different scene related to hiking and nature. In the top left photograph, there is a text overlay that reads ”EXPLORE VIRGINIA’S HIKING TRAILS” in a bold, sans-serif font. The text is green with a slight shadow effect, making it stand out against the white background. The top right photograph features a man wearing a wide-brimmed hat and a light-colored shirt. He is smiling and looking directly at the camera. A green parrot is perched on his shoulder, adding a vibrant splash of color to the scene. The man appears to be outdoors, surrounded by lush greenery, suggesting a natural, possibly tropical, environment. The bottom left photograph shows two individuals, a man and a woman, who are engaged in a hiking activity. The man is wearing a hat and is holding a large, rolled-up map or document, which he seems to be examining. The woman is standing next to him, also wearing a hat, and is looking in the same direction as the man. They are both dressed in casual, outdoor-appropriate clothing. The background is filled with dense foliage, indicating that they are in a forested area. The bottom right photograph contains text that reads ”EXO TRAVEL BOOKING ONLINE” in a similar style to the text in the top left photograph. The text is green with a slight shadow effect, and it is positioned against a white background. Overall, the collage seems to be promoting outdoor activities, specifically hiking in Virginia, and is likely associated with a travel company or service. The images are designed to evoke a sense of adventure and connection with nature.



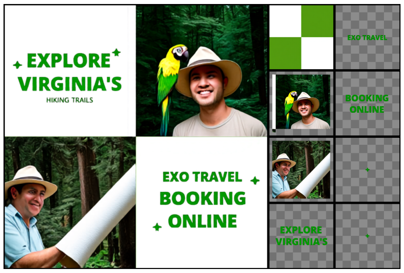
Layout A



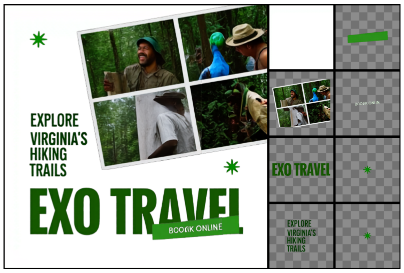
Layout B



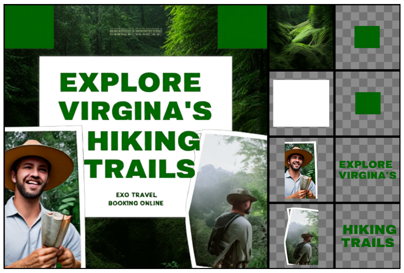
Layout C



Generated A


EXPLORE VIRGINIA'S HIKING TRAILS
EXO TRAVEL
BOOKIK ONLINE
EXO TRAVEL
EXPLORE VIRGINIA'S HIKING TRAILS

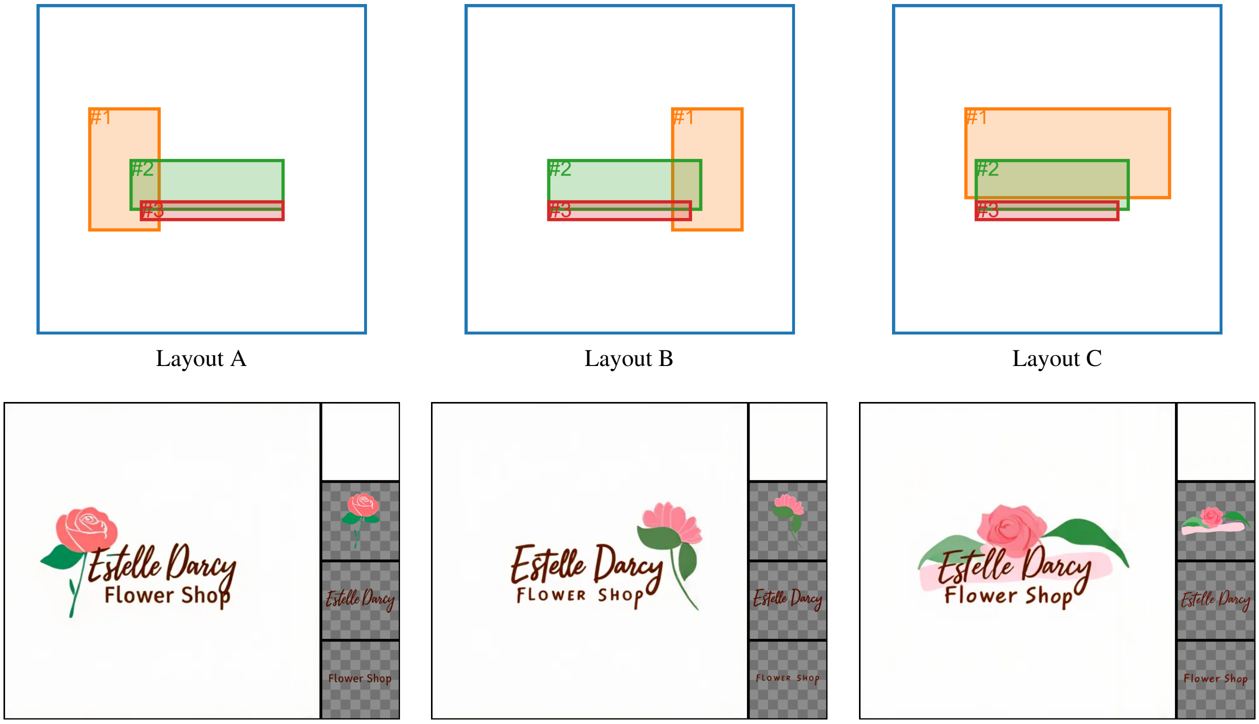
Generated B



Generated C

Table 16. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index ‘#0’ is omitted) at the second row and the generated results at the last row. (Case 8)

