Alfie:Democratising RGBA Image Generation With No $$$

Fabio Quattrini, Vittorio Pippi,
Silvia Cascianelli, and Rita Cucchiara

University of Modena and Reggio Emilia, Modena, Italy
{name.surname}@unimore.it

Abstract. Designs and artworks are ubiquitous across various creative fields, requiring graphic design skills and dedicated software to create compositions that include many graphical elements, such as logos, icons, symbols, and art scenes, which are integral to visual storytelling. Automating the generation of such visual elements improves graphic designers' productivity, democratizes and innovates the creative industry, and helps generate more realistic synthetic data for related tasks. These illustration elements are mostly RGBA images with irregular shapes and cutouts, facilitating blending and scene composition. However, most image generation models are incapable of generating such images and achieving this capability requires expensive computational resources, specific training recipes, or post-processing solutions. In this work, we propose a fully-automated approach for obtaining RGBA illustrations by modifying the inference-time behavior of a pre-trained Diffusion Transformer model, exploiting the prompt-guided controllability and visual quality offered by such models with no additional computational cost. We force the generation of entire subjects without sharp croppings, whose background is easily removed for seamless integration into design projects or artistic scenes. We show with a user study that, in most cases, users prefer our solution over generating and then matting an image, and we show that our generated illustrations yield good results when used as inputs for composite scene generation pipelines. We release the code at https://github.com/aimagelab/Alfie.

Keywords: Diffusion Transformers · Generative AI · Graphic Design

1 Introduction

We are surrounded by a lot of visual content, including visually-rich documents in everyday life, such as pamphlets and banners, technical or scientific papers, business reports, storybooks, or novels with intricate and meaningful illustrations. Much effort has been dedicated to generating and integrating multiple design elements into visually-rich documents [26, 36, 74, 80]. Note that, when composing visually-rich documents (*e.g.* slides for a presentation, banners, cards, pamphlets, social media posts), it is not uncommon that designers, content creators, and even non-expert users would like to insert images that should integrate and blend nicely with the rest. Those users rely more and more often on AI-generated

Fig. 1: We propose a fully-automated pipeline to generate RGBA illustrations by adapting the inference-time behavior of a Diffusion Transformer model.

images for their designs. Generative models, especially prompt-driven diffusion models, have achieved impressive performance, and the increasing availability of user-friendly interfaces also eases non-experts' use of them. Moreover, these models allow for obtaining control over the generated content and producing images free from copyright. Nonetheless, when integrating an illustration inside a design, the designer often wants just the subject of interest to be superimposed on an existing background (possibly patterned or consisting of another image).

Large-scale image generation models have been widely adopted and deeply impacted visual arts. Still, layered image generation has not been extensively tackled, mainly because of the lack of training data. While RGB image generation can leverage open datasets of billions of samples [59], RGBA image datasets are of limited size [55, 65, 73], the biggest being the recently proposed MAGICK [8] which contains only 150k elements. These datasets are commonly used to train matting models to estimate the alpha channel of a given image [25, 65, 73, 75, 76]. Another solution to obtain RGBA images consists in directly predicting all four channels at once by fine-tuning a pre-trained generative model [77]. However, these solutions are costly for the required training resources.

To overcome these limitations, in this work, we propose *Alfie*, a pipeline to obtain high-quality, prompt-driven illustrations to be seamlessly integrated into any design or document with limited to no editing effort by the user (see Fig. 1). We exploit a pre-trained Diffusion Transformer model [12, 48], *i.e.* the freely-available PixArt-Σ [11], and modify its diffusion process in two ways. First, we mask and combine two latents (one for the subject and one for the background) in order to force the generation of the subject in the center of the canvas and with no sharp crops so that these can be realistic in a design or document. Second, we use the informative cross- and self-attention maps [13, 66] computed during the generation process to extract the foreground regions and estimate the alpha channel values. We opt for a Diffusion Transformer for the greater flexibility of such models in generating images with different sizes and aspect ratios compared to standard U-Net-based diffusion models. This arguably makes

them a more suitable starting point for the task at hand. Moreover, they are currently state-of-the-art both in generation quality and computational requirements [11, 12, 15, 48]. PixArt-Σ [11] (our baseline model) is composed of 0.6B parameters, compared to the 2.6B of Stable Diffusion XL [49]. We quantitatively and qualitatively assess the capability of our approach to generate subjects that are fully contained within the image and its capability to adhere to the prompt. We analyze the cross- and self-attention maps computed during the generation process, exploring various alternatives for estimating the α (transparency) channel. Furthermore, we present a user study, which shows that users mostly prefer our method over generating and then matting (63% of the time). Finally, we demonstrate the versatility of our generated illustrations by integrating them into a scene composition pipeline, achieving results of similar quality to those obtained by using illustrations from Adobe Stock. To facilitate further research on this task, we release the code of our approach and the evaluation setup[1].

