MagicFusion: Boosting Text-to-Image Generation Performance by Fusing Diffusion Models

Jing Zhao[1*], Heliang Zheng[2], Chaoyue Wang[2], Long Lan[1], Wenjing Yang[1†]
[1]National University of Defense Technology, Changsha, China,
[2]JD Explore Academy, Beijing, China
{zhaojing,long.lan,wenjing.yang}@nudt.edu.cn, zhenghllj@gmail.com, chaoyue.wang@outlook.com

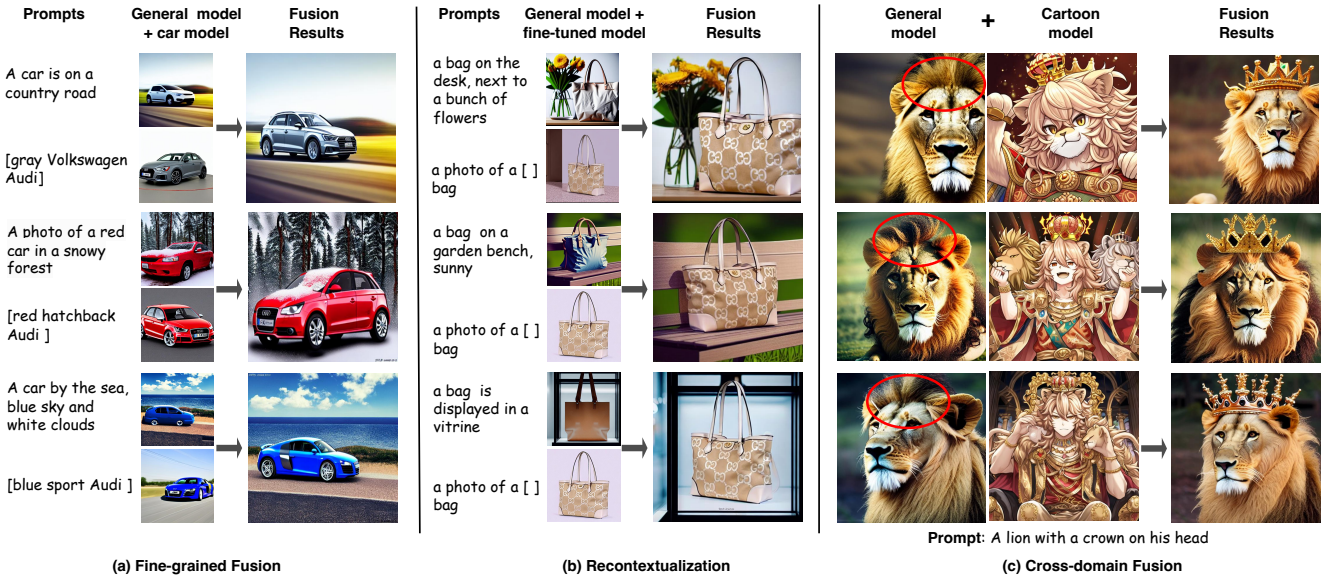

Figure 1: "MagicFusion". Given two diffusion models, our method can preserve the strengths of each individual model. (a) A general model [27] + a fine-grained car model to achieve fine-grained generation with complex scenes. (b) A general model + a DreamBooth model [28] to recontextualize specific objects with well-preserved details. (c) A general model + a cartoon model to generate creative scenes with photorealistic fidelity.

Abstract

The advent of open-source AI communities has produced a cornucopia of powerful text-guided diffusion models that are trained on various datasets. While few explorations have been conducted on ensembling such models to combine their strengths. In this work, we propose a simple yet effective method called Saliency-aware Noise Blending (SNB) that can empower the fused text-guided diffusion models to achieve more controllable generation. Specifically, we experimentally find that the responses of classifier-free guidance are highly related to the saliency of generated images. Thus we propose to trust different models in their areas of expertise by blending the predicted noises of two diffusion models in a saliency-aware manner. SNB is training-free and can be completed within a DDIM sampling process. Additionally, it can automatically align the semantics of two noise spaces without requiring additional annotations such as masks. Extensive experiments show the impressive effectiveness of SNB in various applications. The project page is available at https://magicfusion.github.io/.

