

---

# LayerFusion: Harmonized Multi-Layer Text-to-Image Generation with Generative Priors

---

Anonymous Author(s)

Affiliation

Address

email

## A Disclaimer

In the provided qualitative results throughout this paper, we apply blurring to any trademark logos visible in the generated samples for copyright issues.

## B Related Work

**Denoising Probabilistic Diffusion Models** Diffusion models contributed significantly in the field of image generation, specifically for the task of text-to-image generation. In early efforts, [11, 21, 22] made significant contributions to the area, where significant improvements in generation performance have been experienced with diffusion models on the pixel level. In another paradigm [20] proposed operating in a latent space, which enabled the generation of high-quality images with a lower computational cost compared to models operating at the pixel level, which built the foundation of the state-of-the-art image generation models [18, 8, 16]. Although such approaches differ in terms of their architecture designs, they all follow a paradigm that prioritizes building blocks based on attention blocks [24].

**Transparent Image Layer Processing** In terms of obtaining a single foreground layer, the work of [5] presents PP-Matting, a trimap-free natural image matting method that achieves high accuracy without requiring auxiliary inputs such as user-supplied trimaps. Meanwhile, [19] propose Alfie, a method for generating high-quality RGBA images using a pretrained diffusion transformer model, designed to provide fully automatic, prompt-driven illustrations for seamless integration into design projects or artistic scenes. It modifies the inference-time behavior of a diffusion model to ensure that the generated subjects are centered and fully contained without sharp cropping. It utilizes cross-attention and self-attention maps to estimate the alpha channel, enhancing the flexibility to integrate generated illustrations into complex scenes. In terms of multi-layer, [23] recently introduced MuLAn, a novel dataset comprising more than 44,000 multilayer RGBA decompositions of RGB images, designed to provide a resource for controllable text-to-image generation. MuLAn is constructed using a training-free pipeline that decomposes a monocular RGB image into a stack of layers, including background and isolated instances.

Although these methods have made significant progress, precise control over image layers and their harmonization remain challenging. The most related effort for layered content synthesis is done by [25]. This approach is notable for its ability to generate both single and multiple transparent image layers with minimal alteration to the original latent space of a pre-trained diffusion model. The method utilizes a 'latent transparency' that encodes the alpha channel transparency into the latent manifold of the model. It offers two main workflows. One is jointly generating foreground and background layers by attention sharing. The other one is a sequential approach that generates one layer first and then another layer based on the previous layer. Both requires heavy model training relying on synthetic training data in less satisfying quality (obtained by a pretrained inpainting model). In contrast, our framework provides a training-free solution that offers generation of layered

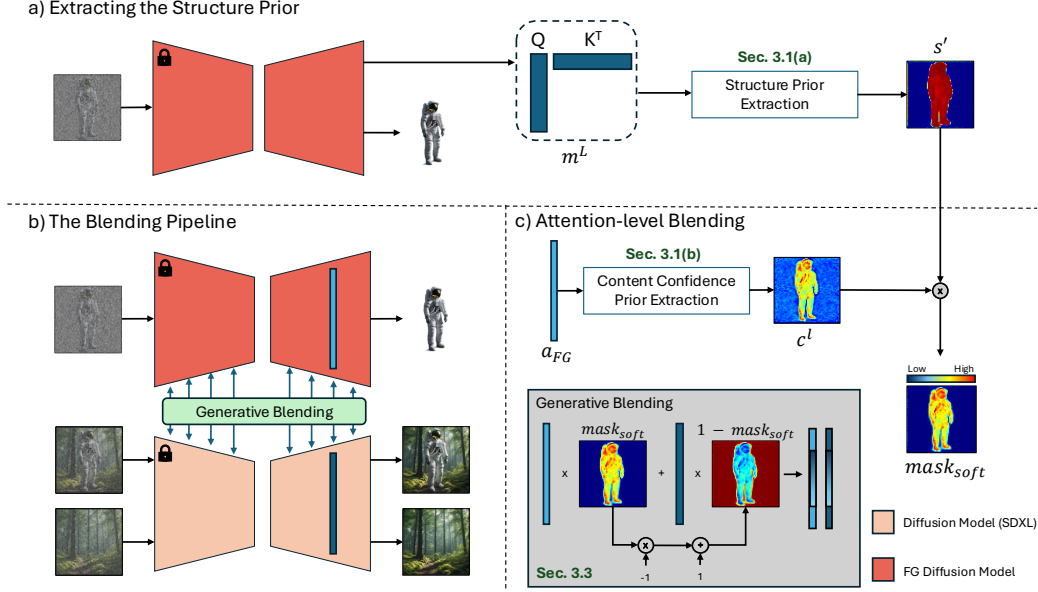


Figure 1: **LayerFusion Framework.** By making use of the generative priors extracted from transparent generation model  $\epsilon_{\theta,FG}$ , LayerFusion is able to generate image triplets consisting a foreground (RGBA), a background, and a blended image. Our framework involves three fundamental components that are connected with each other. First we introduce a prior pass on  $\epsilon_{\theta,FG}$  (a) for extracting the structure prior, and then introduce an attention-level interaction between two denoising networks ( $\epsilon_{\theta,FG}$  and  $\epsilon_{\theta}$ ) (b), with an attention level blending scheme with layer-wise content confidence prior, combined with the structure prior (c).

content in a simultaneous manner, which benefits both from layer transparency and achieves harmony between layers.

**Compositional Image Generation with Diffusion Models** Text-to-image models are known to be limited in generating images with multiple objects specified in the prompt. To tackle it, [14] introduces a compositional generation scheme using conjunction and negation operations. Additionally, [4] resolves the issue of neglected objects with an objective targeting the cross attention maps directly. Extending on this paradigm, [1, 15, 3] offer additional approaches to solve the text-image alignment problem for multiple image elements. However, such approaches all operate on text level, which offers limited control on the spatial positions of the objects in the image.

To enable spatial control over the generated images, [7] utilizes the cross attention maps to determine the object boundaries and introduces spatial edits on latent level. Despite enabling spatial editing for the desired objects, they suffer from the identity preservation problem, where they do not define an explicit layer representation for objects but operate in the RGB space. By addressing this limitation, our approach enables such editing by generating foreground and background pairs that are in harmony with each other, where the foreground object can be moved freely with its transparency properties.

## C Ablation Study

**Influence on BG on FG** We explore how changes in the background prompt affect the generated foreground content. As shown in Fig. 3 (a), by varying the background conditions, such as changing weather scenarios, leads to corresponding adjustments in the foreground details, like the clothing or accessories of a person, as well as fine-grained details such as adding snow on the boots (see rightmost image in Fig. 3 (a)). All experiments are conducted using the same seed, allowing for the preservation of the subject’s identity while adapting other features to match the changing background context. This demonstrates the dynamic adaptability of our method, where the foreground is influenced by the background for more contextually appropriate outputs. We provide additional examples on the impact of the background in the supplementary material.



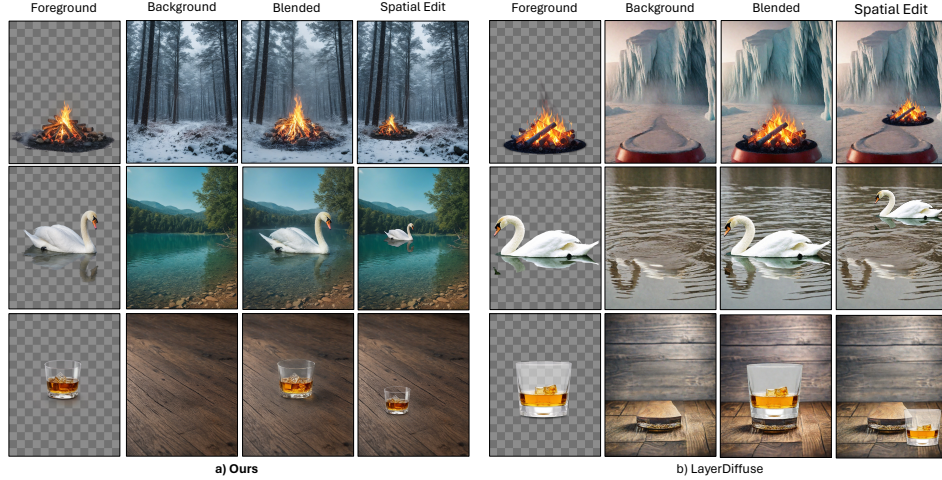


Figure 2: **Qualitative Comparisons on Layered Generation.** We compare our proposed framework with [25] to evaluate the performance in terms of layered image generation (e.g. foreground, background, blended). It clearly shows that [25] propagates the background completion issues observed in SDXL-Inpainting, which degrades the spatial editing quality with the outputted layers. In contrast, our method can provide both harmonized blending results and isolated foreground and background, which enables spatial editing in a straight-forward manner. In our comparisons, we use identical foreground and background prompts for both methods.