2 Related Work

Diffusion Models. Since their introduction, diffusion models [16, 23, 47, 60, 62–64] have been widely adopted for their impressive performance, even more in the text-to-image setting, where the generation is guided by a prompt in natural language (as for, *e.g.*, DALL·E 2 [54], SD-XL [49], and the PixArt family [11, 12]). Such popularity has been further boosted by the introduction of latent diffusion models [56], which save computational resources in training and inference by working in the more compact latent space instead of the pixel space. Moreover, recently proposed Diffusion Transformers [11,12,15,45,48] have further increased the scalability and obtainable visual quality of diffusion models by exploiting a fully-attentive Transformer model instead of the convolutional-attentive U-Net-like noise estimators adopted in previous works. Lots of research efforts have also been dedicated to devising efficient fine-tuning strategies to obtain better controllability over the generation process of diffusion models based on additional guiding signals [2, 7, 18, 78] while maintaining the visual quality of the output. Another line of research focuses on achieving zero-shot capabilities by proposing strategies to alter the inference-time behavior of pre-trained diffusion models [1, 42–44, 51, 67]. Some of these approaches entail modifying the noisy latent vectors at each inference step. Some approaches perform mask-guided inpainting by combining the noised image with generated latent vectors [1, 42], or combine multiple independently generated latent vectors for region-based image generation [4]. Other works exploit the gradient of a downstream task-related loss or score [32, 35], possibly computed on the foreseen image obtained from the latent vector. A line of work has focused on manipulating the cross- and self-attention activations for image editing [14, 21], open-vocabulary segmentation [38, 46], and zero-shot video generation [33]. The role of attention layers in U-Net-based diffusion models has been extensively studied also from an explainability perspective by DAAM [66]. In this work, we use the cross-attention maps to identify

[1] `https://github.com/aimagelab/Alfie`

the foreground pixels of the generated image and combine them with the self-attention maps to obtain the alpha channel. In the context of generating designs and visually-rich documents, training, fine-tuning, and inference-time adaptation approaches for diffusion models have been proposed. These focus on generating aesthetically-pleasing document layouts [9, 10, 20, 28, 37], layout-aware designs backgrounds [70], user interfaces [27], SVGs [29, 72], and collages [4, 58, 79]. In this work, we propose to adapt the inference-time behavior of a Diffusion Transformer [11] for a fully automated pipeline to obtain illustration-like images that are easy to add to any existing background. In particular, we explore the inference-time adaptation of such a model by combining multiple masked latent vectors to obtain the desired characteristics in the generated images.

RGBA Images. Several approaches exist to estimate the α channel of a given RGB image [81], which can be binary and indicate foreground and background only (as in segmentation strategies), or continuous between 0 and 1 to indicate the opacity of each pixel (as in from matting algorithms). Proposed approaches range from classical ones exploiting perceptual features of the image (such as color [5] and edges [57]) to more recent learning-based solutions [25,34,50,65,73, 75]. Note that those approaches require some additional input to guide the prediction of the α channel. For example, image matting approaches [25, 65, 73, 76] need a map of foreground/background/unknown regions (a trimap), while even recent class-free segmentation approaches [34,41,75] need anchor points or visual prompts indicating the subject of interest. In LayerDiffuse [77], the authors propose a fine-tuning strategy to enable transparent image generation using large-scale pre-trained diffusion models. In particular, they adjust the latent space of Stable Diffusion XL [49] and finetune both the U-Net and the VAE, by using a loss that preserves the original latent distribution. This method, while effective, requires about 350 A100 hours for fine-tuning (as specified by the authors) and is not flexible to changes in the backbone model. In this work, we exploit the cross- and self-attention maps of our inference-time adapted Diffusion Transformer to obtain a guidance signal for generating RGBA images with satisfying results, especially when the output is used for further processing or scene composition.

3 Preliminaries

Diffusion Models. By diffusion models [23, 60, 61, 64] we refer to a class of generative probabilistic models trained to transform Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ into samples belonging to a certain data distribution $\mathbf{x}_0 \sim q$ in T steps. This is obtained by learning to approximate the data distribution q. To this end, during the so-defined forward process, Gaussian noise is injected into the data to transform the data distribution into the marginal distribution, *i.e.*

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}),$$

where the parameters (α_t, σ_t) define a differentiable noise schedule ensuring that $q(\mathbf{x}_t) \approx \mathcal{N}(0, \mathbf{I})$. Then, the diffusion models framework entails a reverse process in which the model is trained to denoise $\mathbf{x}_T \sim q(\mathbf{x}_t|\mathbf{x}_0)$. To this end, $\mathbf{x}_0$ is predicted

iteratively by estimating $\mathbf{x}_{t-1}$ starting from $\mathbf{x}_t$. The most common approach, proposed in [23], entails using a parameterization based on the prediction of the noise ϵ for sampling: both $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are parametrized as a combination of $\mathbf{x}_0$ and ϵ, scheduled according to the noise schedule. As a result, the model is trained to estimate the noise ϵ by optimizing

$$\mathbb{E}_{q(\mathbf{x}_0)}[\|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon\|_2^2].$$

Latent Diffusion Models. When working with high-resolution image data, training diffusion models in the pixel space is very computationally heavy. To face this issue, latent diffusion models [49, 56] have been introduced. These models work with vectors in the latent space of an autoencoder (*e.g.* a VQ-VAE [68] or a VQ-GAN [17]) and thus are more efficient than models working in the pixel-space in terms of GFlops [12, 48, 49, 56]. Here, we rely on the pre-trained, latent Diffusion Transformer model PixArt-Σ [11].

Conditional Generation. The generation process of diffusion models can be conditioned on a guidance signal. The most popular diffusion models to date are text-to-image models that can be conditioned on a natural language prompt, embedded by the text encoder of a pre-trained multimodal model (*e.g.* CLIP [52] or T5 [53]) in a vector $\mathbf{e}$. The conditioning is achieved by performing cross-attention with $\mathbf{e}$ inside the noise estimation network. Note that, in this work, we use as backbone a diffusion model trained with the Classifier-Free Guidance [24] conditioning strategy, in which the noise prediction $\hat{\epsilon}_\theta$ is obtained by combining the conditional prediction $\epsilon_\theta(\mathbf{x}_t, \mathbf{e}, t)$ and the unconditional prediction $\epsilon_\theta(\mathbf{x}_t, \emptyset, t)$, where $\emptyset$ is the embedding vector of the null prompt, with weight s as

$$\hat{\epsilon}_\theta(\mathbf{x}_t, \mathbf{e}, t) = \epsilon_\theta(\mathbf{x}_t, \emptyset, t) + s(\epsilon_\theta(\mathbf{x}_t, \mathbf{e}, t) - \epsilon_\theta(\mathbf{x}_t, \emptyset, t)).$$

