Harmonized Multi-Layer Text-to-Image Generation with Generative Priors

Supplementary Material

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. Limitations

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

C. Analyses on Structure Priors from Different Layers

In all of the experiments we provide, we utilize the structure prior extracted from the last attention map of the foreground diffusion model, $\epsilon_{\theta,FG}$. As a justification of this decision and to clearly illustrate what different self attention layers focus on throughout the generation process, we provide structure priors extracted from different layers in Fig. 9. As it can also be observed visually, the structure prior extracted from the last self attention layer provides a more precise estimate of the shape of the foreground being generated.

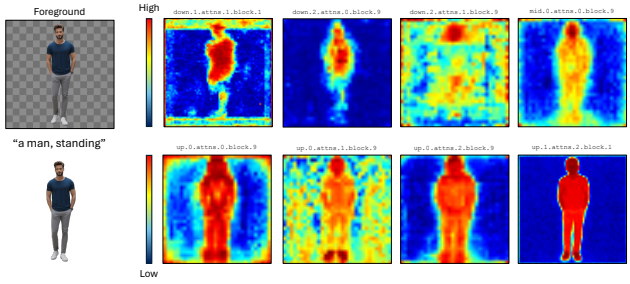

Figure 9. **Visualization of the structure priors from different self attention layers.** We visualize the structure priors extracted from different self attention layers of the foreground diffusion model, where the diffusion timestep is set as $t = 0.8T$. We visualize the structure priors from the self attention layer of each model block, follow the block definition of [11]. We follow the naming convention of diffusers ([31]). In all of our experiments, we use the structure prior from self attention layer `up.1.attns.2.block.1`.

D. Detailed Blending Algorithm

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

E. User Study Details

We conduct our user study over 50 participants with 40 image triplets generated by LayerFusion and [33]. For the generation of the subjected triplets, we generate examples with animal, vehicle, matte objects, person and objects with transparency properties as the foreground to get samples representing a diverse distribution of subjects. Following sample generation, we ask users to rate each image triplet from a scale of 1-to-5, with the following question: "Please rate the following image triplet from a scale of 1-to-5 (1 - unsatisfactory, 5 - very satisfactory) considering how realistic each image is and how naturally blended they are". The users are also supplied the foreground and background prompts used to generate the image triplet, for each method. We provide an example question from the conducted user study in Fig. 10.

Algorithm 1 Extracting the Structure Prior

Require: Foreground denoiser $\epsilon_{\theta,\mathrm{FG}}$, latent feature map $\mathbf{z}_t$, prompt $\mathbf{p}_{\mathrm{FG}}$
Ensure: Structure prior s'

function EXTRACTSTRUCTUREPRIOR($\epsilon_{\theta,\mathrm{FG}}, \mathbf{z}_t, \mathbf{p}_{\mathrm{FG}}$)
  $\hat{\epsilon}$, $m^L \leftarrow \epsilon_{\theta,\mathrm{FG}}(\mathbf{z}_t, \mathbf{p}_{\mathrm{FG}})$ ▷ Retrieve Noise Prediction and Last Self Attention Map
  for $i = 1 \to M$ **do** ▷ Row-wise sparsity (Eq. 6 in main text)
    $s_i \leftarrow \left(\sum_{j=1}^{M} (m^L_{i,j})^2\right)^{-1}$
  end for
  $s' \leftarrow 1 - \mathrm{norm}\left(\{s_i\}_{i=1}^{M}\right)$ ▷ Convert sparsity to density and normalize
  return s'
end function

Algorithm 2 Extracting the Content-Confidence Prior

Require: Foreground denoiser $\epsilon_{\theta,\mathrm{FG}}$, hidden states h, prompt $\mathbf{p}_{\mathrm{FG}}$
Ensure: Content prior c

function EXTRACTCONTENTPRIOR($\epsilon_{\theta,\mathrm{FG}}, h, \mathbf{p}_{\mathrm{FG}}$)
  attn_out, attn_probs $\leftarrow$ ATTENTION$_{\theta,\mathrm{FG}}(h, \mathbf{p}_{\mathrm{FG}})$ ▷ Cross-attention forward pass
  $n \leftarrow$ attn_probs
  $c \leftarrow \frac{1}{H}\sum_{k=1}^{H} n_{k,:,\texttt{<EOS>}}$ ▷ Average EOS channel over H heads
  return c
end function

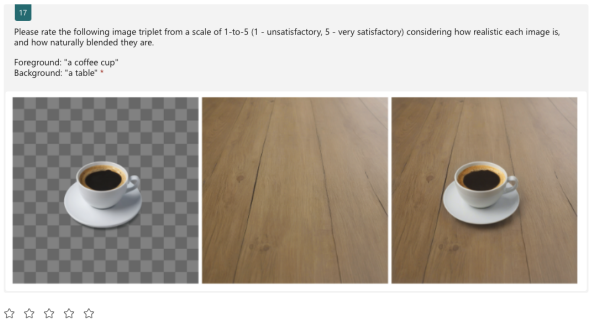

Figure 10. **Example Question from the User Study.** To evaluate the effectiveness our method perceptually, we conduct a user study over 40 generated image triplets. We provide an example question from this study for clarity. The users are shown an image triplet in the order of foreground, background and blended image and then asked to rate it from a scale of 1-to-5 (1 - unsatisfactory, 5 - very satisfactory).

F. Qualitative Comparisons on Blending Quality

To visually supplement the quantitative results presented in the main paper, Fig. 11 provides a direct qualitative comparison of our method against two latent blending baselines. Our attention-guided fusion excels at creating photorealistic compositions in which the subject naturally belongs to the environment. Notice how the hedgehog is not directly placed on the leaves but appears nestled within them, with the lighting on its spines perfectly matching the light of the forest floor. Similarly, the chameleon is grounded in its branch through coherent shadowing, and the monkey's fur correctly picks up the ambient green reflections of the dense jungle. This seamless integration of lighting, shadows, and environmental color is a direct result of our harmonization approach.