62 **Alpha Blending vs. Generative Blending** We compare two blending strategies: Alpha Blending,  
63 which guarantees a complete match between the generated foreground and the blended result,  
64 and Generative Blending, which aims for a more realistic composition by considering shadows,  
65 lighting, and contextual harmonization. As can be seen from Fig. 3 (b), the Alpha Blending  
66 is more deterministic, ensuring that the foreground remains consistent with the original output  
67 without considering the interactions between foreground and background. Meanwhile, the Generative  
68 Blending produces more visually appealing results by better handling subtle elements like shadows  
69 and lighting, making the generated content appear more natural and harmonized with the background.  
70 Note how the feet of the cow is harmonized with the grassy surface in Generative Blending as opposed  
71 to Alpha Blending.

72 **Self-Attention vs. Cross-Attention** The use of attention masks plays a crucial role in controlling  
73 the interaction between the foreground and background layers. As can be seen from Fig. 3 (c),  
74 when the self-attention map is used alone, there are risks of unwanted leaks from the premultiplied  
75 image (i.e., the output from the foreground generation model with a gray background), resulting in  
76 imprecise boundaries. The cross-attention map, on the other hand, provides more precise information,  
77 sharpening the bounding map. However, if the cross-attention map is used in isolation, the regions that  
78 are not voted by the structure prior(from the self attention map) may create artifacts. By combining  
79 both attention maps, we are able to balance these effects, where the cross-attention sharpens the  
80 boundary, and the self-attention ensures coherence within the bounded region.

81 **Soft Decision Boundary Coefficient** We investigate the effect of varying the soft decision boundary  
82 coefficient, which is used to derive the hard mask in our blending formulation. Lower coefficients  
83 result in softer decision boundaries, causing leaks into the foreground and leading to imprecise alpha  
84 channel predictions, as seen in the first image of Fig. 3 (d). As the coefficient value increases,  
85 the boundary becomes more defined, allowing for more accurate capture of foreground details and  
86 improving consistency between the foreground and blended image. This is particularly evident in the  
87 pocket area of the man’s clothing in the second and third images, where higher coefficients result in  
88 more precise blending and alignment.

Table 1: **Latent-blending benchmark.** Higher aesthetic scores and lower KID are better.

Method	Aest. (SigLIP) $\uparrow$	Aest. (CLIP) $\uparrow$	KID $\downarrow$
LayerDiffuse	5.64	5.84	0.0047
BLD	6.13	6.02	0.0156
<b>Ours</b>	<b>6.33</b>	<b>6.24</b>	<b>0.0032</b>

## D Qualitative Results

**Comparisons with Layered Generation Methods** We compare our proposed method against LayerDiffuse to evaluate the quality of the generated foreground (FG), background (BG), and the blended image (see Fig. 2). As shown in the results, our model achieves harmonious blending with smooth FG and BG images. In contrast, LayerDiffuse (Generation) struggles to produce a smooth and consistent background (see the artifacts in Fig. 2 (b) in the background images). This limitation arises from the sequential approach used to curate the training dataset of LayerDiffuse [25], where given a foreground and a blended image, the background is generated by outpainting the foreground from the blended image with SDXL-Inpainting [18]. As a result of this strategy on dataset generation, the background generation model experiences artifacts in the outpainted region, which propagates from the inpainting model. As it is also highlighted in Fig. 2, such artifacts effect the ability of performing spatial edits with the generated foreground and background layers.

**Foreground Extraction Methods** As another baseline, we compare our proposed framework with foreground extraction methods given the blended image (background and blended for LayerDiffuse [25]) to outline the advantages of simultaneous generation of the foreground and background images (layers) in Fig. 5. In addition to background and blended image conditioned foreground extraction pipeline of [25], we also consider PPMating [5] and MattingAnything [13] as competitors as they apply matting to extract the foreground layer from the blended image. As we demonstrate qualitatively in Fig 5, simultaneous generation results in more precise foreground for the cases that include interaction between foreground and background layers (e.g. legs of the horse occluded in the grass) compared to state-of-the art foreground extraction/matting methods.

**Harmonization Quality** For the evaluation of the blending capabilities of our framework, we compare our generative blending result with state-of-the-art image harmonization methods. In our comparisons, we investigate the realism of the harmonized output considering the object (foreground) getting harmonized in the process. To get the harmonized outputs from the competing methods, we give the alpha blending result obtained from our pipeline to each of the competitor methods, and qualitatively evaluate the obtained outputs in Fig. 4. Specifically, we compare our framework with [12, 6, 10].

## E Supplementary Quantitative Results

**Comparison with Latent-Blending Baselines** To gauge the gains of our attention-level fusion over approaches that merge features directly in latent space, we benchmark against *LayerDiffuse (Generation)* and *Blended Latent Diffusion (BLD)*[2]. Using the same 150 prompts and seeds for all methods, we generate blended images and score them with two public aesthetic predictors—one that uses a CLIP backbone and another that uses a SigLIP backbone. We also compute Kernel Inception Distance (KID) to a reference set of SDXL generations conditioned on the same prompts. For fairness, BLD [2] receives the exact  $\alpha$ -mask produced by our pipeline instead of a user-drawn mask. The results in Table 1 show that our method attains the highest aesthetic scores and the lowest KID, outperforming LayerDiffuse on every metric and reducing BLD’s KID by more than 5x. Qualitative comparisons in Fig.6 highlight these differences: our attention-guided fusion delivers crisp object boundaries and consistent colors, whereas BLD introduces visible color fringing and soft halos along the foreground edge. The cleaner edges and seamless backgrounds in our outputs corroborate the quantitative gains and show that attention-level fusion produces composites closer to native SDXL generations.

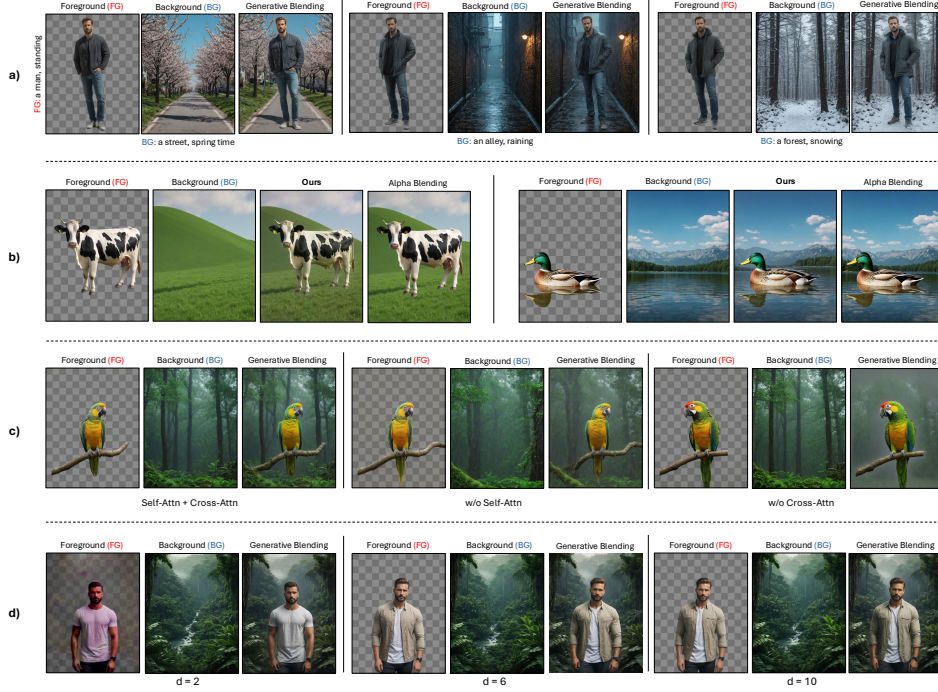


Figure 3: We perform extensive ablation studies on the effect of (a) **Background Influence on Foreground**: Background changes (e.g., weather) dynamically adjust the foreground (e.g., outfit) while preserving identity. (b) **Alpha vs. Generative Blending**: Alpha Blending ensures a perfect match, while Generative Blending creates more realistic harmonization by handling shadows and lighting. (c) **Self-Attention vs. Combined Attention Masks**: Self-attention alone causes leaks; cross-attention alone affects the entire image. Combining both achieves sharper boundaries and better coherence. (d) **Soft Decision Boundary Coefficient**: Lower coefficients cause leaks; higher coefficients yield more precise alpha and consistent blending (e.g., the pocket of the man’s clothing).