1. Introduction

In recent years, significant progress has been made in image generation [29, 6, 24, 23, 27, 20, 29, 5] thanks to breakthroughs in diffusion models [32, 29, 33] and large-

*Work done during an internship at JD Explore Academy.
†Corresponding authors.

scale training [24, 20, 29, 5], as well as the contributions of open-source AI communities. Pre-trained large models have become an invaluable resource in this field. One of the most exciting developments has been text-guided diffusion models [24, 20, 29, 27]. A wide range of powerful text-guided diffusion models trained on various datasets has been publicly released. For example, general models (e.g., stable diffusion v1-4, v1-5, v2-1, etc. [27]) trained on large-scale multimodal datasets like LAION 5B [31], as well as more specialized models (e.g., Anything-v3) trained on cartoon and anime datasets or fine-grained categories such as cars [38]. There are even fine-tuned models designed for specific objects [28]. The vast amounts of data and computational cost have enabled these models to achieve impressive capabilities in various fields. However, few explorations have been conducted on ensembling such models to combine their strengths.

Some works propose to add special symbols or signature phrases when fine-tuning models on new datasets [1]. This approach enables the model to generate novel image distributions while retaining its ability to generate from the original data distribution. However, there has been limited discussion on how to effectively combine the generation capabilities of these two distributions. One intuitive method for integrating the capabilities of two models involves taking a weighted average of their predicted noises [2]. However, such kind of fusions often fails to fully preserve the strengths of each model. Blended diffusion [3] proposes to spatially blend a noisy image and a predicted one, which has been explored in image editing tasks. However, this typically requires specifying a mask to edit particular objects, and few discussions have been conducted to blend the noises of two diffusion models.

In this work, we propose a simple yet effective method called Saliency-aware Noise Blending (SNB) that can empower the fused text-guided diffusion models to achieve more controllable generation. Specifically, we integrate two diffusion models by spatially blending the predicted noises. Our insight is to trust different models in their areas of expertise, thus the strengths of each individual model can be preserved. To obtain diffusion models' areas of expertise, we revisit the classifier-free guidance [13], which is widely adopted in text-guided diffusion models to enhance the difference between a given text and a null text in the predicted noise space. We experimentally find that the responses of classifier-free guidance are highly related to the saliency of generated images. To this end, we propose Saliency-aware Noise Blending that blends the predicted noises of two diffusion models based on their responses to classifier-free guidance.

SNB is training-free and can be completed within a DDIM sampling [33] process. Additionally, it can automatically align the semantics of two noise spaces without requiring additional annotations such as masks. Our main contributions can be summarised as follows:

- We propose to fuse two well-trained diffusion models to achieve more powerful image generation, which is a novel and valuable topic.
- We propose a simple yet effective Saliency-aware Noise Blending method for text-guided diffusion models fusion, which can preserve the strengths of each individual model.
- We conduct extensive experiments on three challenging applications (*i.e.*, a general model + a cartoon model, a fine-grained car model, and a DreamBooth [28] model), and prove that SNB can significantly empower pre-trained diffusion models.

The remainder of the paper is organized as follows. We describe related work in Section 2 and introduce our proposed SNB method in Section 3. An evaluation of three applications is presented in Section 4, followed by the results and comparisons with other methods in Section 5. Finally, a comprehensive summary of the paper is presented and an analysis of the limitations is provided in Section 6.

2. Related Works

2.1. Text-to-image synthesis

Text-guided image generation plays a significant role in image generation [43, 16, 7, 12, 36, 15, 17, 21, 22, 24, 42].Previous works mainly focus on GAN-based [10] models and small-scale image-text datasets [45, 34, 37, 41, 40]. Since Transformer-based autoregressive models [8, 24] were proposed, more and more attention has been attracted to large-scale training. The emergence of denoising diffusion models is another milestone, which significantly boosts the generation fidelity [23, 30, 20, 27]. Notably, Stable Diffusion [27] is publicly released, enabling a large number of variants that are fine-tuned on different datasets. The vast amounts of data and computational cost have enabled these models to achieve impressive capabilities in various fields. However, few explorations have been conducted on ensembling such models to combine their strengths.

2.2. Model Ensembling

Model ensembling is a powerful technique to distill the knowledge of multiple models and boost the performance, which is widely obtained in image understanding tasks, i.e., classification problems [44, 26, 11, 39], regression problems [19, 25] and clustering [35]. White such methods are hard to be adapted to generative models due to the large and complex image pixel space. Vision-aided GAN [14] proposes ensembling pre-trained vision models as a loss to guide the optimization of a generator. eDiff-I [4] proposes to ensemble different denoisers in different timesteps to improve the overall performance of image generation. In this