Diffusion Transformers. Diffusion models originally featured a convolutional-attentive U-Net-like model as a noise estimator. The seminal work [48] proposes to replace the U-Net with a multi-block Diffusion Transformer model (DiT) to achieve better performance and scalability. The noise estimation Transformer takes as input a sequence of tokens obtained from squared patches of the latent vectors added to a positional embedding. At each block of the Transformer, the generation is conditioned on the timestep and the class. After the last block, the tokens are decoded and rearranged into the final image spatial dimension. Subsequent works [11, 12] adapt the class-guided Diffusion Transformer to work with text guidance. In this work, we exploit the recently-proposed PixArt-Σ Diffusion Transformer [11], which is directly conditioned on the textual prompt.

4 Inference-Time Illustration Generation

We propose a fully automated pipeline to obtain a high-quality RGBA illustration from a textual prompt. We argue that such illustration should have the following characteristics:

1. Contain the subjects specified in the prompt;
2. Be fully contained in the canvas without crops;
3. Have a precise α channel.

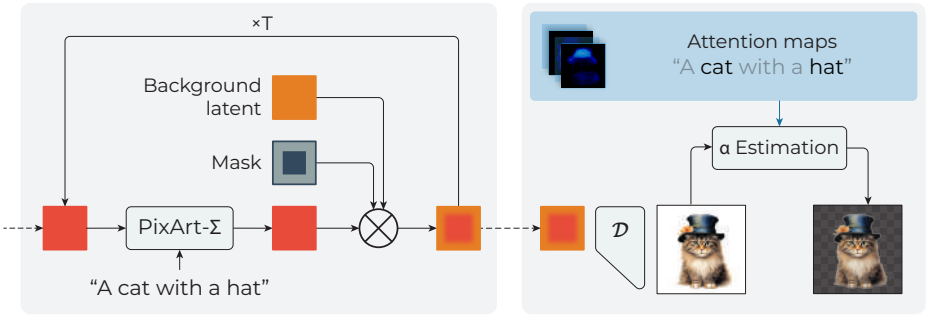

Fig. 2: Schematic representation of our fully-automated prompt-guided pipeline to obtain RGBA illustrations. The core element is a diffusion model, for which we devise an inference-time adaptation strategy aimed at making the generated images illustration-like. Then, we process the generation attention maps to estimate the α channel.

In our proposed inference-time method, we tackle all three aspects. In particular, we exploit the generative power of a pre-trained DiT (1), altering its inference-time behavior so that it always generates entire subjects (2) and extracting and cleaning an α channel estimate (3). Our approach is summarized in Fig. 2. We denote the RGB image as $\hat{x} \in \mathbb{R}^{h \times w \times 3}$, with values in range $[-1, 1]$. The transparency channel is denoted as $\alpha \in \mathbb{R}^{h \times w \times 1}$, with values in range $[0, 1]$. The complete RGBA image is obtained by concatenating the RGB image with its α channel and has shape $h \times w \times 4$.

In particular, we propose to obtain the α channel by exploiting the relevant foreground subject's cross-attention maps and the pixels' self-attention map relative to the foreground subject. The rationale is that the cross-attention map will have higher values for the pixels of the subject than those of the background, providing a coarse localization of the subject. On the other hand, the self-attention values will help render the texture and appearance of the foreground subject. In fact, when a foreground pixel is relative to an object of solid material, it will attend mainly to neighboring pixels of the same object, resulting in a cohesive appearance. In this case, all the self-attention weights for the pixels of that object will be high. Conversely, the appearance of foreground pixels of objects in a translucent, see-through material (*e.g.*, water, fire, glass) will also depend on the background. Therefore, that pixel will attend to both the pixels of the translucent object and the background. As a result, the self-attention weights for the pixels of that object will be lower than in the case of solid objects. This behavior is in line with what the α channel represents. For this reason, we propose to use the cross-attention and self-attention maps to define the α channel of our generated RGBA illustrations.

Centering the Subject. Diffusion models are trained on billions of images [59] and learn to generate all kinds of styles, from natural to artworks. While some works analyze the disentanglement capability of such models [71], recent works argue that the low quality of the captions in most large-scale datasets negatively affects the generation capabilities and train on LLaVA [39]-refined captions [11, 12]. Still, complete disentangling has yet to be achieved and the models have

difficulties with generating meta-descriptions or combinations of subjects and styles not seen in the training data. Note that by meta-descriptions, we define indications on the characteristics of the image, such as the background color or the fact that the full subject should be contained.

When generating illustrations, it is important that the subjects are wholly contained within the image canvas and are not cut or cropped. To achieve this, we chose to employ a method inspired by [1, 4], where we jointly generate the illustration (described by the textual prompt) and a uniform background, and we blend them together at each denoising step. Formally, to generate an image $\mathbf{x_0}$ with the text prompt $\mathbf{e}$, we first define a squared mask $\mathbf{m}$, covering the inner area of the canvas. Then, we start from the Gaussian noise $\mathbf{x}_t$ at the timestep $t \in [0, T]$ and duplicate it into a foreground latent $\mathbf{x}_{t,fg}$ and a background latent $\mathbf{x}_{t,bg}$, assigning a background textual prompt $\mathbf{e}_{bg}$ to the respective latent. After the denoising step, we merge the obtained predictions into

$$\mathbf{x}_{t-1} = \mathbf{x}_{t-1,fg} \cdot \mathbf{m} + \mathbf{x}_{t-1,bg} \cdot (1 - \mathbf{m}).$$

This process constrains the illustration subjects to be generated inside the mask while ensuring that they are not cropped. After the denoising chain, the image is obtained from the final latent $\mathbf{x}_0$ by using the decoder $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{x}_0)$.