## 132 F Limitations

133 While our proposed image generation pipeline based on Latent Diffusion Models (LDMs) demon-  
 134 strates significant advancements in generating harmonized foreground (RGBA) and background  
 135 (RGB) layers, there are several limitations that warrant discussion. Our current approach focuses on  
 136 generating images with two distinct layers—a foreground and a background. While this is suitable for  
 137 many creative workflows, it does not extend to more complex scenarios involving multiple layers or  
 138 hierarchical relationships among multiple visual elements, which we intent to explore for future work.  
 139 Moreover, the harmonization between foreground and background layers in our framework relies  
 140 heavily on the quality of the cross-attention and self-attention masks extracted from the generation  
 141 model. In cases where these masks are suboptimal or noisy, the blending of layers may not be as  
 142 effective, leading to artifacts or less coherent outputs. Finally, our method depends on pre-trained  
 143 Latent Diffusion Models both for foreground and background generation, which may carry inherent  
 144 biases from their training data (such as generating centered foregrounds for the RGBA component).  
 145 These biases can affect the generated content, potentially leading to outputs that are not entirely  
 146 aligned with user expectations or specific requirements in diverse applications. Nevertheless, our  
 147 method provides a structured framework for generating transparent images and layered compositions,  
 148 which are crucial for many creative tasks.

## 149 G Analyses on Structure Priors from Different Layers

150 In all of the experiments we provide, we utilize the structure prior extracted from the last attention  
 151 map of the foreground diffusion model,  $\epsilon_{\theta, FG}$ . As a justification of this decision and to clearly  
 152 illustrate what different self attention layers focus on throughout the generation process, we provide



Figure 4: **Comparisons on Image Harmonization.** We qualitatively evaluate our methods blending capabilities by comparing with image harmonization methods Harmonizer [12], INR-Harmonization [6], and PCT-Net [10]. Our proposed generative blending approach results in adaptation of the foreground object to the background scene (e.g. snow effect on the campfire), in addition to harmonization methods.

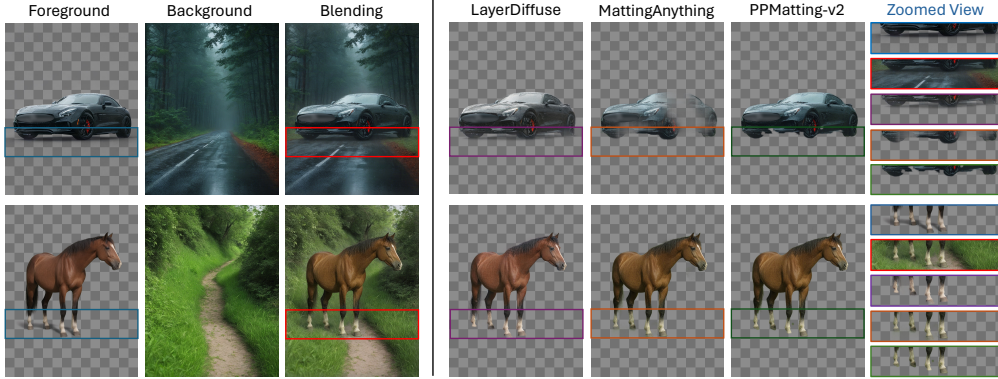


Figure 5: **Comparisons with Foreground Extraction Methods** To illustrate the advantage of our method over the task of foreground extraction given a blended image, we qualitatively compare our approach with LayerDiffuse [25], Matting Anything [13], and PPMatting [5]. As also highlighted by [25], simultaneous generation of the foreground layer is more advantageous compared to extracting from the blended image in terms of the quality of the foreground image.

153 structure priors extracted from different layers in Fig. 7. As it can also be observed visually, the  
 154 structure prior extracted from the last self attention layer provides a more precise estimate of the  
 155 shape of the foreground being generated.

## 156 H Detailed Blending Algorithm

157 Supplementary to the definition of the blending algorithm provided in the methodology section,  
 158 we provide a more detailed description in this here, for clarity. Our proposed blending approach  
 159 involves three sub-procedures, which are the extraction of the structure prior, extraction of the content  
 160 confidence prior and the attention blending step. In this section, we provide the pseudo-code for all  
 161 three procedures as Alg. 1, 2 and 3.

## 162 I User Study Details

163 We conduct our user study over 50 participants with 40 image triplets generated by LayerFusion and  
 164 [25]. For the generation of the subjected triplets, we generate examples with animal, vehicle, matte  
 165 objects, person and objects with transparency properties as the foreground to get samples representing  
 166 a diverse distribution of subjects. Following sample generation, we ask users to rate each image



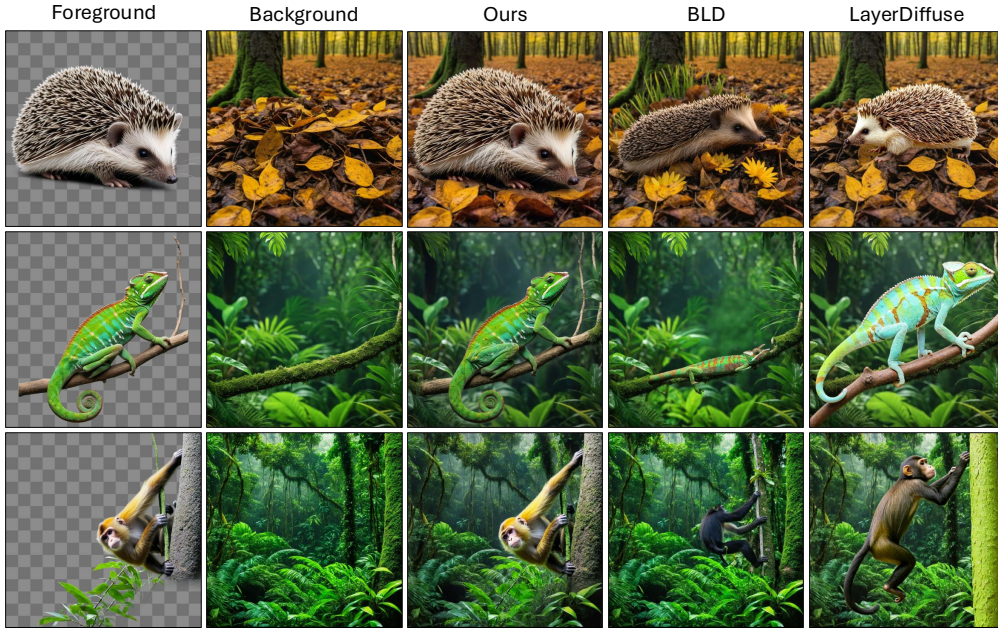


Figure 6: **Comparisons with Latent Blending Baselines.** We provide qualitative comparisons with latent blending baselines. In correspondence to a triplet generated by our method, we perform background-conditioned generations using the same generation prompts for BLD [2] and LayerDiffuse. Supplementary to the background image, we also provide the binarized alpha channel of the foreground layer to BLD [2]. As can be observed qualitatively, our method provides blending superior capabilities compared to latent blending baselines.