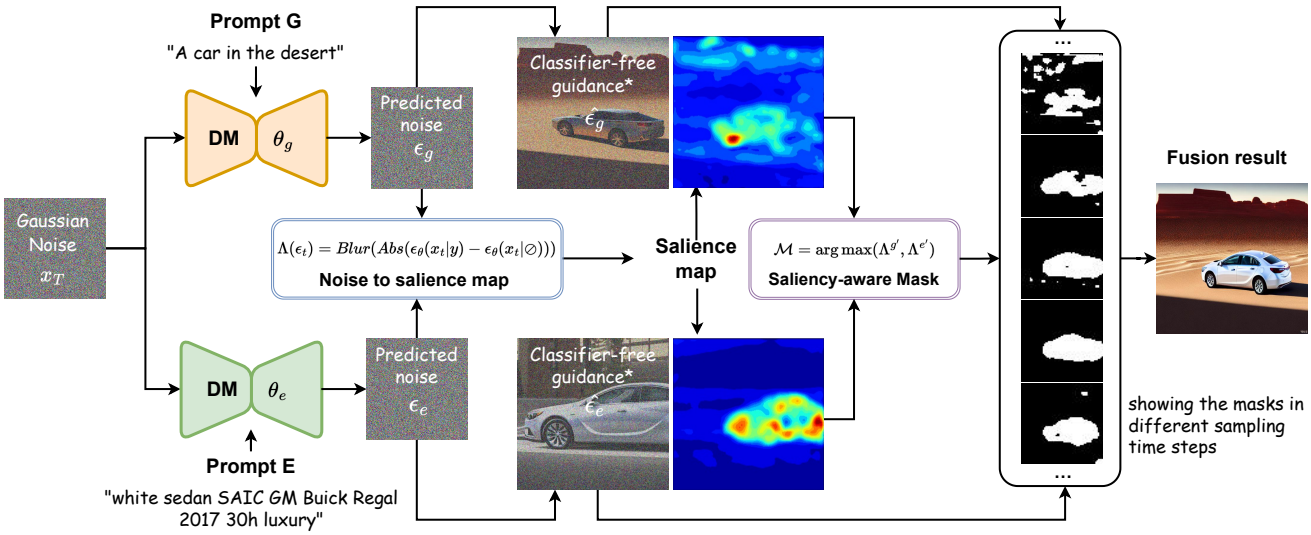

Figure 2: An overview of our Saiency-aware Noise Blending (SNB). Given two diffusion models, we first design a "Noise to salience map" module to obtain salience maps. After that, we can generate saliency-aware masks based on the salience maps. Finally, we blend the diffusion models in the noise space according to the mask. (*) $\hat{\epsilon}_g$ and $\hat{\epsilon}_e$ are noises instead of noisy images, and we add the image here just for visualization.

work, we propose to ensemble different pre-trained diffusion models in a novel dimension, *i.e.*, spatial, which can be applied to various scenarios.

3. Method

In this section, we introduce our proposed saliency-aware noise blending. Specifically, we first review the widely obtained classifier-free guidance, after that, we revisit such guidance and experimentally find that the classifier-free guidance is secretly a saliency indicator. Based on the salience map, we can obtain a saliency-aware mask, which is further used to guide the blending of the noise of two diffusion models. Figure 2 show the whole pipeline, and more details can be found in the following.

3.1. Preliminaries

Given a pre-trained text-guided diffusion model, we can generate images by a DDPM/DDIM sampling process, which progressively converts a Gaussian noise into an image for T timesteps. Take the DDIM sampling for example, a denoising step can be denoted as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}, \quad (1)$$

where t indicates the timestep, x is the noisy image, $\bar{\alpha}$ is related to a pre-defined variance schedule, and $\hat{\epsilon}$ is the predicted noise. The predicted noise can be re-modulated by classifier-free guidance [13], which is designed to extrapolate the output of the model in the direction of $\epsilon_\theta(x_t|c)$ and away from $\epsilon_\theta(x_t|\varnothing)$ as follows:

$$\hat{\epsilon} = \epsilon_\theta(x_t|\varnothing) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\varnothing)), \quad (2)$$

where ϵ_θ is the pretrained model, c is the text condition, $\varnothing$ is a null text, s is the guidance weight and increasing $s > 1$ strengthens the effect of guidance.

3.2. Noise to Salience Map

We experimentally find that the classifier-free guidance introduced above is secretly a saliency indicator. Specifically, $\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\varnothing)$ in Eqn. 2 indicates the difference between a conditional prediction and an unconditional prediction, thus the objects and scenes appeared in the text condition would be emphasized with a large value, especially when we adopt a large guidance scale s (*e.g.*, 10-100). We visualize the re-modulated noises of Eqn. 2 and find that the important region does have high responses. To this end, we propose to obtain a salience map by the following operation:

$$\Lambda(\epsilon_t) = Blur(Abs(\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\varnothing))) \quad (3)$$

where $\Lambda(\epsilon_t)$ represents the salience map, $Abs(\cdot)$ calculates the absolute value of the input variables, and $Blur(\cdot)$ is used to smooth the high-frequency noise, which can eliminate local interference responses and leverage the coherence of adjacent regions.