In contrast, baseline methods struggle to achieve this level of realism. Blended Latent Diffusion (BLD) [2] consistently produces poorly lit and severely discolored figures that appear disconnected from the scene. Although LayerDiffuse [33] achieves better blending, it suffers from a critical lack of color fidelity, drastically altering the natural hues of the foreground object, as seen with the chameleon. These examples clearly illustrate the superior coherence and realism of our method.

G. Supplementary Generation Results

In addition to the results provided in the main paper, we provide supplementary generation results in this section. Below, we include harmonized generations of a variety of subjects. We provide Fig. 12 to Fig. 22 as supplementary results.

Algorithm 3 Attention-Level Blending (one transformer block)

Require: Foreground denoiser $\epsilon_{\theta,\mathrm{FG}}$, RGB denoiser ϵ_θ, hidden states h_{FG}, h_{BL}, h_{BG}, prompts $\mathbf{p}_{\mathrm{FG}}$, $\mathbf{p}_{\mathrm{BG}}$, boundary coefficient d, structure prior s'
Ensure: Updated attention outputs a'_{FG}, a'_{BL}, a_{BG}

function AttnBlend($\epsilon_{\theta,\mathrm{FG}}, \epsilon_\theta, h_{\mathrm{FG}}, h_{\mathrm{BL}}, h_{\mathrm{BG}}, \mathbf{p}_{\mathrm{FG}}, \mathbf{p}_{\mathrm{BG}}, d, s'$)
  $h^\ell_{\mathrm{FG}}, h^\ell_{\mathrm{BL}}, h^\ell_{\mathrm{BG}} \leftarrow \mathrm{LayerNormCrossAttn}(h_{\mathrm{FG}}, h_{\mathrm{BL}}, h_{\mathrm{BG}})$ ▷ Layer normalization (shared parameters)
  $c \leftarrow \mathrm{ExtractContentPrior}(\epsilon_{\theta,\mathrm{FG}}, h^\ell_{\mathrm{FG}}, \mathbf{p}_{\mathrm{FG}})$ ▷ Layer-specific content prior
  ▷ Soft and hard masks (Eq. 7)
  $\mathrm{mask}_{\mathrm{soft}} \leftarrow \mathrm{norm}(s' \odot c)$
  $\mathrm{mask}_{\mathrm{hard}} \leftarrow \sigma\big(d\,(\mathrm{mask}_{\mathrm{soft}} - 0.5)\big)$
  ▷ Cross-attention for each branch
  $a_{\mathrm{BG}}, a_{\mathrm{BL}} \leftarrow \mathrm{Attention}_\theta\big([h^\ell_{\mathrm{BG}}, h^\ell_{\mathrm{BL}}], \mathbf{p}_{\mathrm{BG}}\big)$
  $a_{\mathrm{FG}} \leftarrow \mathrm{Attention}_{\theta,\mathrm{FG}}(h^\ell_{\mathrm{FG}}, \mathbf{p}_{\mathrm{FG}})$
  ▷ Blending rules (Eqs. 8–9)
  $a'_{\mathrm{BL}} \leftarrow a_{\mathrm{FG}} \odot \mathrm{mask}_{\mathrm{soft}} + a_{\mathrm{BL}} \odot (1 - \mathrm{mask}_{\mathrm{soft}})$
  $a'_{\mathrm{FG}} \leftarrow a'_{\mathrm{BL}} \odot \mathrm{mask}_{\mathrm{hard}} + a_{\mathrm{FG}} \odot (1 - \mathrm{mask}_{\mathrm{hard}})$
  return $a'_{\mathrm{FG}}, a'_{\mathrm{BL}}, a_{\mathrm{BG}}$
end function

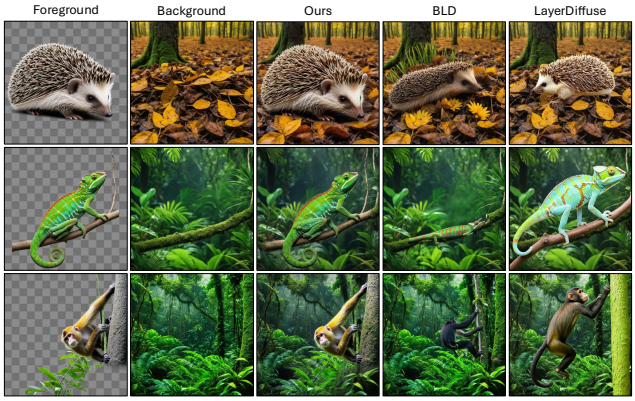

Figure 11. **Comparisons with Latent Blending Baselines.** We provide qualitative comparisons with latent blending baselines. In correspondence to a triplet generated by our method, we perform background-conditioned generations using the same generation prompts for BLD [2] and LayerDiffuse. Supplementary 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.