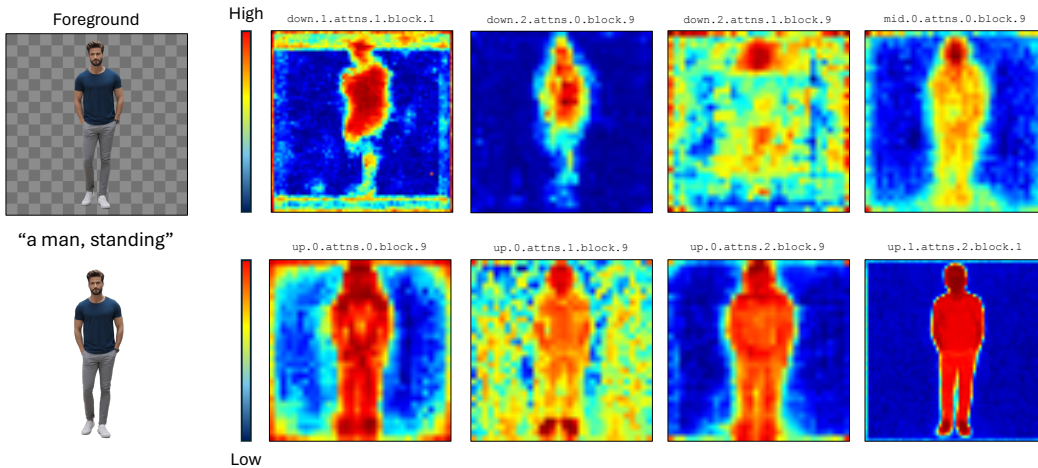


Figure 7: **Visualization of the structure priors from different self attention layers.** We visualize the structure priors extracted from different self attention layers of the foreground diffusion model, where the diffusion timestep is set as  $t = 0.8T$ . We visualize the structure priors from the self attention layer of each model block, follow the block definition of [9]. We follow the naming convention of diffusers ([17]). In all of our experiments, we use the structure prior from self attention layer `up.1.attns.2.block.1`.

---

**Algorithm 1:** Extracting the Structure Prior

---

**Input:** Foreground denoiser  $\epsilon_{\theta, \text{FG}}$ , latent feature map  $\mathbf{z}_t$ , prompt  $\mathbf{p}_{\text{FG}}$

**Output:** Structure prior  $s'$

**Procedure:** EXTRACTSTRUCTUREPRIOR( $\epsilon_{\theta, \text{FG}}$ ,  $\mathbf{z}_t$ ,  $\mathbf{p}_{\text{FG}}$ )

**begin**

    // Retrieve Noise Prediction (unused) and Last Self Attention Map

$\hat{\epsilon}, m^L \leftarrow \epsilon_{\theta, \text{FG}}(\mathbf{z}_t, \mathbf{p}_{\text{FG}})$

    // Row-wise sparsity (Eq. 6 in main text)

**for**  $i = 1$  **to**  $M$  **do**

$s_i \leftarrow \left( \sum_{j=1}^M (m_{i,j}^L)^2 \right)^{-1}$

**end**

    // Convert sparsity to density and normalize

$s' \leftarrow 1 - \text{norm}(\{s_i\}_{i=1}^M)$

**return**  $s'$

**end**

---

---

**Algorithm 2:** Extracting the Content-Confidence Prior

---

**Input:** Foreground denoiser  $\epsilon_{\theta, \text{FG}}$ , hidden states  $h$ , prompt  $\mathbf{p}_{\text{FG}}$

**Output:** Content prior  $c$

**Procedure:** EXTRACTCONTENTPRIOR( $\epsilon_{\theta, \text{FG}}$ ,  $h$ ,  $\mathbf{p}_{\text{FG}}$ )

**begin**

    // Cross-attention forward pass

$\text{attn\_out}, \text{attn\_probs} \leftarrow \text{ATTENTION}_{\theta, \text{FG}}(h, \mathbf{p}_{\text{FG}})$

$n = \text{attn\_probs}$

    // Average EOS channel over  $H$  heads

$c \leftarrow \frac{1}{H} \sum_{k=1}^H n_{k, :, \langle \text{EOS} \rangle}$

**return**  $c$

**end**

---

167 triplet from a scale of 1-to-5, with the following question: “Please rate the following image triplet  
168 from a scale of 1-to-5 (1 - unsatisfactory, 5 - very satisfactory) considering how realistic each image  
169 is and how naturally blended they are”. The users are also supplied the foreground and background  
170 prompts used to generate the image triplet, for each method. We provide an example question from  
171 the conducted user study in Fig. 8.

## 172 J Supplementary Generation Results

173 In addition to the results provided in the main paper, we provide supplementary generation results in  
174 this section. Below, we include harmonized generations of a variety of subjects. We provide Fig. 9 to  
175 Fig. 19 as supplementary results.

## 176 References

- 177 [1] Agarwal, A., Karanam, S., Joseph, K., Saxena, A., Goswami, K., Srinivasan, B.V.: A-star:  
178 Test-time attention segregation and retention for text-to-image synthesis. In: Proceedings of the  
179 IEEE/CVF International Conference on Computer Vision. pp. 2283–2293 (2023)
- 180 [2] Avrahami, O., Fried, O., Lischinski, D.: Blended latent diffusion. ACM transactions on graphics  
181 (TOG) **42**(4), 1–11 (2023)
- 182 [3] Bao, Z., Li, Y., Singh, K.K., Wang, Y.X., Hebert, M.: Separate-and-enhance: Compositional  
183 finetuning for text-to-image diffusion models. In: ACM SIGGRAPH 2024 Conference Papers.  
184 pp. 1–10 (2024)

---

**Algorithm 3:** Attention-Level Blending (one transformer block)

---

**Input:** Foreground denoiser  $\epsilon_{\theta,FG}$ , RGB denoiser  $\epsilon_{\theta}$ , hidden states  $h_{FG}, h_{BL}, h_{BG}$ , prompts  $\mathbf{p}_{FG}, \mathbf{p}_{BG}$ , boundary coefficient  $d$ , structure prior  $s'$

**Output:** Updated attention outputs  $a'_{FG}, a'_{BL}, a_{BG}$

**Procedure:** ATTNBLENDD( $\epsilon_{\theta,FG}, \epsilon_{\theta}, h_{FG}, h_{BL}, h_{BG}, \mathbf{p}_{FG}, \mathbf{p}_{BG}, d, s'$ )

**begin**

```
// Layer normalization (shared parameters)
 $h_{FG}^{\ell}, h_{BL}^{\ell}, h_{BG}^{\ell} \leftarrow \text{LAYERNORMCROSSATTN}(h_{FG}, h_{BL}, h_{BG})$ 
// Layer-specific content prior
 $c \leftarrow \text{EXTRACTCONTENTPRIOR}(\epsilon_{\theta,FG}, h_{FG}^{\ell}, \mathbf{p}_{FG})$ 
// Soft and hard masks (Eq. 7)
 $\text{mask}_{\text{soft}} \leftarrow \text{norm}(s' \odot c)$ 
 $\text{mask}_{\text{hard}} \leftarrow \sigma(d(\text{mask}_{\text{soft}} - 0.5))$ 
// Cross-attention for each branch
 $a_{BG}, a_{BL} \leftarrow \text{ATTENTION}_{\theta}([h_{BG}^{\ell}, h_{BL}^{\ell}], \mathbf{p}_{BG})$ 
 $a_{FG} \leftarrow \text{ATTENTION}_{\theta,FG}(h_{FG}^{\ell}, \mathbf{p}_{FG})$ 
// Blending rules (Eqs. 8-9)
 $a'_{BL} = a_{FG} \odot \text{mask}_{\text{soft}} + a_{BL} \odot (1 - \text{mask}_{\text{soft}})$ 
 $a'_{FG} = a'_{BL} \odot \text{mask}_{\text{hard}} + a_{FG} \odot (1 - \text{mask}_{\text{hard}})$ 
return  $a'_{FG}, a'_{BL}, a_{BG}$ 
```