Isolating the Subject. During each forward process generation step t, each layer l performs cross-attention between the latent $\mathbf{z}_{t,l}$ and the encoded prompt $\mathbf{e}$ to guide the subject creation and self-attention to correlate image features. Following DAAM [66] and DAAM-I2I [13], we extract the cross- and self-attention maps. Formally, for a given t and l, the cross-attention map $\mathcal{A}_C^{t,l} \in [0,1]^{h \times w \times N}$ will represent the relation between each pixel and each of the N prompt tokens, while the self-attention map $\mathcal{A}_S^{t,l} \in [0,1]^{hw \times h \times w}$ will represent the relation between each pixel to all the other latent pixels:

$$\mathcal{A}_C^{t,l} := \text{softmax}\left(Q_{\mathbf{z}_{t,l}} K_{\mathbf{e}}^T / \sqrt{d}\right),$$
$$\mathcal{A}_S^{t,l} := \text{softmax}\left(Q_{\mathbf{z}_{t,l}} K_{\mathbf{z}_{t,l}}^T / \sqrt{d}\right),$$

where d is the feature size of the layer l.

Note that by performing subject-centering and classifier-free guidance, we have a batch size of four, respectively representing the latent for the null (θ) and text prompts ($\mathbf{e}$) for the background and the subject. We discard all the attention maps except those corresponding to the prompt-guided subject latent. We reserve further investigation on integrating the background and null-prompt guided attention maps for future work. Since diffusion models define the coarse image layout during the first steps of the denoising chain and later define the details, we keep the maps of the last ten out of thirty timesteps to obtain more precise localization. Finally, we average across timesteps, layers, and attention heads to obtain the global maps $\mathcal{A}_C \in [0,1]^{h \times w \times N}$ and $\mathcal{A}_S \in [0,1]^{hw \times h \times w}$.

Once the final RGB image $\mathbf{x}_0$ is obtained, we isolate the cross-attention maps of the noise estimator relative to the nouns in the input prompt $\mathbf{e}$. The list of nouns $\mathbf{n} = [\mathbf{n}_1, ..., \mathbf{n}_N]$ is obtained by analyzing the prompt with the standard tokenizer from the NLTK library [6] and excluding generic, uninformative nouns

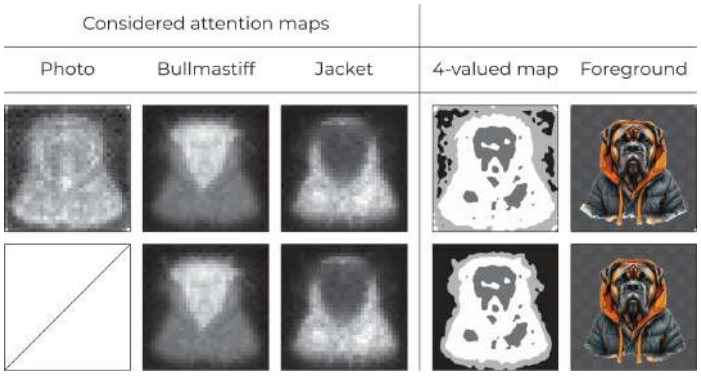

Fig. 3: Cross-attention map analysis of the prompt *A photo of a bullmastiff with a jacket*. Left to right: maps of the three prompt nouns, candidate region mask computed w/ and w/o the generic noun *photo*, and foreground extraction results.

such as *image*, *photo*, and *picture*. We release the complete list of excluded nouns in the provided repository. Empirically, we found that the cross-attention maps associated with these nouns refer to generic areas of the image and not always to the subject, hindering the final quality. This is in line with what is done in [30, 66], where the authors performed analyses on the cross-attention maps of text-to-image convolutional-attentive diffusion models and avoided spurious correlations between text tokens and feature space by removing this type of nouns. To showcase the effect of the generic nouns exclusion, in Fig. 3 we visualize the attention maps of different nouns and the effect on the obtained illustration of excluding generic nouns for the prompt *A photo of a bullmastiff with a jacket*. As we can see, the attention maps related to meaningful object nouns (*e.g.*, *dog*, *jacket*) have high activation values in the spatial locations aligned with the specified subjects. In contrast, the attention map of the generic noun *photo* has high activation values also in border regions. In the second-last column, we display the mask of the candidate foreground regions that we use to obtain the guidance masks for a foreground extraction algorithm, namely GrabCut [57]. In particular, we average the maps of each subject and normalize the values between 0 and 1, obtaining the foreground cross-attention map $\overline{\mathcal{CA}}_{\text{fg}}$ and quantizing it into a mask. For foreground extraction, this mask has four values: background (black), probable background (light gray), probable foreground (white), and sure foreground (dark gray). As we can see, this mask is much cleaner when not considering generic, non-descriptive nouns, and the final illustration is of higher quality. In summary, as also shown in [21, 66] for U-Net-based diffusion models, cross-attention maps associated with prompt subjects contain high activation values in the foreground regions. We argue that the attention maps of a Transformer-based model exhibit the same behavior. Therefore, we use them to compute the candidate foreground pixels.