3.3. Saliency-aware Blending

The above discussions are all about a single diffusion model, and now let us move to the next stage, i.e., obtaining a blending mask based on two models' salience maps.

Given a general model and an expert model, we can obtain the corresponding salience maps by Eqn. 3, which are denoted as Λ^g and Λ^e, respectively. We first normalize these salience maps by the $softmax$ function:

$$\begin{aligned} \Lambda^{g'} &= softmax(k^g * \Lambda^g) \\ \Lambda^{e'} &= softmax(k^e * \Lambda^e), \end{aligned} \tag{4}$$

where k^g and k^e are hyper-parameters, *i.e.*, temperature of the $softmax$. Note that the $softmax$ here ensures the sum of each salience map to be a constant (*i.e.*, 1), which means each model must focus on some regions instead of having a high response everywhere. When we integrate a general model and an expert model, the salience map of the general model tends to cover multiple objects to compose the whole scene, while the expert model tends to have a higher response to a specific object. Thus we obtain a blending mask $\mathcal{M}$ by comparing these two salience maps:

$$\mathcal{M} = \arg\max(\Lambda^{g'}, \Lambda^{e'}) \tag{5}$$

The saliency-aware mask is an effective guide to perform noise blending, which consists of binary values of 0 and 1, corresponding to the noise ϵ_g and ϵ_e respectively. We can obtain the fused noise as follows:

$$\hat{\epsilon} = \mathcal{M} \odot \hat{\epsilon}_g + (1 - \mathcal{M}) \odot \hat{\epsilon}_e, \tag{6}$$

where $\odot$ denotes Hadamard (element-wise) Product, and we omit t for simplicity. In the specific implementation process, the fusion strategy is executed only when $t \leq t_s$, ensuring that the fusion results possess the fundamental structure of images generated by the general model. Algorithm 1 summarizes the process of the saliency-aware noise blending algorithm.

Additional explanations on the three applications. In the three applications (*i.e.*, a general model + a fine-grained model, a DreamBooth model, and a cartoon model) of this work, the expert model of the former two focuses on a specific object, while the last one contributes to the global structure of the generated image. In the last application, the cartoon model tends to focus on low-frequency structure and the general model focuses on high-frequency details. Thus the blended mask is not object-level, and we remove the blur operation in Eqn. 3 to facilitate such blending.

Clarification of technique novelty. The overall process of this algorithm is quite simple, yet it is non-trivial by solving two challenges. Firstly, we leverage the classifier-free guidance to automatically identify each model's areas of expertise. The introduction of the hyper-parameter k provides improved controllability for blending the two sources of images, thereby enabling greater creativity and flexibility in image generation based on SNB. Secondly, the task of obtaining saliency response values that are closely linked to prompt content is highly non-trivial. Secondly, in each sampling step, the two models take as input the blended x_t, enabling automatic semantic alignment of the two models' noise space. We believe our exploration would contribute to the community and benefit the leveraging of pre-trained diffusion models.

Algorithm 1 Saliency-aware Noise Blending

Input: Two pre-trained models ϵ_{θ_g} and ϵ_{θ_e}, along with two prompts, y_g and y_e. gradient scale s in Eq.(2). Hyper-parameters k^a and k^b in Eq. (4) and fusion time t_s.
Output: The fused image x_0

$x_T \sim \mathcal{N}(0, I)$
for t from T to 0 **do**
  if $t > t_s$ **then**
    $\epsilon_t = \epsilon_{t_g}$
  else
    $\epsilon_{t_g} = \epsilon_{\theta_g}(x_t|c_g)$,$\epsilon_{t_e} = \epsilon_{\theta_e}(x_t|c_e)$
    get $\Lambda^g(\epsilon_{t_g})$ and $\Lambda^e(\epsilon_{t_e})$ according to Eq.(3).
    get $\Lambda^{g'}(\epsilon_{t_g})$ and $\Lambda^{e'}(\epsilon_{t_e})$ via Eq. (4).
    $\mathcal{M} = \arg\max(\Lambda^{a'}, \Lambda^{b'})$
    get the noise generated by classifier-free guidance $\hat{\epsilon}_{t_g}$ and $\hat{\epsilon}_{t_e}$ via Eq.(2).
    $\epsilon_t = \mathcal{M} \odot \hat{\epsilon}_{t_g} + (1 - \mathcal{M}) \odot \hat{\epsilon}_{t_e}$
  end if
  $x_{t-1} \leftarrow \epsilon_t$ via Eq.(1)
end for
return x_0.