**end**

---

- 185 [4] Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based  
186 semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*  
187 **42**(4), 1–10 (2023)
- 188 [5] Chen, G., Liu, Y., Wang, J., Peng, J., Hao, Y., Chu, L., Tang, S., Wu, Z., Chen, Z., Yu, Z., et al.:  
189 Pp-matting: High-accuracy natural image matting. *arXiv preprint arXiv:2204.09433* (2022),  
190 <https://arxiv.org/pdf/2204.09433>
- 191 [6] Chen, J., Zhang, Y., Zou, Z., Chen, K., Shi, Z.: Dense pixel-to-pixel harmonization via continu-  
192 ous image representation. *IEEE Transactions on Circuits and Systems for Video Technology*  
193 pp. 1–1 (2023). <https://doi.org/10.1109/TCSVT.2023.3324591>
- 194 [7] Epstein, D., Jabri, A., Poole, B., Efros, A., Holynski, A.: Diffusion self-guidance for controllable  
195 image generation. *Advances in Neural Information Processing Systems* **36**, 16222–16239 (2023)
- 196 [8] Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer,  
197 A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In:  
198 Forty-first International Conference on Machine Learning (2023)
- 199 [9] Frenkel, Y., Vinker, Y., Shamir, A., Cohen-Or, D.: Implicit style-content separation using b-lora.  
200 *arXiv preprint arXiv:2403.14572* (2024)
- 201 [10] Guerreiro, J.J.A., Nakazawa, M., Stenger, B.: Pct-net: Full resolution image harmonization  
202 using pixel-wise color transformations. In: *Proceedings of the IEEE/CVF Conference on*  
203 *Computer Vision and Pattern Recognition (CVPR)*. pp. 5917–5926 (June 2023)
- 204 [11] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural*  
205 *information processing systems* **33**, 6840–6851 (2020)
- 206 [12] Ke, Z., Sun, C., Zhu, L., Xu, K., Lau, R.W.: Harmonizer: Learning to perform white-box image  
207 and video harmonization. In: *European Conference on Computer Vision (ECCV)* (2022)
- 208 [13] Li, J., Jain, J., Shi, H.: Matting anything. In: *Proceedings of the IEEE/CVF Conference on*  
209 *Computer Vision and Pattern Recognition*. pp. 1775–1785 (2024)
- 210 [14] Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with  
211 composable diffusion models. In: *Computer Vision–ECCV 2022: 17th European Conference,*  
212 *Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. pp. 423–439. Springer (2022)



Figure 8: **Example Question from the User Study.** To evaluate the effectiveness our method perceptually, we conduct a user study over 40 generated image triplets. We provide an example question from this study for clarity. The users are shown an image triplet in the order of foreground, background and blended image and then asked to rate it from a scale of 1-to-5 (1 - unsatisfactory, 5 - very satisfactory).

- 213 [15] Meral, T.H.S., Simsar, E., Tombari, F., Yanardag, P.: Conform: Contrast is all you need for  
214 high-fidelity text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on  
215 Computer Vision and Pattern Recognition. pp. 9005–9014 (2024)
- 216 [16] Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the  
217 IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023)
- 218 [17] von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair,  
219 D., Paul, S., Berman, W., Xu, Y., Liu, S., Wolf, T.: Diffusers: State-of-the-art diffusion models.  
220 <https://github.com/huggingface/diffusers> (2022)
- 221 [18] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach,  
222 R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: The  
223 Twelfth International Conference on Learning Representations (2023)
- 224 [19] Quattrini, F., Pippi, V., Cascianelli, S., Cucchiara, R.: Alfie: Democratising rgba image  
225 generation with no \$\$\$ arXiv preprint arXiv:2408.14826 (2024), [https://arxiv.org/pdf/](https://arxiv.org/pdf/2408.14826)  
226 2408.14826
- 227 [20] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis  
228 with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision  
229 and pattern recognition. pp. 10684–10695 (2022)
- 230 [21] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Confer-  
231 ence on Learning Representations (2020)
- 232 [22] Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based  
233 generative modeling through stochastic differential equations. In: International Conference on  
234 Learning Representations (2020)
- 235 [23] Tudosi, P.D., Yang, Y., Zhang, S., Chen, F., McDonagh, S., Lampouras, G., Iacobacci,  
236 I., Parisot, S.: Mulan: A multi layer annotated dataset for controllable text-to-image  
237 generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern  
238 Recognition (CVPR) (2024), <https://openaccess.thecvf.com/content/CVPR2024/>



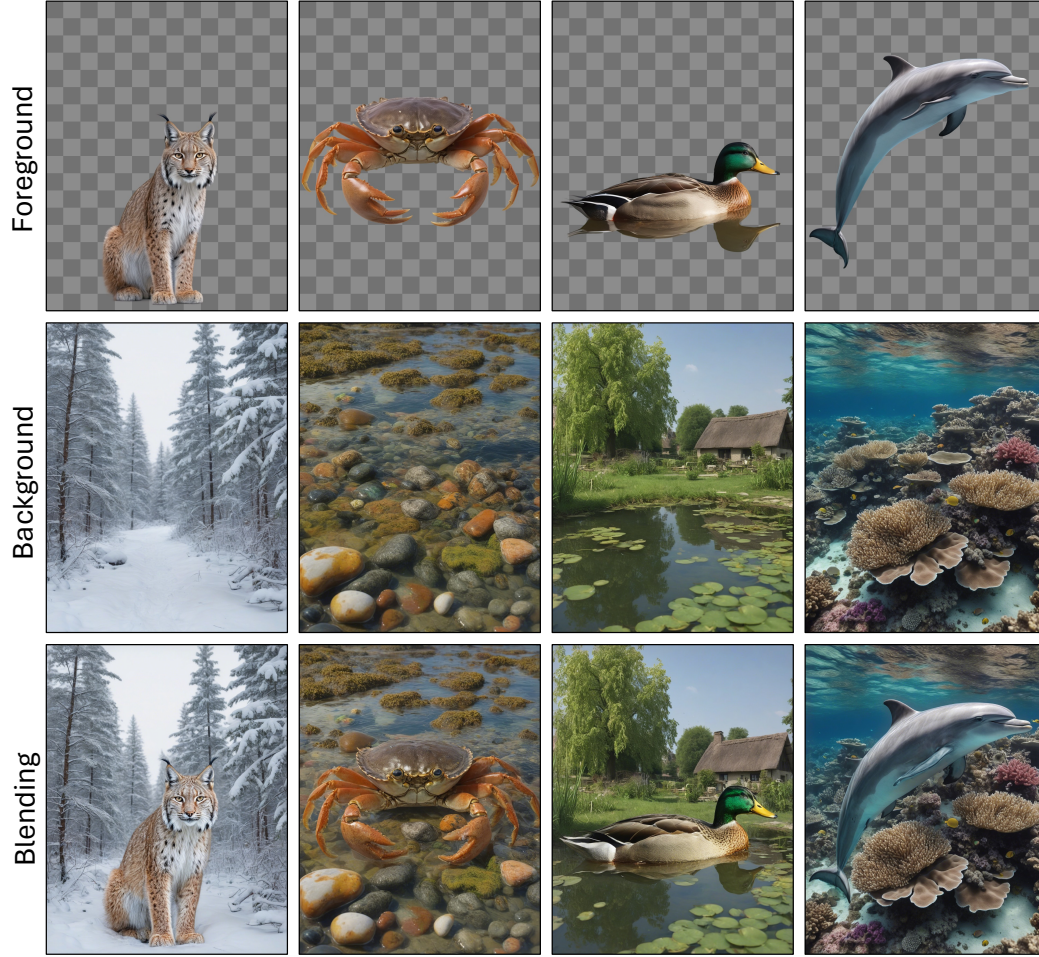


Figure 9: **Supplementary Generation Results with animal subjects.** Supplementary results with image resolution 896x1152. The foreground & background prompt pairs from left to right are: “a lynx”, “a snowy forest”), (“a crab”, “a rocky tide pool”), (“a duck”, “a village pond”), (“a dolphin”, “a crystal-clear coral reef”)

- papers/Tudosiu\_MULAN\_A\_Multi\_Layer\_Annotated\_Dataset\_for\_Controllable\_
- Text-to-Image\_Generation\_CVPR\_2024\_paper.pdf
- [24] Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
- [25] Zhang, L., et al.: Transparent image layer diffusion using latent transparency. arXiv preprint arXiv:2402.17113 (2024), <https://arxiv.org/abs/2402.17113>, last revised 23 Jun 2024