Then, we extract the subject information from the self-attention maps. Inspired by DAAM-I2I [13], we use the foreground map $\overline{\mathcal{CA}}_{\text{fg}}$ to merge all the foreground pixels self-attention maps by performing a weighted average, where

the weights are the activation values of the foreground maps, normalized in the range $[0, 1]$. This map, denoted as $\overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{\text{fg}})}$, provides us with the attention scores of each foreground pixel with respect to the other foreground pixels. As explained before, we want to combine the different aspects captured by cross- and self-attention and use both for our coarse α channel estimation, which is given by the normalized sum of the two maps, *i.e.*,

$$\hat{\alpha} = \overline{\overline{\mathcal{CA}}_{\text{fg}} + \overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{\text{fg}})}}.$$

Obtaining the RGBA Illustration. Once the reverse diffusion process is over, we obtain an image containing the desired subject with no sharp croppings and an estimation of the coarse transparency mask $\hat{\alpha}$. At this point, we need to remove the background to obtain the final illustration. To this end, we exploit the candidate foreground map $\overline{\mathcal{CA}}_{\text{fg}}$ obtained with the cross-attention maps of the noise estimation Transformer. Specifically, we quantize it into four values to obtain a map whose values indicate whether the corresponding pixel is in the sure background, the probable background, the probable foreground, or the sure foreground. We use the so-obtained mask as input to the perceptual GrabCut [57] algorithm, which performs graph optimization by combining color and border information to remove the background from the generated image. We use this algorithm's output foreground mask to clean $\hat{\alpha}$, zeroing the values outside the subject and obtaining our transparency channel α.

As a side note, we remark that one could alter the intensity of the α channel, *i.e.* augmenting or reducing the opacity of the generated illustration in a semantically meaningful way, by simply introducing a hyperparameter k to obtain

$$\hat{\alpha}' = \min(1, (1 + k)\hat{\alpha}),$$

which the user can tune depending on their needs. We set it to 0.5 in all experiments, leaving further exploration for future work.

5 Experiments and Results

Implementation Details. In our experiments, we use the pre-trained PixArt-Σ-512 [11] from HuggingFace [69]. PixArt-Σ is based on DiT [48] and comprises 28 self- and cross-attention layers, each with 16 attention heads. We use the Euler Discrete scheduler [31] with 30 denoising steps. The subject-centering mask has a border of 64 in the pixel space on all image sides. For the attention map quantization into a 4-valued map, we use the percentiles 0.8, 0.3, and 0.1, respectively: for sure foreground, probable foreground, and probable background.

Evaluation Setup. For quantitative evaluation, we collect 3000 of the multi-sentence prompts used to train PixArt-Σ [11, 12] and consider only the first sentence of the prompt, which is the one describing the general visual content of the image. These have been obtained by running the LLaVA Multimodal Large Language Model [40] on images from the SAM dataset [34][2].

[2] We release the list of prompts at `https://github.com/aimagelab/Alfie`.

We quantitatively evaluate how reliably the pipeline generates illustrations of subjects wholly contained within the image, without crops or cuts. To do so, we consider the image canvas's left, right, top, and bottom borders and, for each one, a margin of 4 pixels. If all the pixels in the margin area have a value greater than 0.8, we consider that border empty. We choose this threshold value because we aim to keep the background uniform and light rather than perfectly white. Then, we compute the percentage of generated images having each of the borders empty (denoted as **empty-**{**l**, **r**, **t**, **b**}, respectively), as well as the aggregate percentage of images having all four borders empty (denoted as **empty-a**). Moreover, we consider the CLIP score [22] (denoted as **CLIP-S**) to quantify how much the generated RGB images still respect the original prompt. Regarding the RGBA images, we remark that we do not evaluate in terms of the FID and KID scores, which are commonly adopted in image generation evaluation. This is because these scores need a large reference set (ground truth RGBA images) possibly compatible with the prompts fed to our model and with the characteristics that we enforce in our images. To the best of our knowledge, such a set with the required characteristics is not available.

Moreover, in line with RGBA image generation works [77], we perform a user study to compare the perceived quality of our illustrations compared to the best-performing alternative method. For a fair comparison, we employ the state-of-the-art ViT-B-based ViTMatte [75] and perform matting over our generation results. This matting method requires an input trimap for sure foreground, unsure regions, and sure background. We provide it by quantizing the same map $\overline{\mathcal{CA}}_{\text{fg}}$ that we obtain in our pipeline with the percentiles 0.8 and 0.3 for sure foreground and unsure regions, respectively. We consider the values under 0.3 as sure background. We also provide qualitative results showcasing the effects of centering, α estimation, and comparison with matting.

Considered Variants and Baselines. As baselines, we consider the base PixArt-Σ [11] with the same prompts as we use for our pipeline and a version to which we always append the phrase *on a white background* to the input prompt (we refer to this as **PixArt-Σ+suffix**). As for the variants of our approach, we isolate the role of each component by considering the subject centering (which we refer to as **PixArt-Σ+centering**) and different methods for obtaining the α channel, namely: the normalized cross-attention heatmap $\overline{\mathcal{CA}}_{\text{fg}}$, the normalized self-attention heatmap $\overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{\text{fg}})}$, the normalized combination of cross- and self-attention $\hat{\alpha}$, and the GrabCut-cleaned α (which we refer to as *Alfie*).

Runtime. We run all experiments on a 24GB Nvidia RTX 4090 GPU with half precision. Generating an image with the base model (PixArt-Σ [11]) takes $\sim$15GB of VRAM and $\sim$3.15 seconds, increased to $\sim$4.13 seconds and $\sim$16.7GB of VRAM when we perform the centering. Cleaning the estimate transparency channel $\hat{\alpha}$ with GrabCut [57] requires additional $\sim$0.36 seconds.