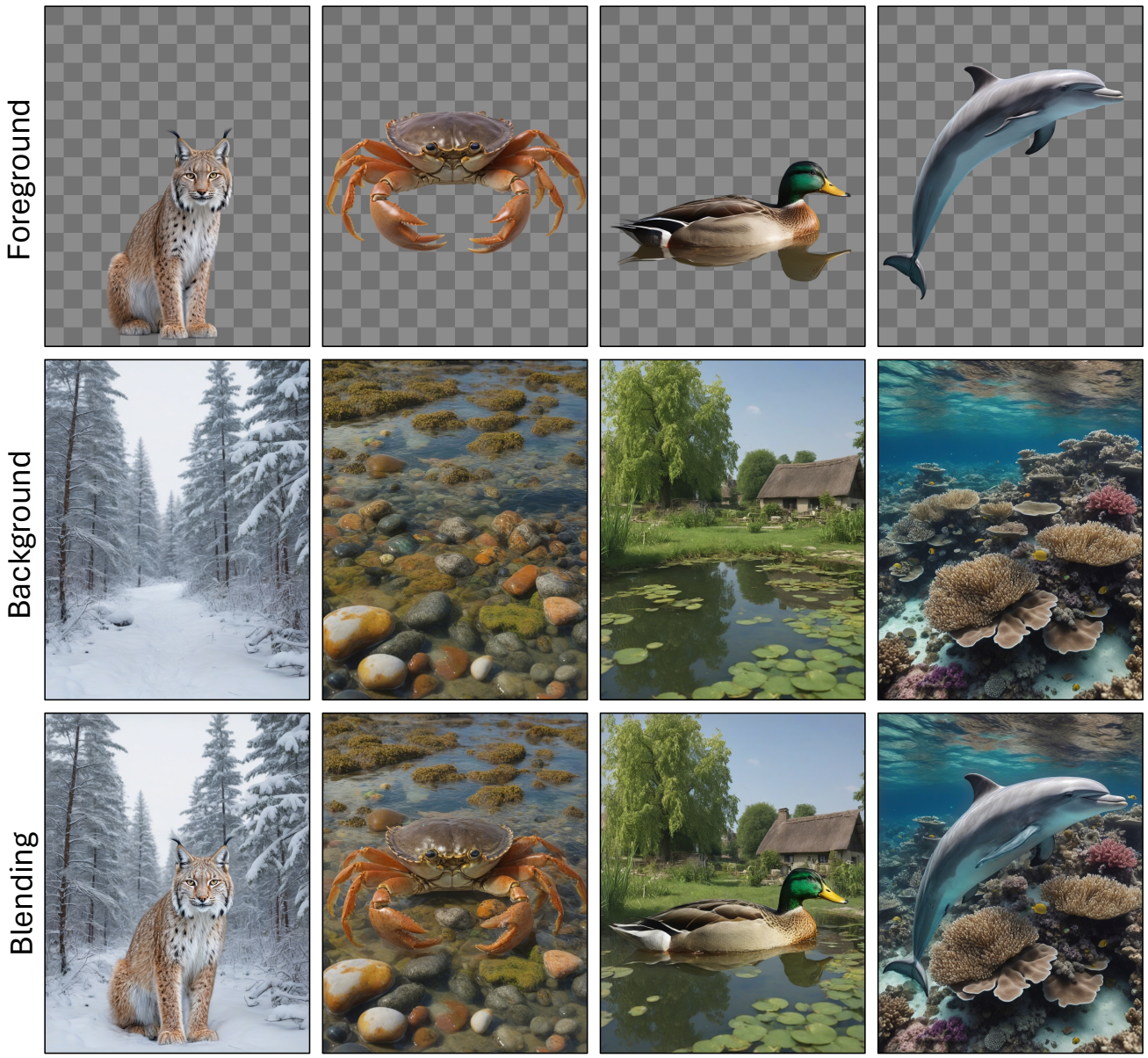

Figure 12. **Supplementary Generation Results with animal subjects.** Supplementary results with image resolution 896x1152. The foreground & background prompt pairs from left to right are: “a lynx”, “a snowy forest”), (“a crab”, “a rocky tide pool”), (“a duck”, “a village pond”), (“a dolphin”, “a crystal-clear coral reef”)