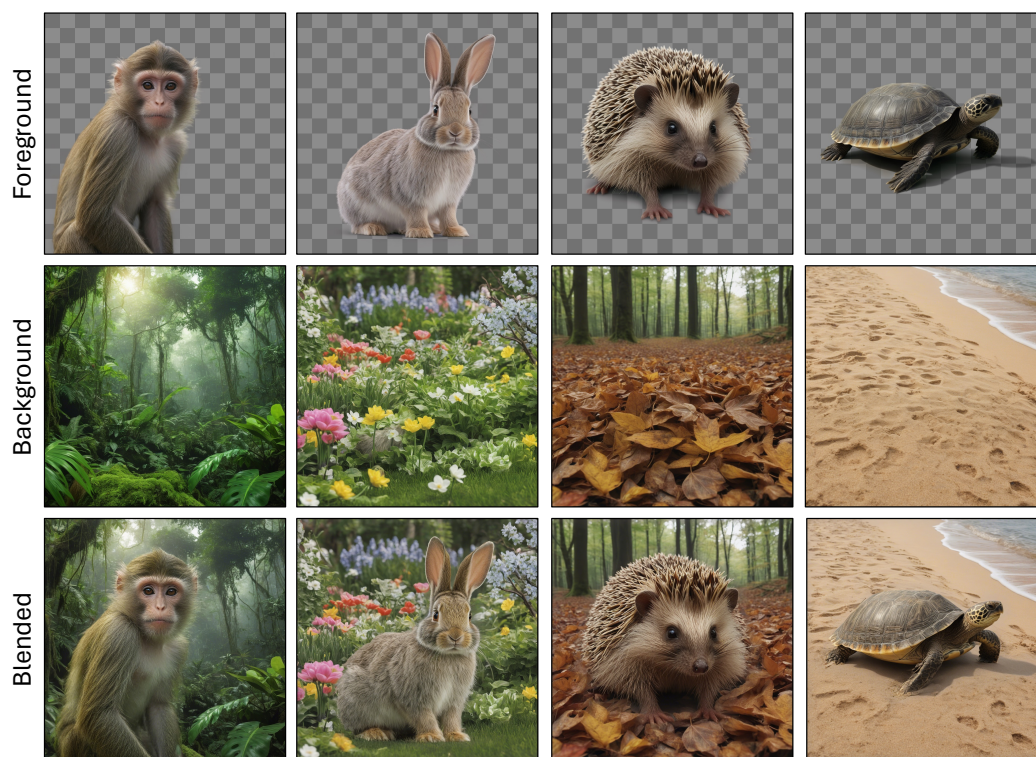


Figure 10: **Supplementary Generation Results with animal subjects.** Supplementary results with image resolution 1024x1024. The foreground & background prompt pairs from left to right are: (“a monkey”, “a vibrant tropical rainforest”), (“a rabbit”, “a backyard garden”), (“a hedgehog”, “a forest floor covered in leaves”), (“a turtle”, “a warm sandy beach”)



Figure 11: **Supplementary Generation Results with stylization prompts.** We provide additional examples with stylization prompts to demonstrate the harmonization capabilities of our method. For each image triplet, we generate the examples with the prompt set (“a man, standing”, “a street, style\_name”) where style\_name is “comics style” for the leftmost column. We provide the label (style\_name) for each style under its respective image triplet. All images have the resolution of 896x1152.





Figure 12: **Supplementary Generation Results for “comics” style.** To demonstrate the stylization capabilities of our layer harmonization approach, we provide additional results with the background prompt “a street, comics style”. The resolution is 896x1152 for all of the images.

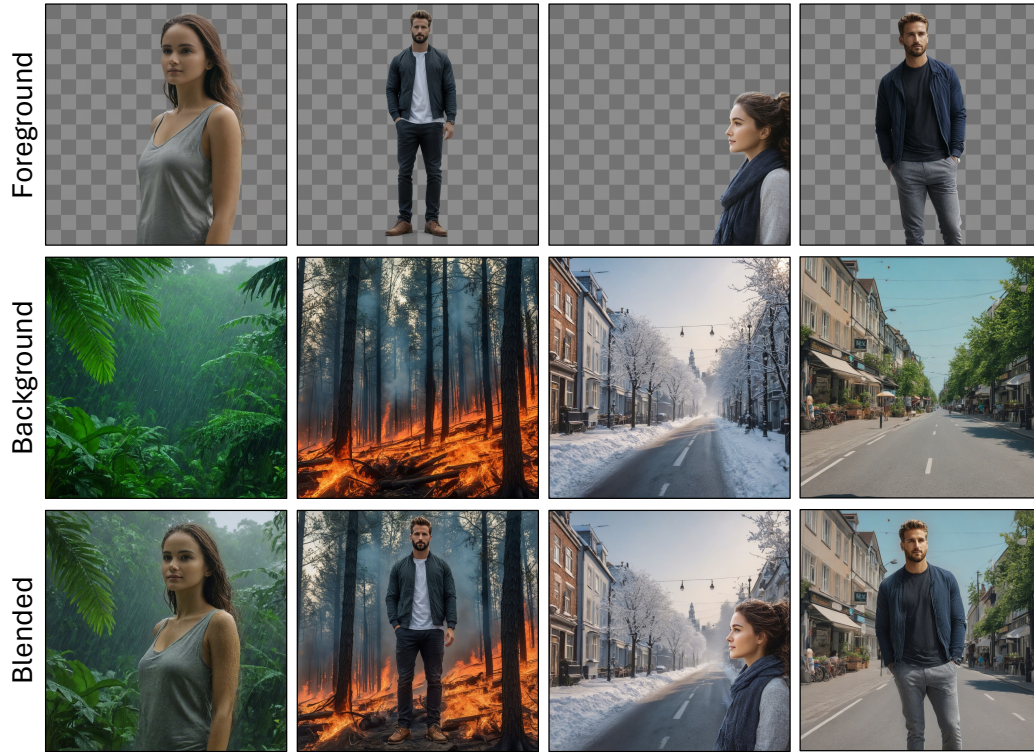


Figure 13: **Supplementary Generation Results with human subjects.** We provide additional examples with human subjects with different background prompts. The background prompts used are “a rainy jungle”, “a forest in fire”, “a street, winter time”, “a street, daytime”. Note that depending on the background prompt, the blending involves an interaction between the background and foreground (e.g. wetness in arm for the left-most image triplet). Image resolution is 1024x1024 for all examples.

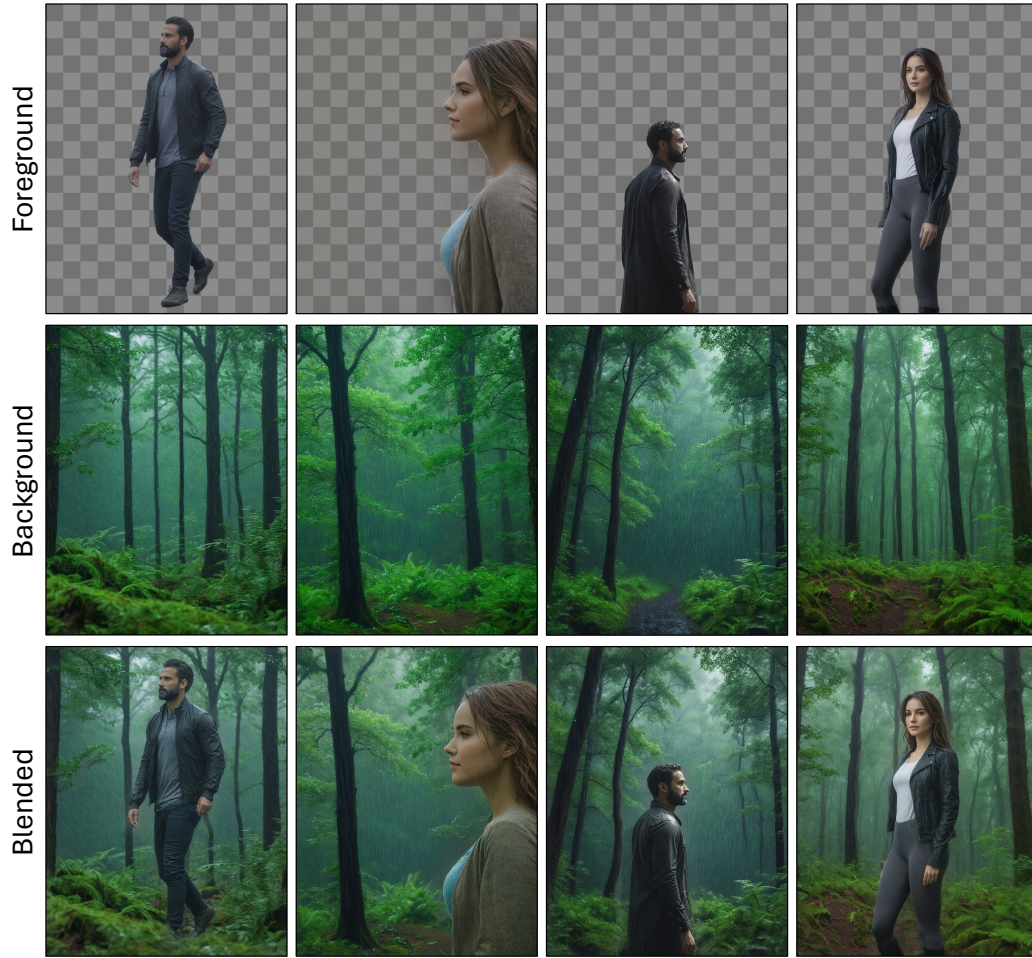


Figure 14: **Supplementary Generation Results for the background “a rainy forest”**. For each of the images, the background prompt "a rainy forest" is used to generate the background image. As it can also be observed from the provided examples, the background creates an influence over the foreground (e.g. wetness effect on the human subjects). The image resolution is 896x1152 for all examples.



