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

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large-scale diffusion models have achieved remarkable success in generating high-quality images from textual descriptions, gaining popularity across various applications. However, the generation of layered content, such as transparent images with foreground and background layers, remains an under-explored area. Layered content generation is crucial for creative workflows in fields like graphic design, animation, and digital art, where layer-based approaches are fundamental for flexible editing and composition. In this paper, we propose a novel image generation pipeline based on Latent Diffusion Models (LDMs) that generates images with two layers: a foreground layer (RGBA) with transparency information and a background layer (RGB). Unlike existing methods that generate these layers sequentially, our approach introduces a harmonized generation mechanism that enables dynamic interactions between the layers for more coherent outputs. We demonstrate the effectiveness of our method through extensive qualitative and quantitative experiments, showing significant improvements in visual coherence, image quality, and layer consistency compared to baseline methods.

1 Introduction

Layered content generation, particularly creating images with transparency, is crucial in creative industries that rely on layer-based composition. While large-scale diffusion models [7, 6, 8] excel at generating single images, their application to compositional layer generation remains underexplored. Recent efforts have shown promise in generating single transparent images [10, 4, 2, 9], highlighting the need to better align these models with established creative workflows.

However, harmonizing distinct layers into a visually coherent composite presents significant challenges. First, realistic composition demands interactions beyond simple alpha blending, such as grounding, shadows, and illumination, which are difficult to achieve when layers are generated independently. Second, preserving continuous transparency for elements like glass or reflections is a problem for pipelines that rely on segmentation, as binary masks cannot capture the nuances of an alpha channel.

Addressing these issues, we propose a novel latent diffusion model pipeline that generates a transparent foreground layer (RGBA) and a background layer (RGB) simultaneously. Our method facilitates harmonized interaction between the layers, in contrast to sequential generation approaches [4, 10] that can lead to inconsistencies. The core of our framework is a novel **attention-level blending** mechanism, which utilizes cross-attention and self-attention masks to guide the co-generation of both layers. This allows for dynamic, fine-grained interactions that improve the realism and coherence of the final composition.

Our contributions are: (1) a new pipeline that generates harmonized foreground (RGBA) and background (RGB) layers with natural interactions; (2) a novel attention-level blending scheme for