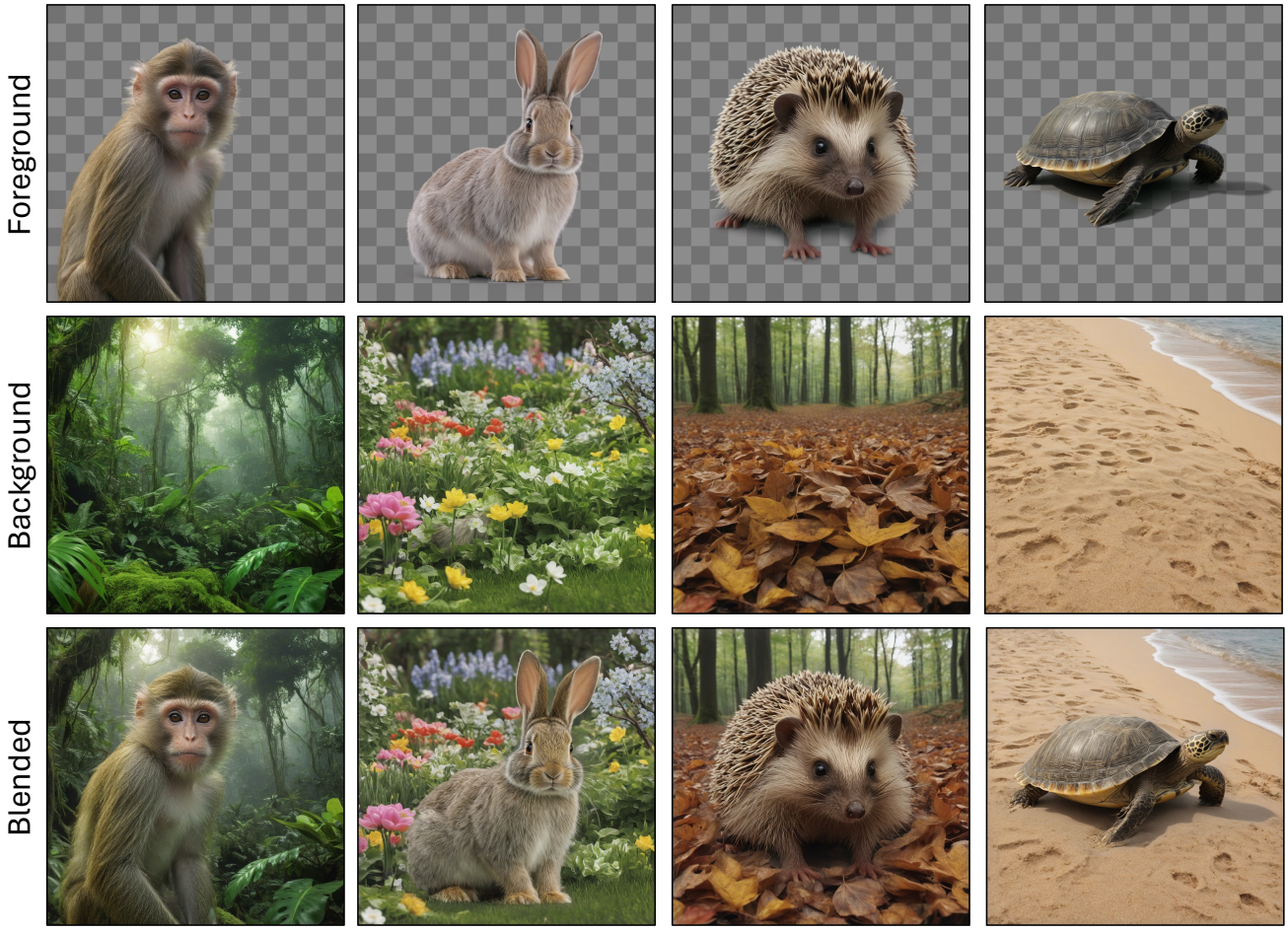

Figure 13. **Supplementary Generation Results with animal subjects.** Supplementary results with image resolution 1024x1024. The foreground & background prompt pairs from left to right are: (“a monkey”, “a vibrant tropical rainforest”), (“a rabbit”, “a backyard garden”), (“a hedgehog”, “a forest floor covered in leaves”), (“a turtle”, “a warm sandy beach”)

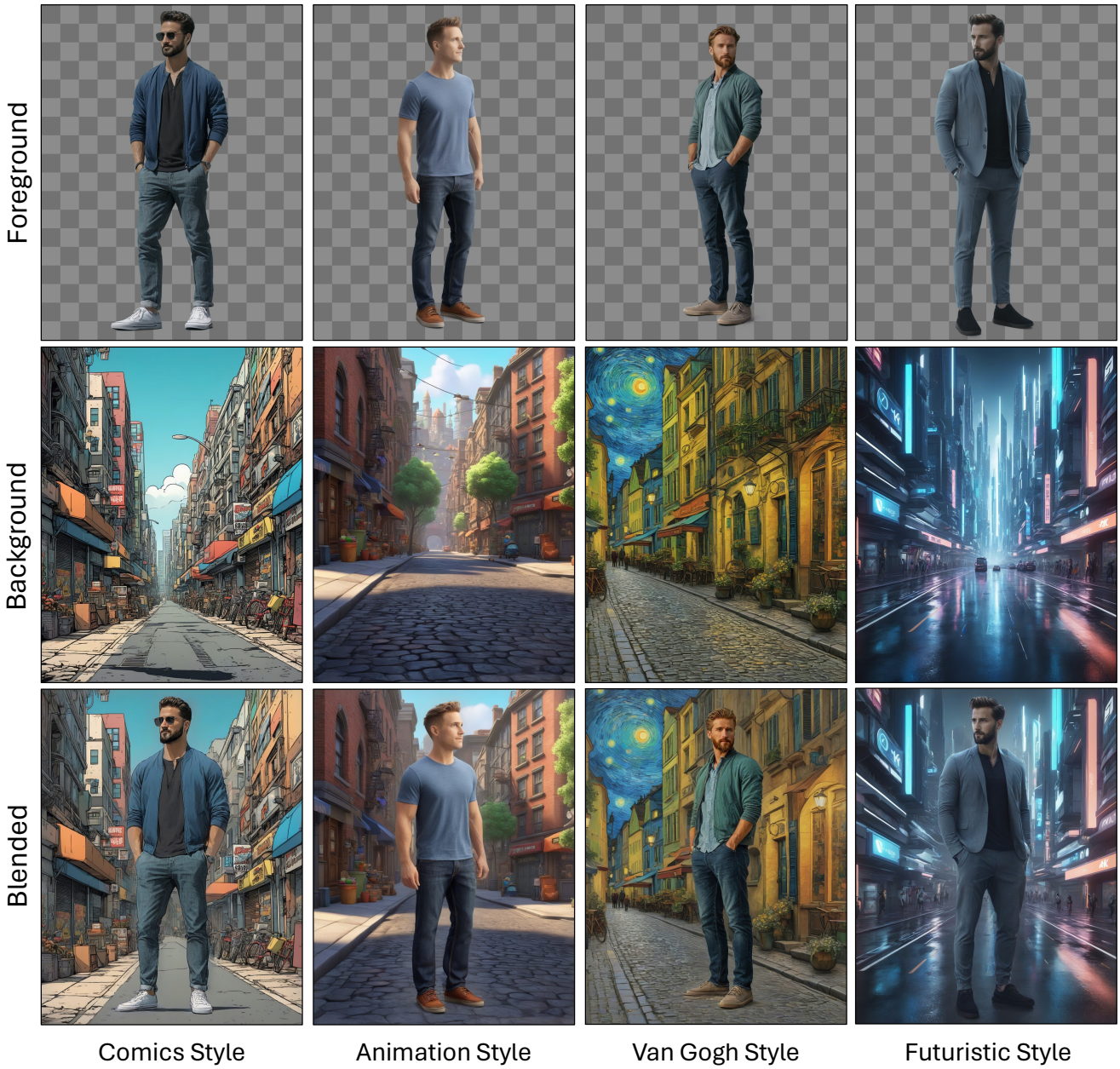

Figure 14. **Supplementary Generation Results with stylization prompts.** We provide additional examples with stylization prompts to demonstrate the harmonization capabilities of our method. For each image triplet, we generate the examples with the prompt set ("a man, standing", "a street, `style_name`") 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 15. **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.