4. Applications

In order to evaluate the effectiveness of our proposed method, we conduct experiments on three challenging applications. 1) Fine-grained Fusion, *i.e.*, fusing a general and a fine-grained model to achieve fine-grained generation with complex scenes. 2) Recontextualization, *i.e.*, fusing a general and a DreamBooth [28] model allows for the recontextualization of specific objects with well-preserved details. 3) Cross-domain Fusion, *i.e.*, fusing a general and a cartoon model to combine the creative advantages of the cartoon model to generate complex scenes and the photorealistic fidelity of the general model. These experiments will allow us to evaluate the performance of our method across a range of scenarios and provide valuable insights into the strengths and limitations of our approach.

4.1. Application 1: Fine-grained Fusion

The stable diffusion model [27] trained on large-scale multimodal datasets like LAION 5B [31] has shown impressive performance on general text-to-image synthesis. In our experiments, we use the publicly released stable diffusion v1-4. Meanwhile, fine-grained car models trained on (an

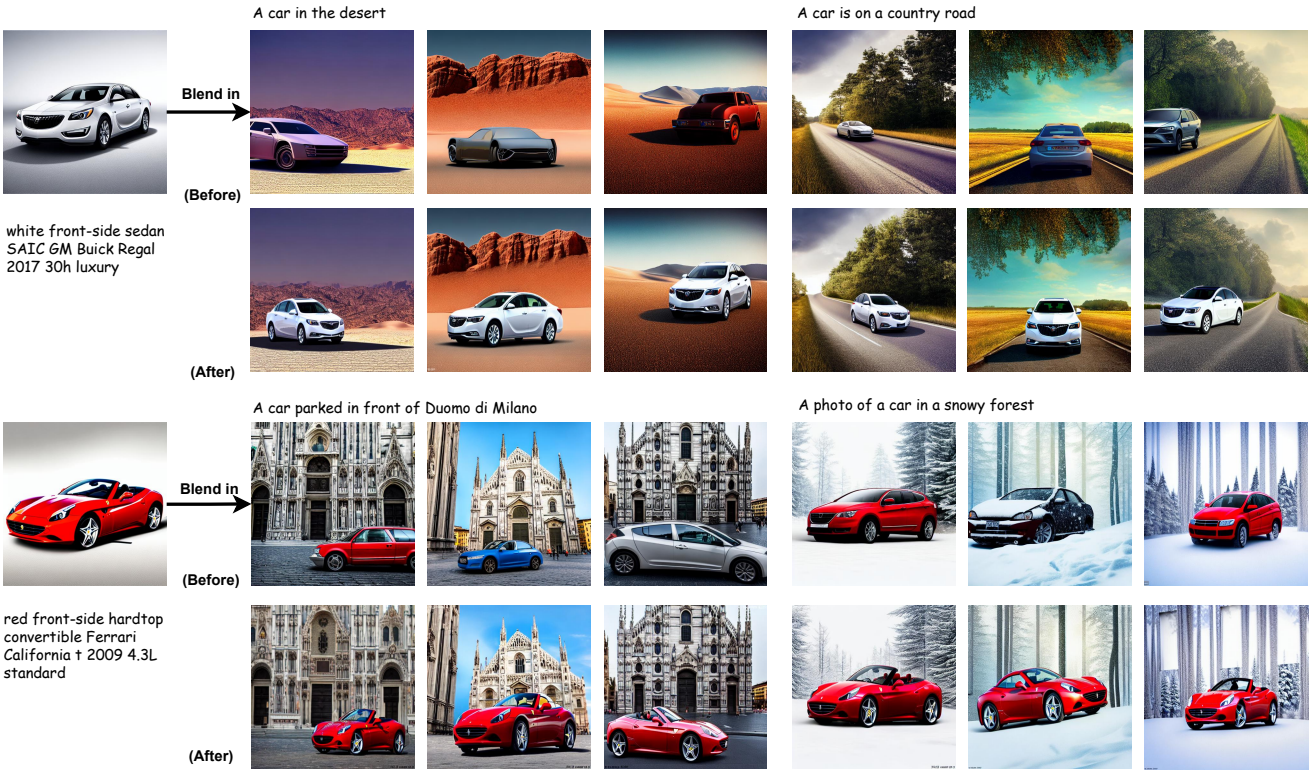

Figure 3: Application 1: visualization results of Fine-grained Fusion. Our method enables fine-grained generation with complex scenes.

extension of) CompCars [38] can generate fine-grained car images with specific colors, viewpoints, types, brands, and models.

For instance, a scene description like "a photo of a red car in a snowy forest" can be fed into the general model, resulting in a corresponding image as shown in the first left column of Figure 3. Similarly, a specific car description like "red hatchback Audi (imported) Audi A1 2010 e-tron" can be fed into the car model, generating an Audi car that matches the prompt. To fuse the noise of the two models during the denoising sampling process, we propose the SNB method, which can produce a fused image of a red Audi hatchback car in a snowy forest. The fusion results of the general model and the fine-grained car model are illustrated in Figure 3, showing the ability to replace the car in the scene with a specific one while retaining the original scene unchanged. Notably, our SNB does not require additional annotations to specify the car's position in the original image, enabling automatic semantic alignment.