5.1 Results

Centering. We consider the background generation statistics and report it in Table 1. We can observe that the PixArt-Σ+suffix variant is not entirely reli-

Fig. 4: Qualitative comparison on the effect of our constrained whole-subject generation compared to meta-descriptions in the input prompt.

able, as it leads to ∼53% probability of generating the image inside the canvas. Conversely, our approach provides a much higher probability of generating the whole subject, consistently reaching more than 95%. In Table 1, we also report the CLIP-S between the generated images and the textual prompt. As we can see, the base model obtains the highest value, our reference score, as the generated images are fully based on the textual prompt, and the generation process is unconstrained. On the other hand, our target images have implicit characteristics that are not part of the textual caption but rather represent meta-descriptions, *e.g.*, whole-object generation. As expected, the second-best CLIP-S is obtained by the base model with the suffix. In fact, this baseline sometimes (almost half of the times) ignores the suffix, and thus, it is less penalized. While respecting the desired characteristics defined in Section 4, our generated images achieve CLIP-S values very close to the reference ones. Thus, we can conclude that our inference-time adaptation strategy guides the generation toward the desired characteristics while maintaining adherence to the textual prompt. In Fig. 4, we showcase the impact of this constraint on the generated images. While the simple appending of the suffix cannot always guarantee whole-subject generation, our method consistently steers the generation towards a correct result and enforces this meta-description on the image.

Table 1: Quantitative comparison of our approach variants in terms of average probability with which the subjects are generated in the center of the image with no sharp croppings and adherence to the prompt.

	empty-a	**empty-l**	**empty-r**	**empty-t**	**empty-b**	**CLIP-S**
PixArt-Σ	3.33	4.33	4.13	8.53	6.06	31.29
PixArt-Σ + suffix	53.10	62.63	62.10	92.53	84.50	30.79
PixArt-Σ + centering	96.50	99.27	99.03	99.53	98.00	30.08

Fig. 5: Qualitative comparison on different α estimates. Combining self- and cross-attention maps provides the best balance of spatial localization and transparency values, and their cleanup using GrabCut [57] (*Alfie*) further increases border precision.

α Channel Estimation. In Fig. 5, we show alternative α estimations obtained using different combinations and processing of the attention maps. As we can see, the normalized cross-attention map $\overline{\mathcal{CA}}_{\text{fg}}$ provides coarse localization of the foreground subjects with uniform transparency values. The normalized self-attention map $\overline{\mathcal{FF}}_{(\overline{\mathcal{CA}}_{\text{fg}})}$ is more spatially precise and has a broader range of values, reflecting how much the objects appearance depends on a localized or sparse image region. By combining the two, the resulting map $\hat{\alpha}$ gives a balance between absolute and relative transparency values. Finally, using Grabcut [57] for cleanup ensures precise border localization, leading to the best overall results.

Comparison with Matting. We compare our method against matting by considering the generated RGB images with the foreground cross-attention map $\overline{\mathcal{CA}}_{\text{fg}}$ and using it to compute the input trimap for ViTMatte [75]. We run a user study with ten individuals and 100 images randomly sampled from our dataset and ask them "Which of the following RGBA illustrations do you prefer?". We find that

Fig. 6: Qualitative comparison of our method against ViTMatte [75]. Without training an additional network, we produce reasonable transparency channels.

users indicated the images from our method ∼63% of the time, demonstrating that we are slightly better than state-of-the-art matting methods while requiring no costs for training. In Figure 6, we provide samples obtained by using the two methods. As we can see, our method provides reasonable α channels, comparable to or better than matting, mainly because using the attention maps reduces prediction errors, discarded image parts, and preserved background areas.

Compositional Artwork Generation. We use the images generated by our method as input for Collage Diffusion [58]. This method allows compositional scene generation by taking sequences of layers as input, which define the spatial arrangement and attributes of objects in the scene. The authors combine SDEdit [43], cross-attention manipulation [3], textual inversion [19], and a proposed extension of ControlNet [78] to find a balance between harmonization (of the input subjects and the scene) and fidelity (of the produced collage w.r.t. the different original subjects). The pipeline runs on Stable Diffusion v2-1-base [56]. In Fig. 7, we show the results on the pipeline with the same input background and three different subjects: an RGBA illustration obtained from Adobe Stock, an image obtained with the base PixArt-Σ model, and an image obtained with Alfie. The outputs of Collage Diffusion showcase that the quality obtained by using our method is similar to that achieved with a commercial RGBA illustration. Conversely, when the input is the base PixArt-Σ squared illustration, the pipeline tries to incorporate it into the scene by rendering it as a big painting, a plausible but unacceptable result.

6 Conclusions

In this work, we have tackled the task of generating illustrations for visual content, artworks, and visually-rich documents. We have devised a fully automated, prompt-guided pipeline for obtaining high-quality RGBA images that can be used by users or integrated into scene compositions with good results at zero additional training cost. As the main component of our pipeline, we have exploited the recently proposed Diffusion Transformer paradigm and explored inference-time adaptation strategies for these models. We have quantitatively and qualitatively validated the effectiveness of the proposed approach, which can serve as