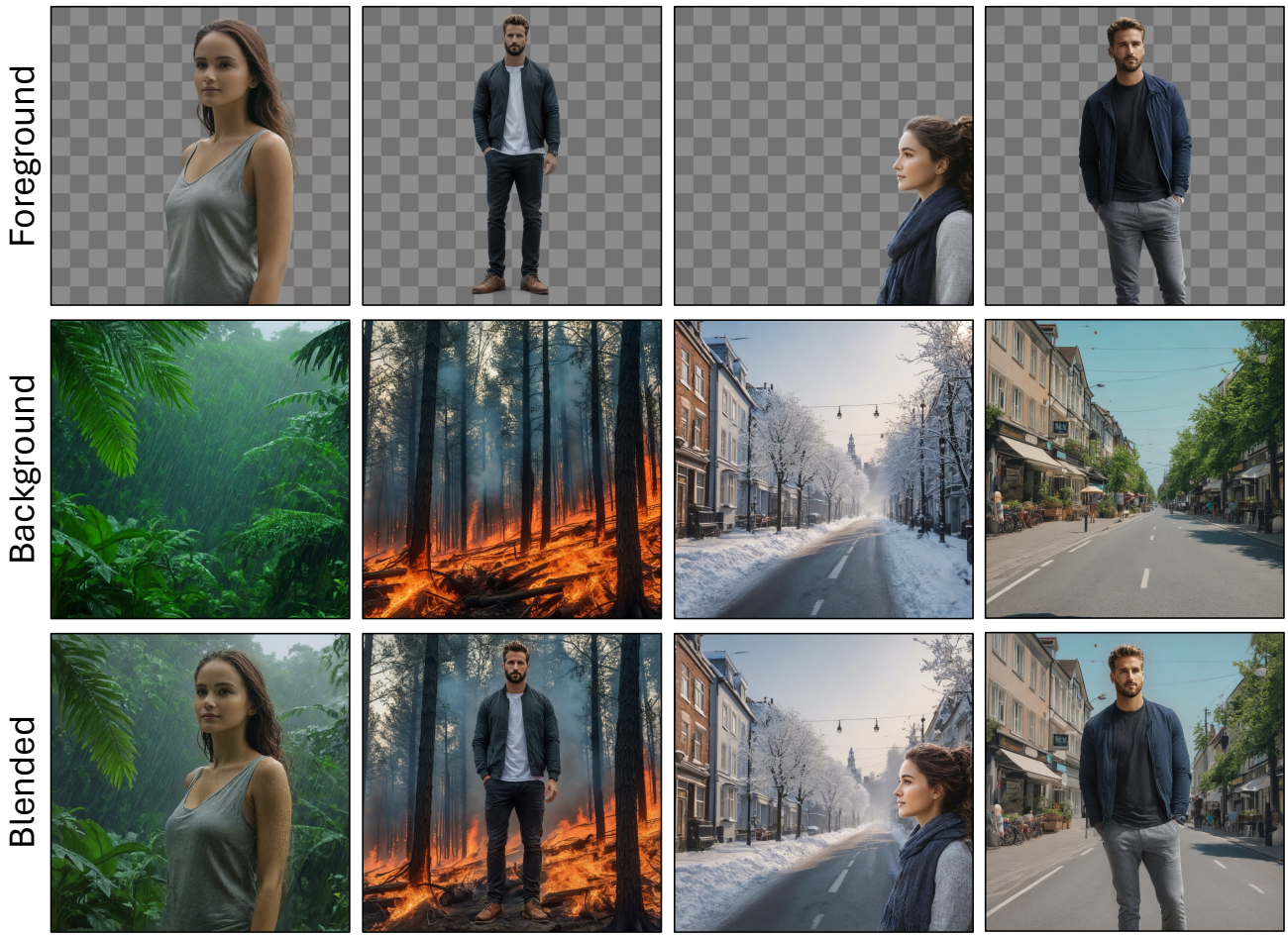

Figure 16. **Supplementary Generation Results with human subjects.** We provide additional examples with human subjects with different background prompts. The background prompts used are "a rainy jungle", "a forest in fire", "a street, winter time", "a street, daytime". Note that depending on the background 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.