4.2. Application 2: Recontextualization

Recontextualization, or named personalizing text-to-image generation, is proposed in previous works [9, 28], which aims to generate a creative scene for a specific object/concept. DreamBooth [28] proposes to fine-tune a diffusion model on several given images together with a placeholder word to enable the model to generate a specific object/concept. In this application, we integrate a general and a DreamBooth [28] model to allow for the recontextualization of specific objects with well-preserved details.

Specifically, we first fine-tune the general model using multiple images of the target object. Thus the fine-tuned model can represent the specific object with the placeholder "[ ]" in the prompt. Next, we get through a sentence that describes a complex scene and "a photo of a [ ] $< class >$" into SNB to integrate the general model and the DreamBooth model. Results can be found in Figure 4. It can be observed that our method can put a specific object with rich details into a complex scene.

4.3. Application 3: Cross-domain Fusion

Different models have distinct advantages when it comes to generating images with specific styles. For example, cartoon models are particularly skilled at creative composition and can generate scenes that are rarely observed in real-world scenarios. When we strive to generate images with unique and imaginative compositions, we can integrate a cartoon model to generate a distinctive scene and a general

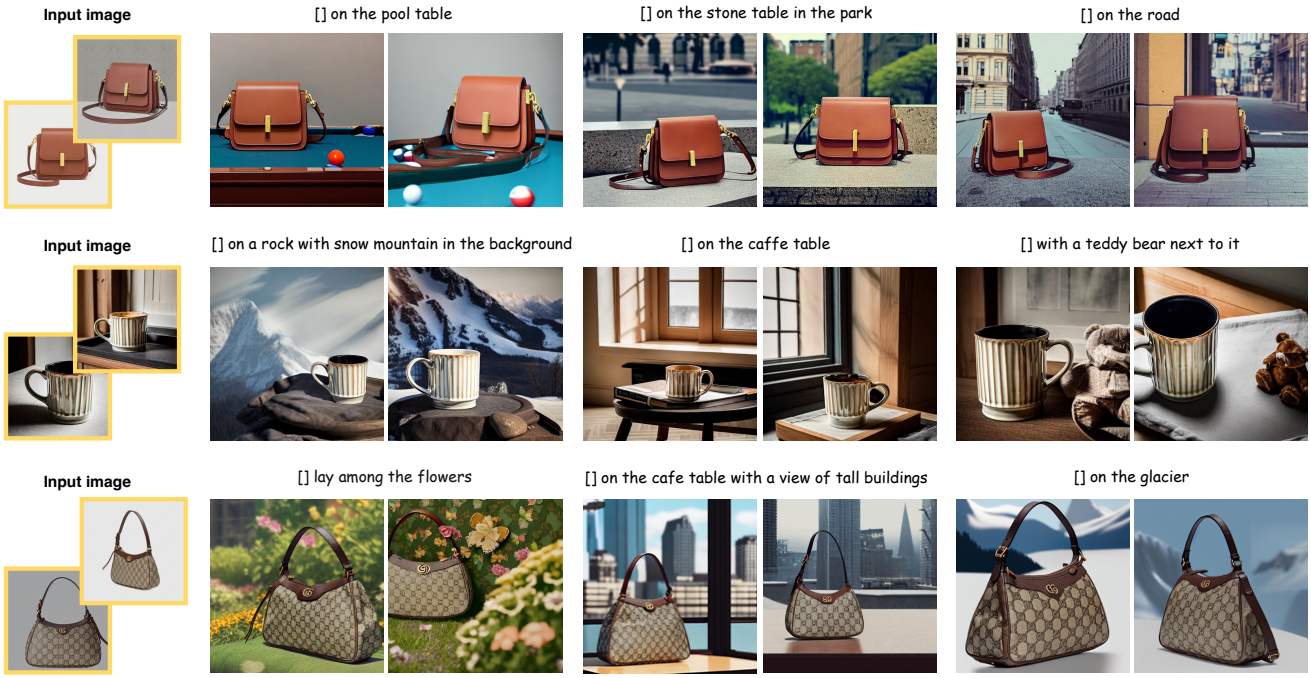

Figure 4: Application 2: visualization results of Recontextualization. Our method can recontextualize specific objects in complex scenes with well-preserved details.