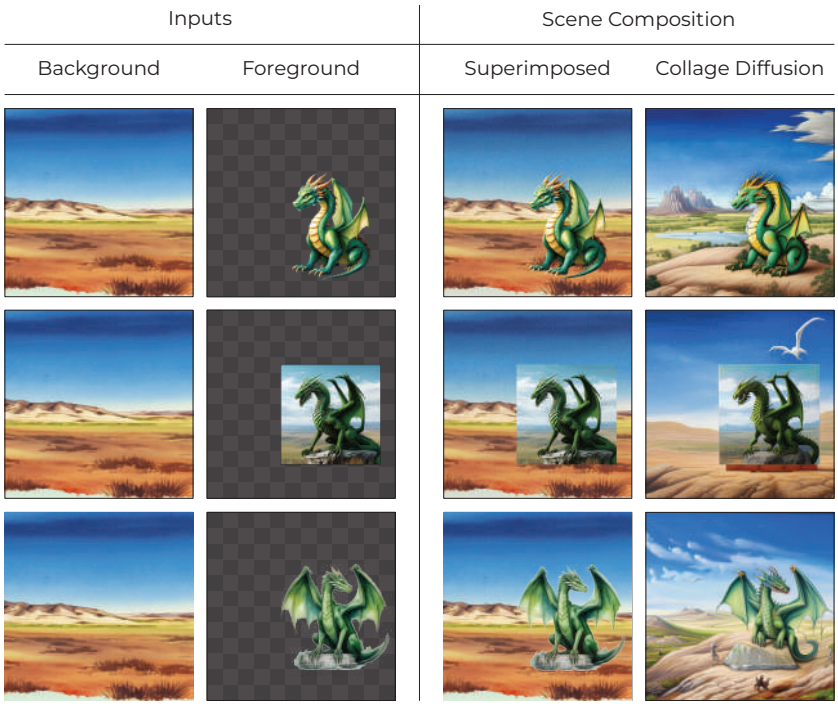

Fig. 7: Results of Collage Diffusion [58] with different input foreground images. Top to bottom: a commercial RGBA image, an image generated by the baseline PixArt-Σ [11], and an RGBA illustration obtained with our approach. With our generated illustration, the output is comparable to that obtained with the commercial image, differently from the pipeline run with the image from the baseline model. To generate the background and the foreground, we use the prompts *A steppe landscape* and *A green dragon*, combined for the composite scene into *A green dragon in a steppe landscape*.

a starting point for the community to continue working on the tackled task. We hope that the promising results obtained can encourage the research towards the development of similar low-cost strategies for such models.

Acknowledgement

This work was supported by the "AI for Digital Humanities" project funded by "Fondazione di Modena" and the PNRR project Italian Strengthening of ESFRI RI Resilience (ITSERR) funded by the European Union – NextGenerationEU.

References

1. Avrahami, O., Fried, O., Lischinski, D.: Blended Latent Diffusion. ACM Trans. Graphics **42**(4), 1–11 (2023)
2. Avrahami, O., Hayes, T., Gafni, O., Gupta, S., Taigman, Y., Parikh, D., Lischinski, D., Fried, O., Yin, X.: SpaText: Spatio-Textual Representation for Controllable ImageGgeneration. In: CVPR (2023)
3. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. arXiv preprint arXiv:2211.01324 (2022)
4. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. ICML (2023)
5. Ben-Ezra, M.: Segmentation with Invisible Keying Signal. In: CVPR (2000)
6. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
7. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to Follow Image Editing Instructions. In: CVPR (2023)
8. 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: CVPR (2024)
9. Chai, S., Zhuang, L., Yan, F.: LayoutDM: Transformer-based Diffusion Model for Layout Generation. In: CVPR (2023)
10. Chen, J., Zhang, R., Zhou, Y., Chen, C.: Towards Aligned Layout Generation via Diffusion Model with Aesthetic Constraints. In: ICLR (2023)
11. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: PixArt-\Sigma: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. arXiv preprint arXiv:2403.04692 (2024)
12. Chen, J., YU, J., GE, C., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. In: ICLR (2024)
13. Chowdhury, R.D.: DAAM-Image2Image: Extension of DAAM for Image Self-Attention in Diffusion Models. `https://github.com/RishiDarkDevil/daam-i2i`
14. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: DiffEdit: Diffusion-based semantic image editing with mask guidance. In: ICLR (2022)
15. Crowson, K., Baumann, S.A., Birch, A., Abraham, T.M., Kaplan, D.Z., Shippole, E.: Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In: ICML (2024)
16. Dhariwal, P., Nichol, A.: Diffusion Models Beat GANs on Image Synthesis. NeurIPS (2021)
17. Esser, P., Rombach, R., Ommer, B.: Taming Transformers for High-Resolution Image Synthesis. In: CVPR (2021)
18. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., Taigman, Y.: Make-a-Scene: Scene-Based Text-to-Image Generation with Human Priors. In: ECCV (2022)
19. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
20. He, L., Lu, Y., Corring, J., Florencio, D., Zhang, C.: Diffusion-Based Document Layout Generation. In: ICDAR (2023)

21. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-Prompt Image Editing with Cross-Attention Control. In: ICLR (2023)
22. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP (2021)
23. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. NeurIPS (2020)
24. Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance. In: NeurIPS Workshop (2021)
25. Hu, Y., Lin, Y., Wang, W., Zhao, Y., Wei, Y., Shi, H.: Diffusion for Natural Image Matting. arXiv preprint arXiv:2312.05915 (2023)
26. Huang, D., Guo, J., Sun, S., Tian, H., Lin, J., Hu, Z., Lin, C.Y., Lou, J.G., Zhang, D.: A Survey for Graphic Design Intelligence. arXiv preprint arXiv:2309.01371 (2023)
27. Hui, M., Zhang, Z., Zhang, X., Xie, W., Wang, Y., Lu, Y.: Unifying Layout Generation with a Decoupled Diffusion Model. In: CVPR (2023)
28. Inoue, N., Kikuchi, K., Simo-Serra, E., Otani, M., Yamaguchi, K.: LayoutDM: Discrete Diffusion Model for Controllable Layout Generation. In: CVPR (2023)
29. Jain, A., Xie, A., Abbeel, P.: VectorFusion: Text-to-SVG by Abstracting Pixel-Based Diffusion Models. In: CVPR (2023)
30. Karazija, L., Laina, I., Vedaldi, A., Rupprecht, C.: Diffusion models for zero-shot open-vocabulary segmentation. arXiv preprint arXiv:2306.09316 (2023)
31. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the Design Space of Diffusion-Based Generative Models. NeurIPS (2022)
32. Katherine Crowson: CLIP guided diffusion HQ 256x256. `https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj`
33. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In: ICCV (2023)
34. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
35. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via Synchronized Joint Diffusions. In: NeurIPS (2023)
36. Li, D.W., Huang, D., Ma, T., Lin, C.Y.: Towards topic-aware slide generation for academic papers with unsupervised mutual learning. In: AAAI (2021)
37. Li, F., Liu, A., Feng, W., Zhu, H., Li, Y., Zhang, Z., Lv, J., Zhu, X., Shen, J., Lin, Z., et al.: Relation-aware diffusion model for controllable poster layout generation. In: CIKM (2023)
38. Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Open-vocabulary object segmentation with diffusion models. In: ICCV (2023)
39. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: NeurIPS (2023)
40. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. NeurIPS (2024)
41. Lüddecke, T., Ecker, A.: Image Segmentation Using Text and Image Prompts. In: CVPR (2022)
42. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In: CVPR (2022)
43. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In: ICLR (2022)
44. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text Inversion for Editing Real Images using Guided Diffusion Models. In: CVPR (2023)

45. Nair, N.G., Valanarasu, J.M.J., Patel, V.M.: Diffscaler: Enhancing the Generative Prowess of Diffusion Transformers. arXiv preprint arXiv:2404.09976 (2024)
46. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. NeurIPS (2024)
47. Nichol, A.Q., Dhariwal, P.: Improved Denoising Diffusion Probabilistic Models. In: ICML (2021)
48. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: ICCV (2023)
49. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In: ICLR (2024)
50. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Van Gool, L.: Highly Accurate Dichotomous Image Segmentation. In: ECCV (2022)
51. Quattrini, F., Pippi, V., Cascianelli, S., Cucchiara, R.: Merging and Splitting Diffusion Paths for Semantically Coherent Panoramas. In: ECCV (2024)
52. 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: ICML (2021)
53. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the Limits of Transfer Learning with a Unified Text-To-Text Transformer. Journal of Machine Learning Research **21**(1), 5485–5551 (2020)
54. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv e-prints pp. arXiv–2204 (2022)
55. Rhemann, C., Rother, C., Wang, J., Gelautz, M., Kohli, P., Rott, P.: A perceptually motivated online benchmark for image matting. In: CVPR (2009)
56. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution Image Synthesis with Latent Diffusion Models. In: CVPR (2022)
57. Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graphics **23**(3), 309–314 (2004)
58. Sarukkai, V., Li, L., Ma, A., Ré, C., Fatahalian, K.: Collage Diffusion. In: WACV (2024)
59. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. NeurIPS (2022)
60. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: ICML (2015)
61. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: ICLR (2021)
62. Song, Y., Ermon, S.: Generative Modeling by Estimating Gradients of the Data Distribution. NeurIPS (2019)
63. Song, Y., Ermon, S.: Improved Techniques for Training Score-Based Generative Models. NeurIPS (2020)
64. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations. In: ICLR (2021)
65. Sun, Y., Tang, C.K., Tai, Y.W.: Semantic image matting. In: CVPR (2021)
66. Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., Ture, F.: What the daam: Interpreting stable diffusion using cross attention. In: Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 5644–5659 (2023)

67. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In: CVPR (2023)
68. Van Den Oord, A., Vinyals, O., et al.: Neural Discrete Representation Learning. NeurIPS (2017)
69. Von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Wolf, T.: Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers` (2022)
70. Weng, H., Huang, D., Qiao, Y., Hu, Z., Lin, C.Y., Zhang, T., Chen, C.: Desigen: A pipeline for controllable design template generation. CVPR (2024)
71. Wu, Q., Liu, Y., Zhao, H., Kale, A., Bui, T., Yu, T., Lin, Z., Zhang, Y., Chang, S.: Uncovering the disentanglement capability in text-to-image diffusion models. In: CVPR (2023)
72. Xing, X., Zhou, H., Wang, C., Zhang, J., Xu, D., Yu, Q.: Svgdreamer: Text guided svg generation with diffusion model. CVPR (20234)
73. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In: CVPR (2017)
74. Yamaguchi, K.: CanvasVAE: Learning to Generate Vector Graphic Documents. In: ICCV. IEEE (2021)
75. Yao, J., Wang, X., Yang, S., Wang, B.: ViTMatte: Boosting image matting with pre-trained plain vision transformers. Inform. Fusion **103**, 102091 (2024)
76. Yao, J., Wang, X., Ye, L., Liu, W.: Matte anything: Interactive natural image matting with segment anything model. Image and Vision Computing **147**, 105067 (2024)
77. Zhang, L., Agrawala, M.: Transparent image layer diffusion using latent transparency. arXiv preprint arXiv:2402.17113 (2024)
78. Zhang, L., Rao, A., Agrawala, M.: Adding Conditional Control to Text-to-Image Diffusion Models. In: ICCV (2023)
79. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: DiffCollage: Parallel Generation of Large Content With Diffusion Models. In: CVPR (2023)
80. Zheng, X., Xiaotian, Q., Ying, C., Rynson, W.: Content-aware Generative Modeling of Graphic Design Layouts. ACM Trans. Graphics **38**(4), 133 (2019)
81. Zhu, Q., Shao, L., Li, X., Wang, L.: Targeting Accurate Object Extraction From an Image: A Comprehensive Study of Natural Image Matting. IEEE Trans. Neural Netw. Learn. Syst. p. 185 (2015)