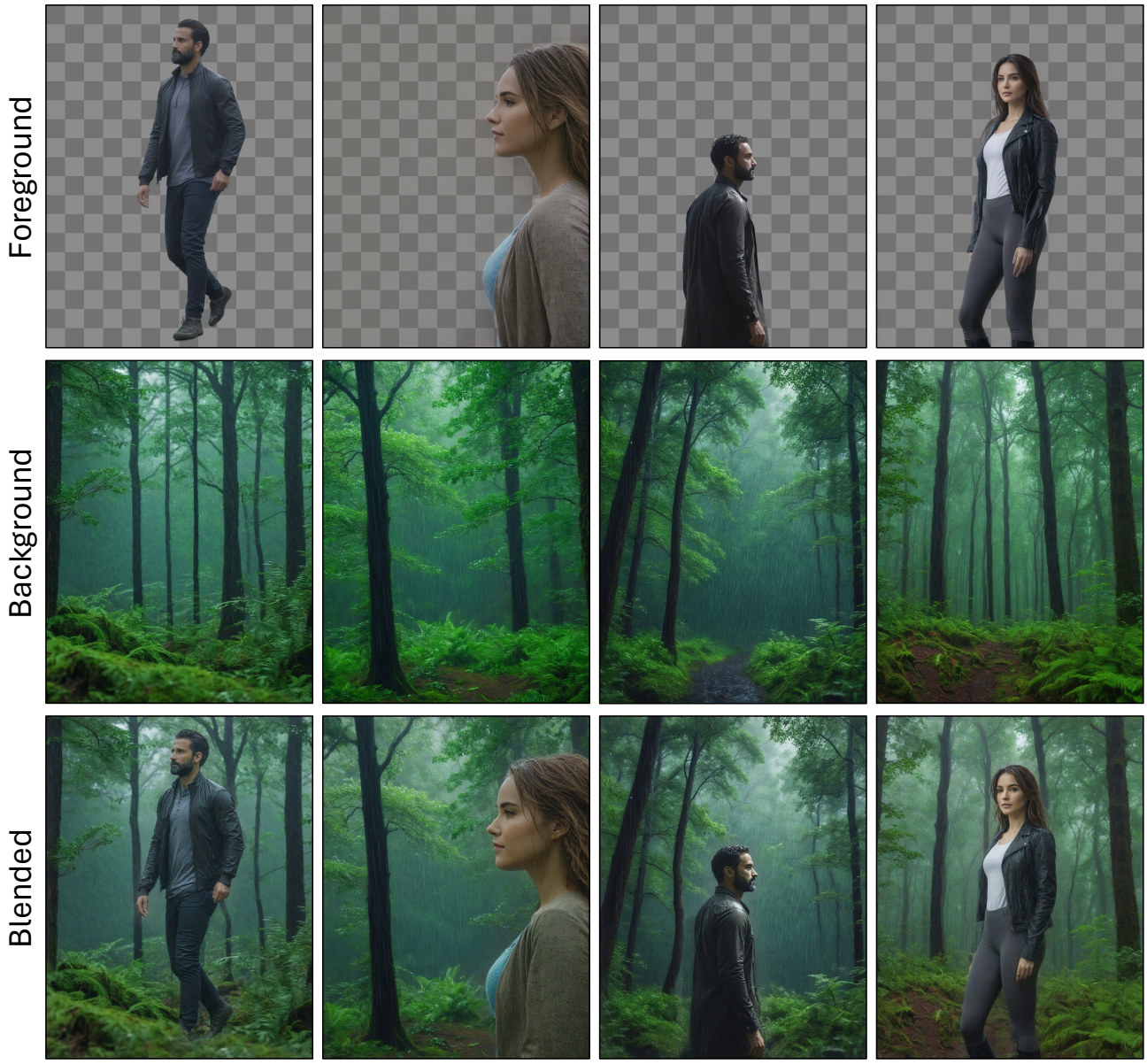

Figure 17. **Supplementary Generation Results for the background "a rainy forest".** For each of the images, the background prompt "a rainy forest" is used to generate the background image. As it can also be observed from the provided examples, the background creates an influence over the foreground (e.g. wetness effect on the human subjects). The image resolution is 896x1152 for all examples.

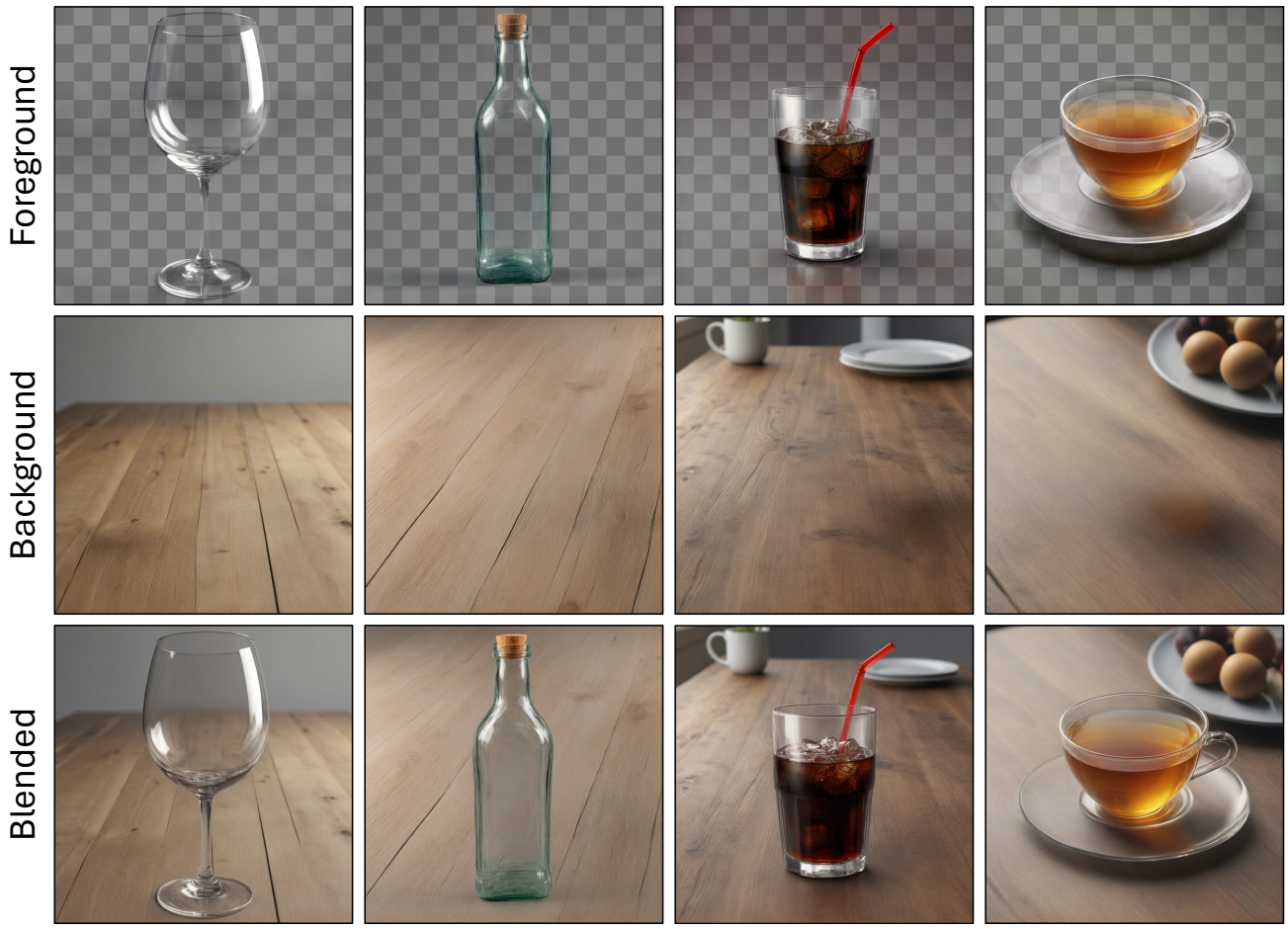

Figure 18. **Supplementary Generation Results for subjects with transparency property.** To demonstrate that our framework is able to preserve the transparency properties of layered image representations, we provide additional results here. With the background prompt ”a table” we use the following foreground prompts: “a wine glass”, “a glass bottle”, “a cup filled with coke”, “a cup of tea”. All images have the resolution of 1024x1024.