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

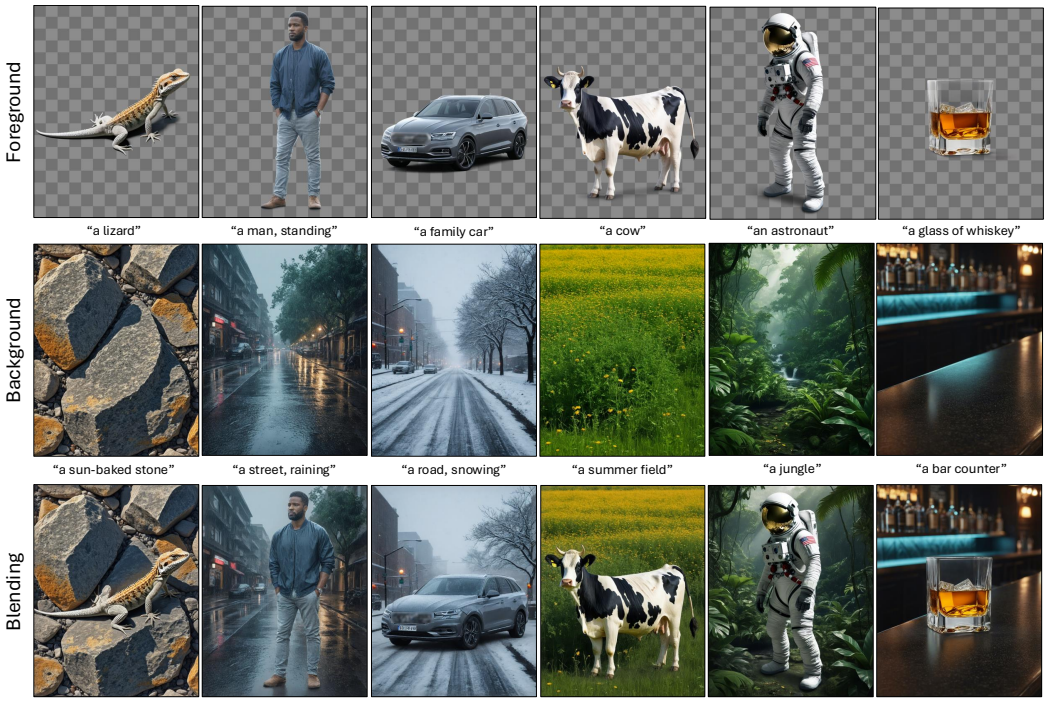

Figure 1: **Qualitative Results.** We present qualitative results on multi-layer generation over different visual concepts. In each column, we show the high-quality results of foreground layer, background layer and their generative blending respectively, in terms of text-image alignment, transparency and harmonization. We present more results in the supplementary material.

seamless and cohesive composition; and (3) extensive experiments demonstrating that our method outperforms baselines in visual coherence, image quality, and layer consistency.

2 Methodology

Our pipeline returns a *foreground image with an α channel* (RGBA), a *background RGB* image, and their *composite RGB* in a single sampling run, doing so with **no weight updates**. We reuse the transparent–foreground branch of **LayerDiffuse** [10], denoted $\epsilon_{\theta,\mathrm{FG}}$, and team it with a frozen, off-the-shelf text-to-image diffusion model ϵ_θ, which is responsible for both the background and the blended view. Both denoisers share the same timestep schedule and exchange information *only* through their attention tensors; therefore the pretrained output distributions of the two networks remain unchanged.

2.1 Priors from attention

Our blending strategy is driven by two complementary guidance maps. One captures *static structure* which locates the boundary of the foreground object (structure prior), while the other captures *layer-wise content confidence*, how strongly the prompt steers each block of the network as the image forms (content confidence prior).

Structure prior. We extract the structure prior once, from the *last* self-attention block of the foreground denoiser $\epsilon_{\theta,\mathrm{FG}}$, because this layer is closest to the point where the object appears sharply in latent space. Let $m^L \in \mathbb{R}^{M\times M}$ be the head-averaged self-attention matrix of that block (L is the deepest layer index). For each spatial token i we compute $s_i = \left(\sum_{j=1}^{M}(m_{ij}^L)^2\right)^{-1}$, $s'_i = 1-\mathrm{norm}(s_i)$, yielding a fixed map $s' \in [0,1]^M$ that highlights tokens whose attention rows are dense, typical for the foreground object (Fig. 2). This map is reused, unchanged, at every transformer depth during fusion.

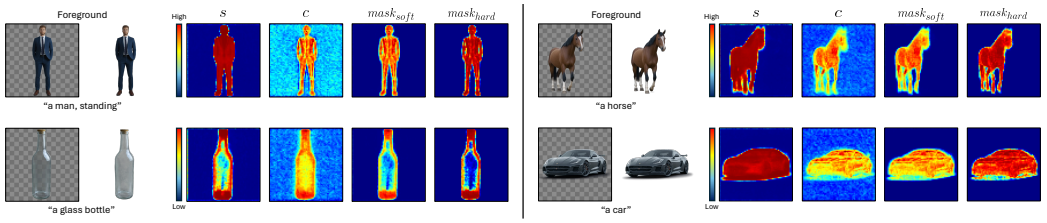

Figure 2: **Visualization of the masks extracted as generative priors.** Throughout the generation process, we extract a structure prior s and a content confidence prior c. To combine the structure and content information, we construct $mask_{soft}$ and $mask_{hard}$ during the blending process. As visible from the provided maps (as priors), We can both capture the overall object structure with the structure prior s and incorporate the content with c, where their combination provides a precise mask reflecting both quantities (see the example "the car"). Also note that the masks we construct also capture transparency information throughout the masking process (see the example "a glass bottle"). We retrieve the provided masks for the diffusion timestep $t = 0.8T$.