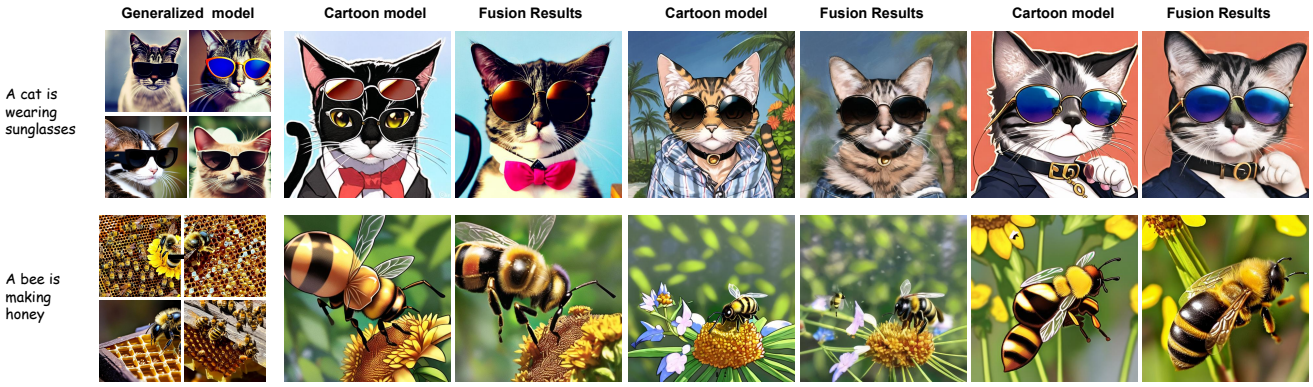

Figure 5: Application 3: visualization results of Cross-domain Fusion. Our method can combine the creative advantages of the cartoon model to generate complex scenes and the photorealistic fidelity of the general model.

model to make the content more realistic.

In this application, the same textual description of a scene is fed into both the general and the cartoon models. This allows us to generate two sets of noise that correspond to the same content but different styles during the sampling process. By fusing these two sets of noise, SNB generates images that exhibit both a creative composition and realistic content. As shown in Figure 5, our proposed SNB provides a powerful tool for achieving the balance of creative and realistic, and we believe that such a tool has the potential to benefit a wide range of applications in various fields, such as art and AI-aided design.

5. Results and Comparisons

5.1. Compared to the General Model

The stable diffusion model is a widely used and versatile generative model in computer vision that can generate a diverse range of images based on prompts. However, when the prompts contain multiple content subjects, the general model's generation performance can become challenging, especially when the combination of these subjects is rare in real-world scenes. This limitation can result in missing subjects in the generated images, as observed in Application 1, where the general model fails to capture both the car and

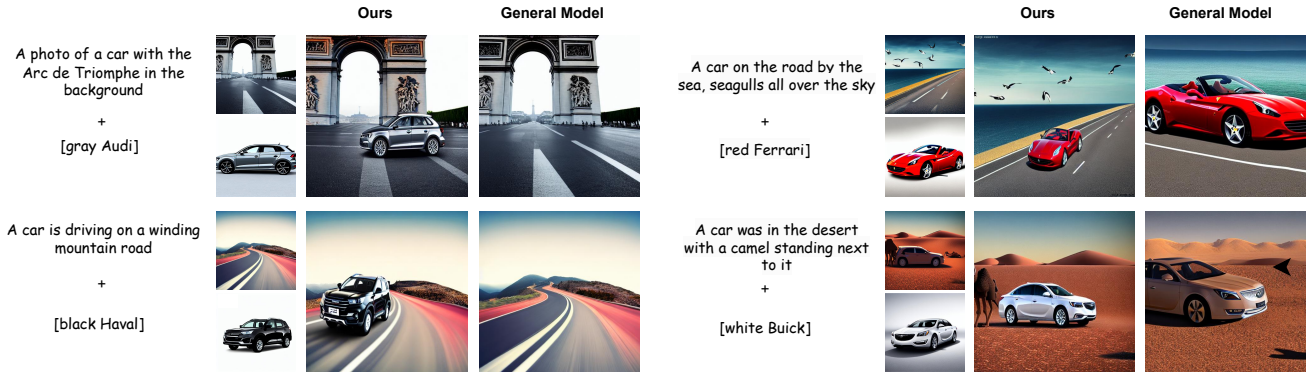

Figure 6: Compared to the general model. The general model fails to precisely generate fine-grained objects and tends to miss some objects when the prompt is full of details.

Figure 7: Compared to DreamBooth [28]. Dreambooth struggles with preserving the details of the given objects, especially for complex and lengthy prompts.

the seagulls in the same scene. A similar issue is observed in Application 3 (Figure 1), where the general model cannot produce scenes that are uncommon in real-life, such as a lion wearing a crown (Figure 6). In addition, the general model has difficulty understanding the semantics of certain prompts, as observed in the case of "a bee is making honey" in Figure 5, where the model only combines the bee and honey in the same scene.