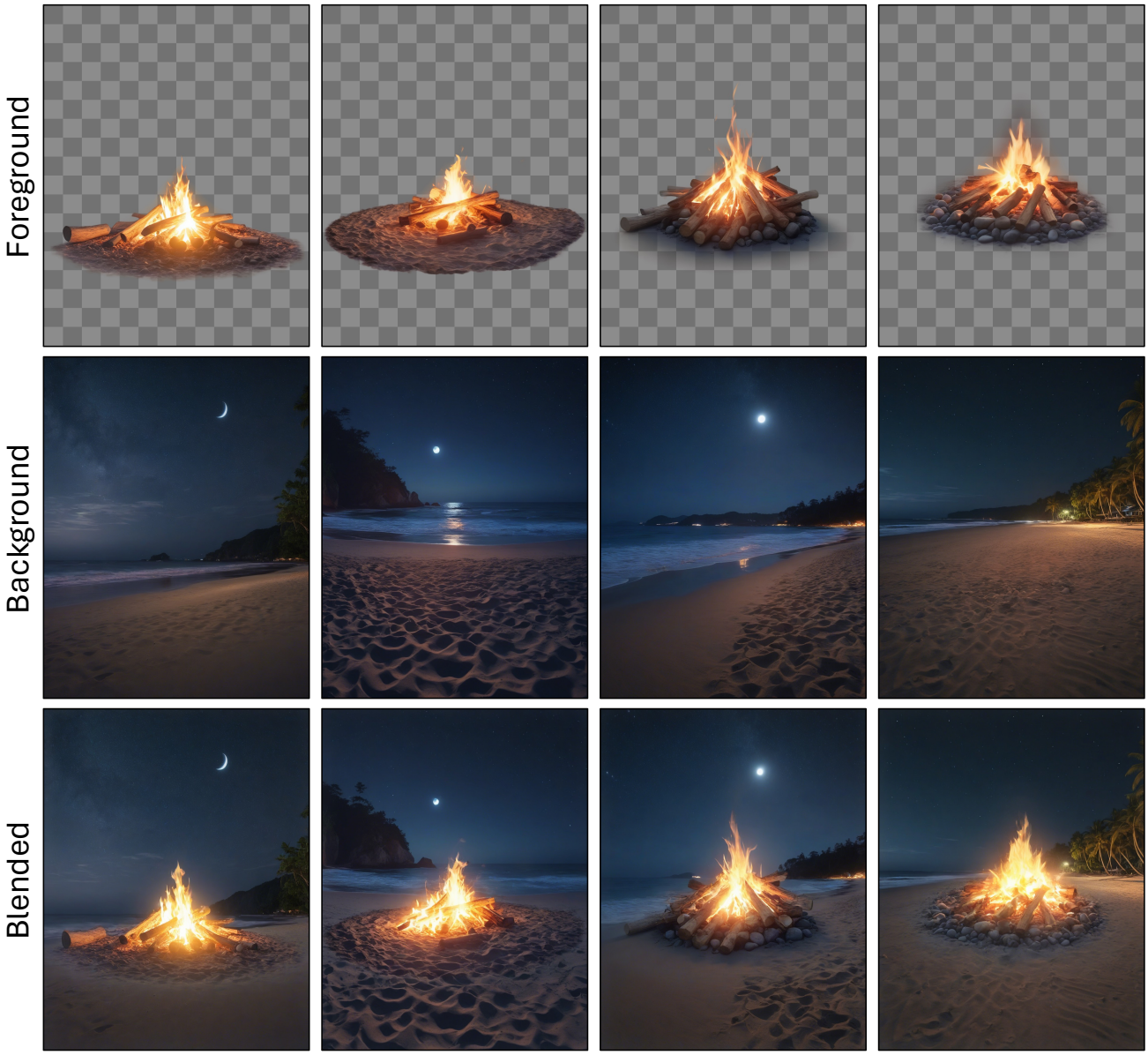

Figure 19. **Supplementary Generation Results for the subject "a campfire".** We provide additional generation results for the foreground prompt "a campfire" and background prompt "a beach, night time." The image resolution is 896x1152 for all examples.

Figure 20. **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.