Content-confidence prior. While foreground location is largely stable across depth, the influence of the text prompt evolves from coarse to fine detail. We therefore compute a content-confidence map for *every* transformer block. For a given block index ℓ, we average the H cross-attention heads and keep the channel that corresponds to the <EOS> text token, obtaining $c^{\ell} \in [0,1]^{M} \quad (\ell = 1, \ldots, L)$. Because CLIP's encoder is unidirectional, the <EOS> token accumulates information from the entire prompt, so c^{ℓ} offers a compact, up-to-date measure of how strongly each token is being driven by the text at depth ℓ (Fig. 2). These attention layer-specific maps allow our masks to adapt as the image content sharpens through the network.

Together, the layer-invariant structure prior s' and the layer-dependent content maps $\{c^{\ell}\}$ supply the spatial weights used to build the soft and hard blending masks in the next section.

2.2 Mask extraction

We combine the two attention-based priors into a pair of spatial masks that will guide all later fusion steps. First, we form a *soft mask* by an element-wise product of the structure scores $s' \in [0,1]^{M}$ and the content-confidence map $c \in [0,1]^{M}$, followed by a min–max rescaling over the token set, $mask_{soft} = \mathrm{norm}\big(s' \odot c\big)$. This mask holds fractional weights in $[0,1]$ and therefore supports smooth interpolation between foreground and background activations. For operations that require a crisper decision we pass the soft mask through a centred sigmoid gate, $mask_{hard} = \sigma\big(d\,(\mathrm{mask}_{\mathrm{soft}} - 0.5)\big)$, d=10 where the slope parameter d controls the hardness of the boundary (see supplement for a sensitivity analysis). Figure 2 visualizes typical soft and hard masks produced by this two-step process.

2.3 Attention-level Blending

During each transformer block our method maintains three latent streams: $\mathcal{L}_{\mathrm{FG}}$ for the foreground branch of $\epsilon_{\theta,\mathrm{FG}}$, $\mathcal{L}_{\mathrm{BL}}$ for the composite branch of ϵ_{θ}, and $\mathcal{L}_{\mathrm{BG}}$ for the background branch of the same model. Let $a_{\mathrm{FG}}, a_{\mathrm{BL}}, a_{\mathrm{BG}} \in \mathbb{R}^{M \times D}$ denote the self-attention outputs of the current block, where M is the number of spatial tokens and D the hidden dimension. Guided by the masks, we first inject foreground information into the composite stream through a soft, convex combination $a'_{\mathrm{BL}} = a_{\mathrm{FG}} \odot \mathrm{mask}_{\mathrm{soft}} + a_{\mathrm{BL}} \odot (1 - \mathrm{mask}_{\mathrm{soft}})$, which copies foreground activations wherever the mask weight is high while leaving the remainder of the composite activations untouched. The updated composite activations then feed back into the foreground stream through the harder binary mask $a'_{\mathrm{FG}} = a'_{\mathrm{BL}} \odot \mathrm{mask}_{\mathrm{hard}} + a_{\mathrm{FG}} \odot (1 - \mathrm{mask}_{\mathrm{hard}})$, ensuring that both branches stay in sync over regions judged to belong to the foreground. Finally, the same soft-mask operation is applied between a'_{BL} and a_{BG}, allowing the background stream to adapt to the evolving composite without overwriting areas that should remain purely background. Identical equations are applied to the cross-attention outputs of the block. Because these updates involve only element-wise arithmetic, and require no gradient computation; full pseudo-code is given in Algorithm 1 of the supplementary material.

Table 1: **Quantitative Results.** We quantitatively evaluate the output distribution for the foreground and background images with CLIP-score, KID, and FID metrics. Furthermore, we also conduct a user study to evaluate the blending performance of our framework perceptually.