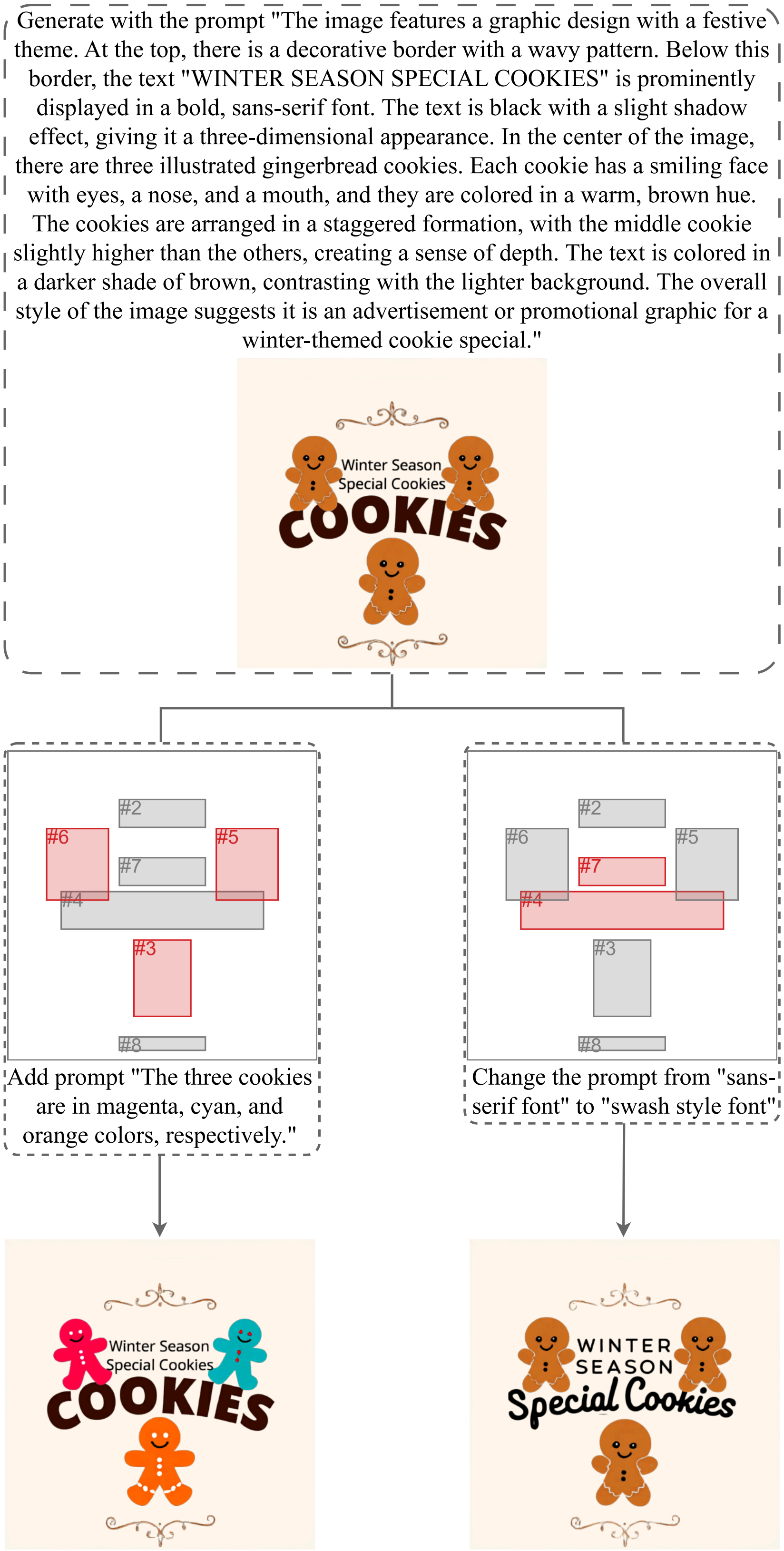
Prompt: The image features a logo for a flower shop named ”Estelle Darcy Flower Shop.” The logo is designed with a stylized flower, which appears to be a rose, in shades of pink and green. The flower is positioned to the left of the text, which is written in a cursive font. The text is in a brown color, and the overall style of the image is simple and elegant, with a clean, light background that does not distract from the logo itself. The logo conveys a sense of freshness and natural beauty, which is fitting for a flower shop.



Layout A Layout B Layout C

Generated A Generated B Generated C

Table 17. Generated results conditioned on the same prompt and variant layouts. We show the prompt at the first row, three different layouts (the background index ‘#0’ is omitted) at the second row and the generated results at the last row. (Case 9)



Generate with the prompt "The image features a graphic design with a festive theme. At the top, there is a decorative border with a wavy pattern. Below this border, the text "WINTER SEASON SPECIAL COOKIES" is prominently displayed in a bold, sans-serif font. The text is black with a slight shadow effect, giving it a three-dimensional appearance. In the center of the image, there are three illustrated gingerbread cookies. Each cookie has a smiling face with eyes, a nose, and a mouth, and they are colored in a warm, brown hue. The cookies are arranged in a staggered formation, with the middle cookie slightly higher than the others, creating a sense of depth. The text is colored in a darker shade of brown, contrasting with the lighter background. The overall style of the image suggests it is an advertisement or promotional graphic for a winter-themed cookie special."

Add prompt "The three cookies are in magenta, cyan, and orange colors, respectively."

Change the prompt from "sans-serif font" to "swash style font"

Figure 7. Layer-wise editing of the generated image.