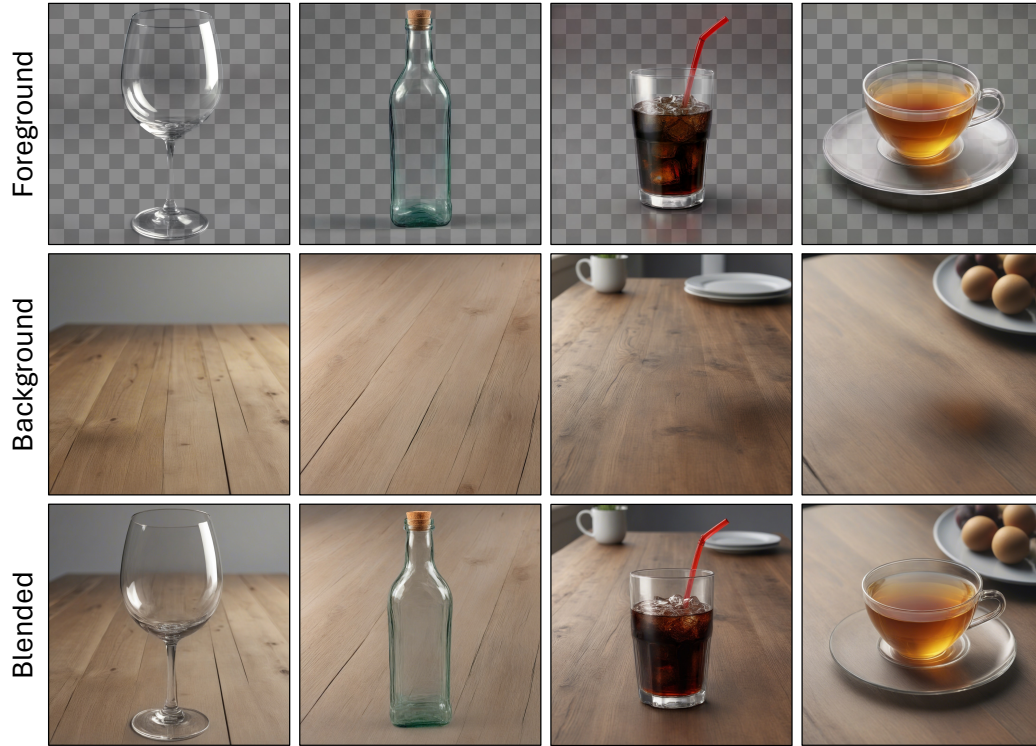


Figure 15: **Supplementary Generation Results for subjects with transparency property.** To demonstrate that our framework is able to preserve the transparency properties of layered image representations, we provide additional results here. With the background prompt "a table" we use the following foreground prompts: "a wine glass", "a glass bottle", "a cup filled with coke", "a cup of tea". All images have the resolution of 1024x1024.

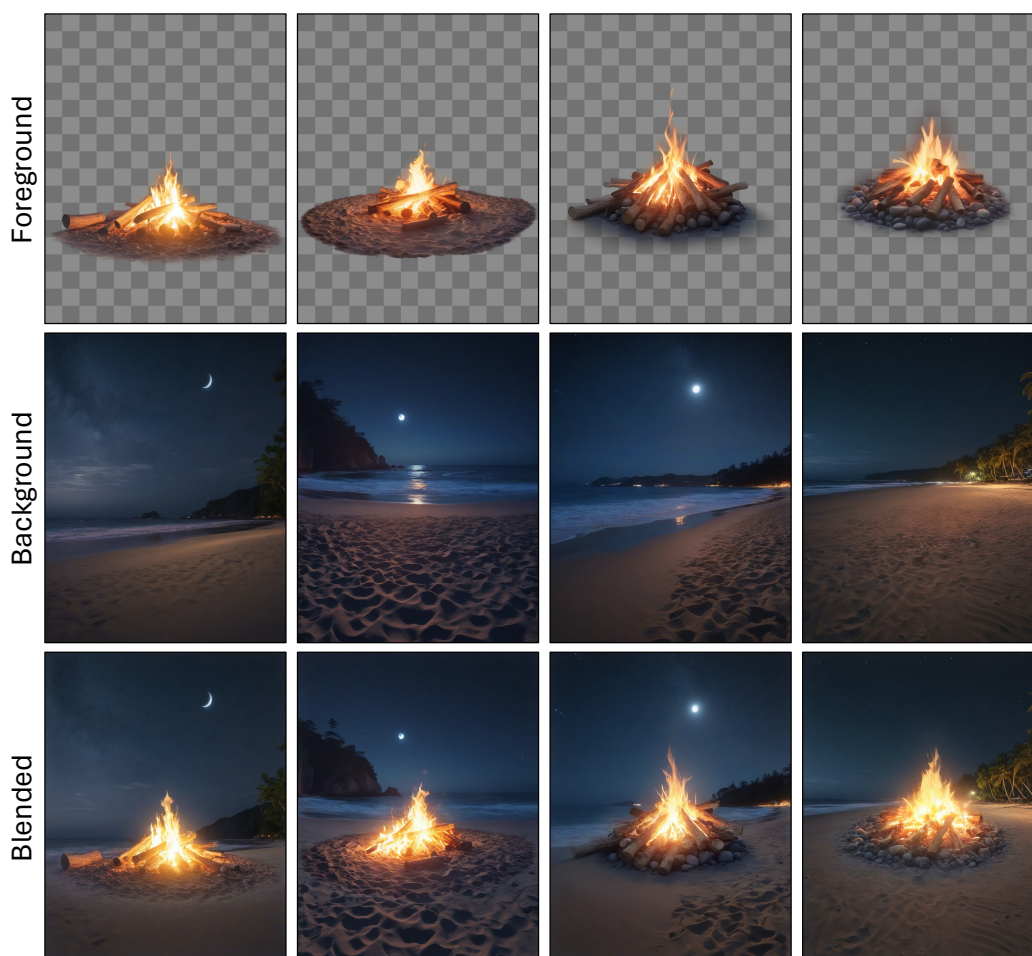


Figure 16: **Supplementary Generation Results for the subject "a campfire"**. We provide additional generation results for the foreground prompt "a campfire" and background prompt "a beach, night time." The image resolution is 896x1152 for all examples.