	Foreground			Background			Blending
	CLIP	KID	FID	CLIP	KID	FID	User Preference
LayerDiffuse [10]	38.46	0.0014	0.09	38.27	0.0400	1.17	2.960 ± 0.692
Ours	**38.97**	**0.0012**	0.09	**41.95**	**0.0058**	**0.14**	$\mathbf{3.233 \pm 0.566}$

3 Experiments

In all of our experiments, we use SDXL model as the diffusion model. Following the implementation released by [10], we use the model checkpoint `RealVisXL_V4.0`[1], unless otherwise stated. While using the non-finetuned SDXL, ϵ_θ as the background and blended image generators, we use the weights released by [10] for the foreground diffusion model $\epsilon_{\theta,FG}$[2]. We conduct all of our experiments on a single NVIDIA L40 GPU.

3.1 Quantitative Results

Quantitative Results We compare our framework with [10], a state-of-the-art method that generates a transparent foreground, an RGB background, and their blended result. Our evaluation assesses the quality of the individual layers and the final composite. For the foreground and background layers, we measure text-prompt alignment using CLIP score [5]. We also evaluate distributional realism using FID [3] and KID [1] scores, comparing our outputs to a reference distribution of foregrounds from [10] and backgrounds from a base SDXL model. This analysis confirms our method improves background quality and alignment with the base model while maintaining high-fidelity foreground generation.

User Study To assess the perceptual harmony of the final blended image, we conducted a user study with 50 participants. Participants were shown 40 image triplets (foreground, background, and our blended composite) and asked to rate the quality of the final blend on a 1-to-5 scale (1=not satisfactory, 5=very satisfactory). The results, presented in Table 1, show our method receives significantly higher ratings, confirming that our compositions are more visually coherent and appealing. Additional details about the study setup are in the supplementary material.

4 Discussion

Limitations and Future Directions The present work concentrates on the widely used two-layer case (foreground + background); extending our attention-guided fusion to richer multi-layer or hierarchical scenes is an exciting next step. Our results already benefit greatly from self- and cross-attention masks, yet further gains are possible with more robust mask extraction or lightweight refinement. Finally, because we rely on frozen latent-diffusion checkpoints, the method inherits their scene priors (for example, a mild bias toward centered subjects)—opening opportunities for bias-mitigation strategies or task-specific fine-tuning. A fuller discussion and illustrative failure cases appear in the supplementary material.

Conclusion We present an attention-guided diffusion pipeline that produces a harmonized foreground RGBA, a clean background RGB, and their composite in a single training-free pass. Given structure and content confidence priors, extracted from the frozen foreground branch, steers an unmodified SDXL backbone so both layers evolve together, yielding visually coherent results. Qualitative, quantitative, and user-study evaluations demonstrate clear gains over layered-generation and latent-blending baselines. Future work will extend the method to multi-layer scenes and release a public layered-image dataset to support tasks such as text-guided inpainting and advanced harmonization.

[1] `https://huggingface.co/SG161222/RealVisXL_V4.0`

[2] `https://huggingface.co/lllyasviel/LayerDiffuse_Diffusers`

References

[1] Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference on Learning Representations (2018)

[2] Burgert, R.D., Price, B.L., Kuen, J., Li, Y., Ryoo, M.S.: Magick: A large-scale captioned dataset from matting generated images using chroma keying. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22595–22604 (2024)

[3] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

[4] Quattrini, F., Pippi, V., Cascianelli, S., Cucchiara, R.: Alfie: Democratising rgba image generation with no $$$. arXiv preprint arXiv:2408.14826 (2024), https://arxiv.org/pdf/2408.14826

[5] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[6] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)

[7] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

[8] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)

[9] Wang, L., Li, Y., Chen, Z., Wang, J.H., Zhang, Z., Zhang, H., Lin, Z., Chen, Y.: Transpixar: Advancing text-to-video generation with transparency. arXiv preprint arXiv:2501.03006 (2025)

[10] 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