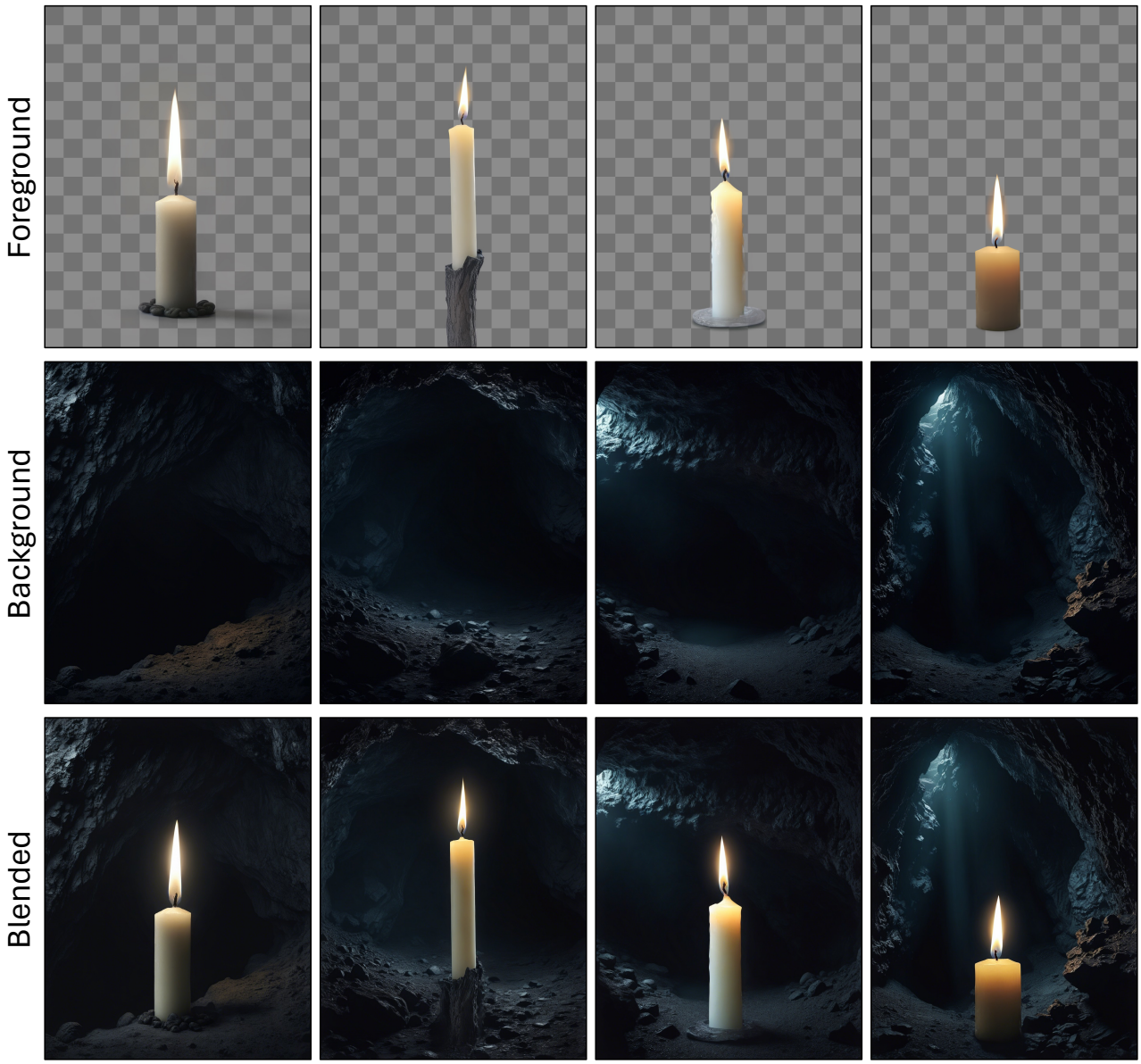

Figure 21. **Supplementary Generation Results for the subject "a candle".** We provide additional generation results for the foreground prompt "a candle" and background prompt "a dark cave". The image resolution is 896x1152 for all examples.

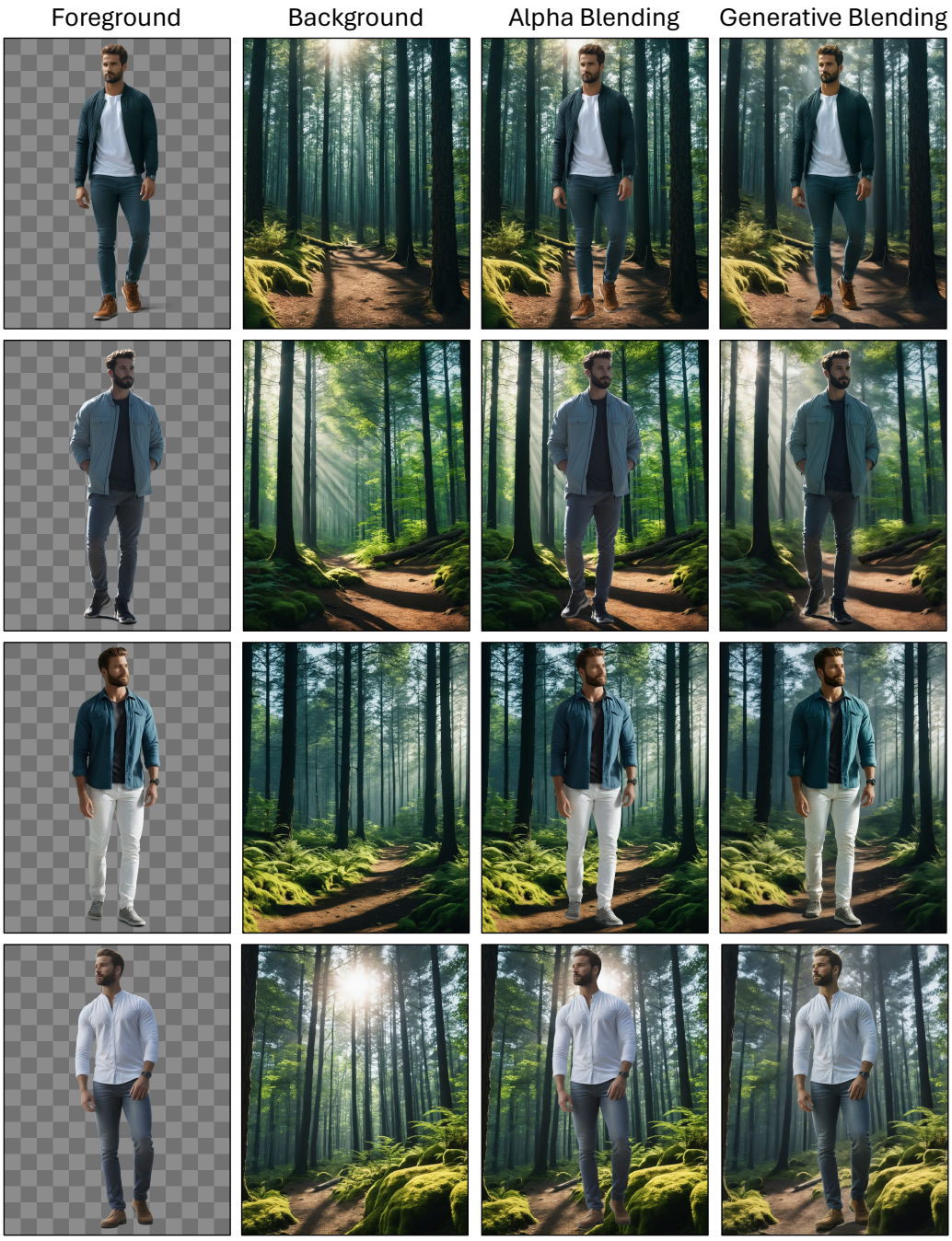

Figure 22. **Supplementary Generation Results for Grounding and Shadowing Effects.** We provide additional generation examples to demonstrate the grounding and shadowing capabilities of our framework. Our approach succeeds in both appropriate lighting compared to alpha blending (see rows 1, 2, 3), and 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.