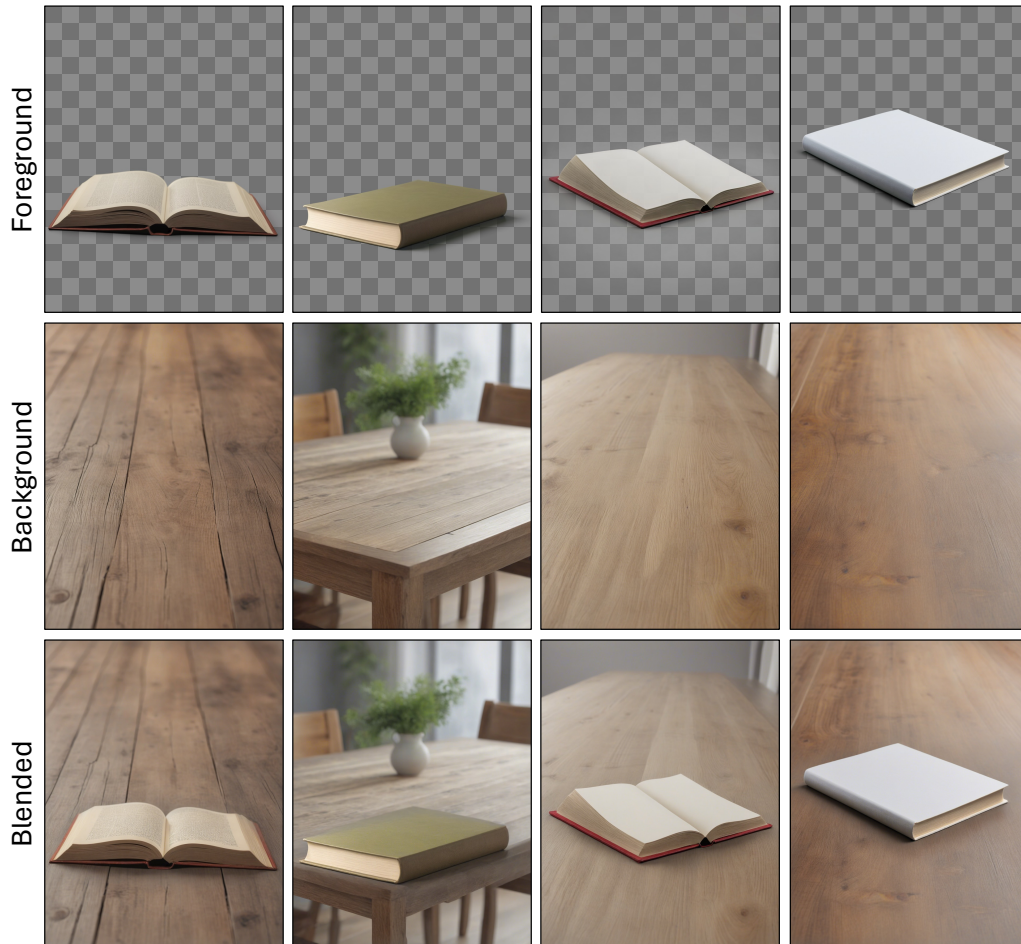


Figure 17: **Supplementary Generation Results for the subject “a book”**. We provide additional generation results for the foreground prompt “a book” and background prompt “a table”. The image resolution is 896x1152 for all examples.

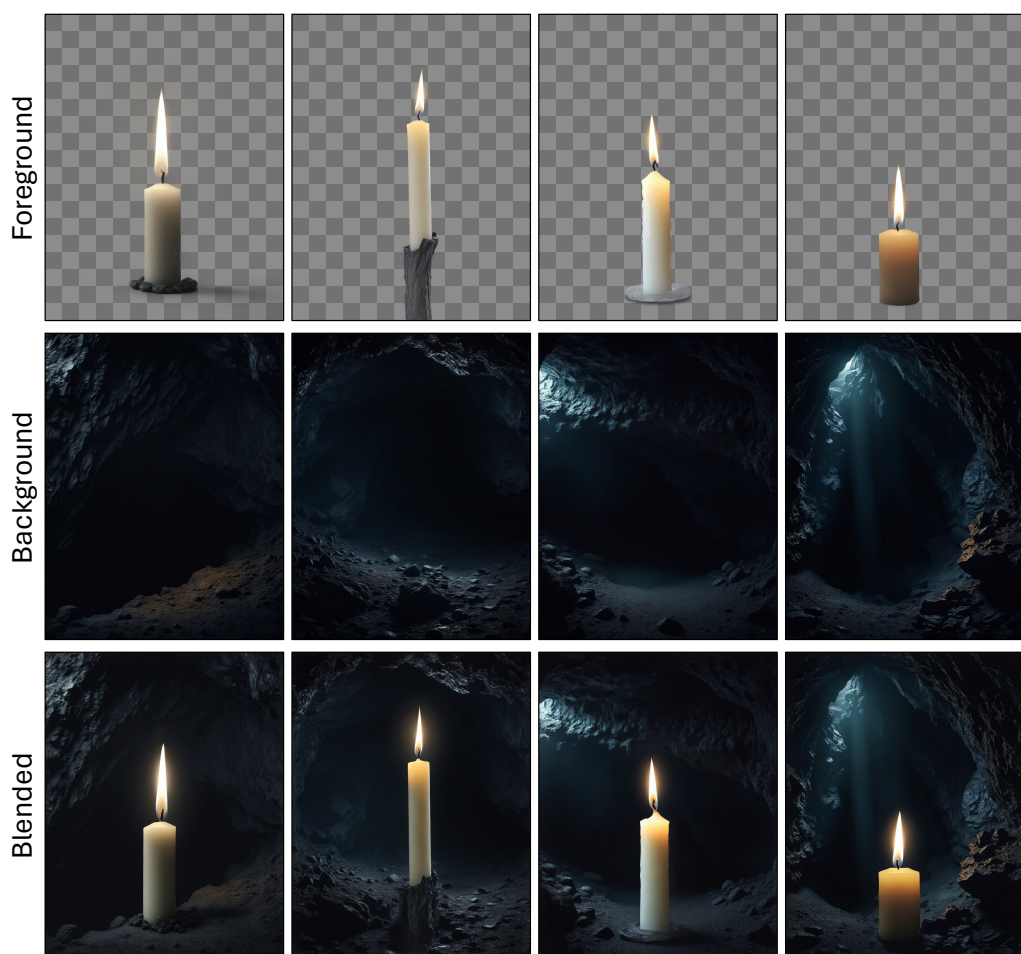


Figure 18: **Supplementary Generation Results for the subject “a candle”**. We provide additional generation results for the foreground prompt “a candle” and background prompt “a dark cave”. The image resolution is 896x1152 for all examples.

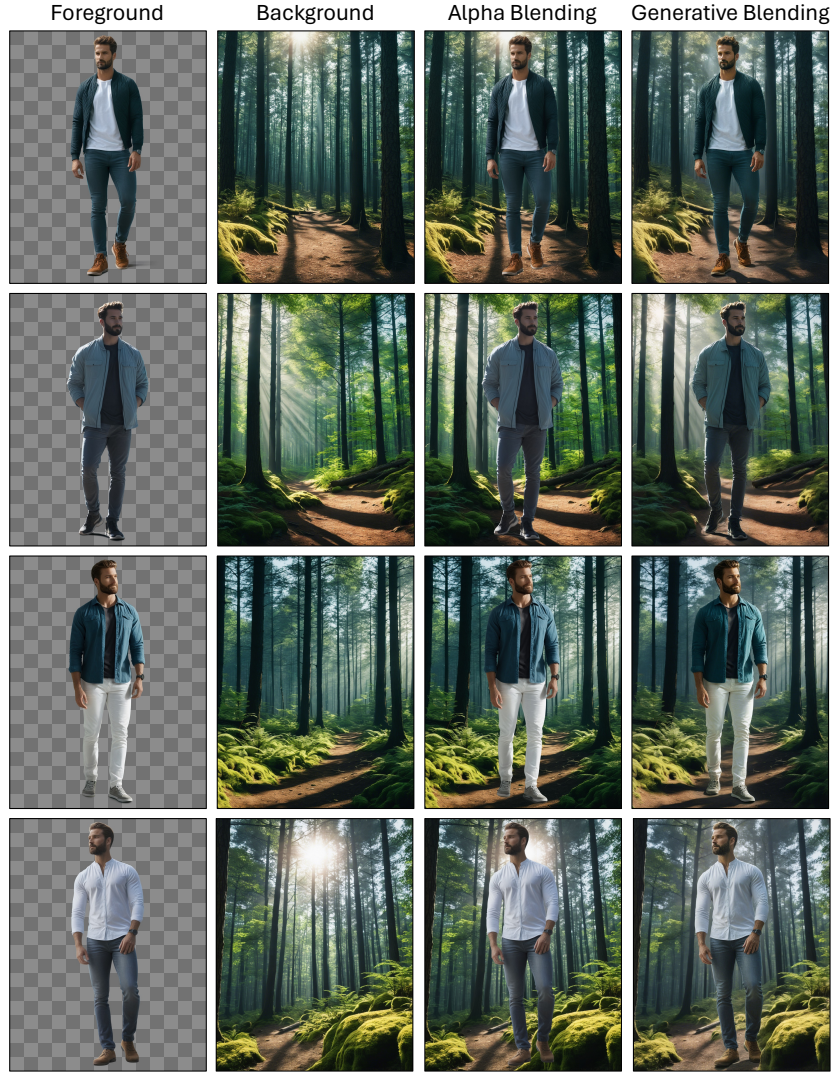


Figure 19: **Supplementary Generation Results for Grounding and Shadowing Effects.** We provide additional generation examples to demonstrate the grounding and shadowing capabilities of our framework. Our approach succeeds in both appropriate lighting compared to alpha blending (see rows 1, 2, 3), and can successfully ground the foreground on the background (row 4). We perform our generations with foreground prompt “a man, standing” and background prompt “a forest, daytime”. The image resolution is 896x1152 for all examples.