The proposed SNB can address these limitations, which integrates a specialized model, such as a model that focuses on cars in Application 1, or a cartoon model in Application 3, to improve the accuracy and realism of image generation. By leveraging the strengths of both the general and specialized models, our proposed approach can generate visually appealing and semantically rich images. This approach has the potential to advance the field of computer vision by enabling more sophisticated and realistic image generation.

5.2. Compared to Dreambooth

Dreambooth [28] presents an approach for synthesizing novel renditions of a given subject using a few images of the subject and the guidance of a text prompt, *i.e.*, recontextualization. However, it is challenging for Dreambooth to handle certain creative scene compositions, such as "a [] backpack on the moon." Our empirical investigation reveals that Dreambooth struggles with integrating specific subjects into the scene, particularly in cases where the scenario is uncommon in reality. As illustrated in Figure 7, Dreambooth frequently produces images that either lack or exhibit inconsistencies in terms of specific subjects in such scenarios. In contrast, our approach excels at integrating a specific subject into the scene while maintaining fidelity to the prompt. Although Dreambooth can generate contextually relevant images for brief prompts like "a [] bag on a garden bench," it may not accurately capture the finer details of the subject, such as the graphic texture of the bag. In contrast, our approach not only generates images that adhere to the prompt but also reproduces subject details with greater accuracy.

5.3. Compared to Annotated Masks

Blended-Diffusion [3] firstly propose to blend the noisy image in each sampling step based on a given mask for image editing. Such a method can neither handle applica-

Figure 8: Compared to annotated masks. Our saliency-aware mask achieves better composing performance, e.g., generating snow on the car.

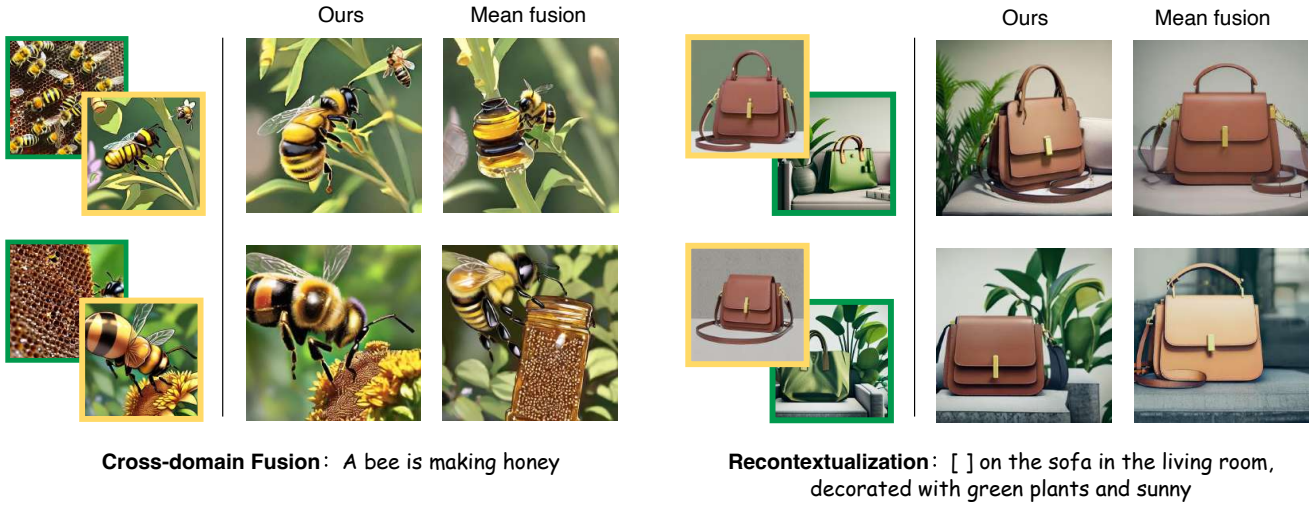

Figure 9: Compared to weighted sum. Directly averaging the two noises cannot well preserve each model's strengths and leads to cluttered image content.

tions like cross-domain fusion nor perform well on a mixed object and scene content, such as "a car in a snowy forest." As depicted in Figure 8, a comparison between our method and a mask-based approach demonstrates that our SNB method accurately preserves detailed subject features influenced by the scene, like the snow on the car's hood and wheels. In contrast, the mask-based method replaces the entire car area, resulting in a loss of subject details within the scene. Overall, our SNB yields more natural and realistic results when editing or replacing scene content, particularly in scenarios with mixed object and scene content.

5.4. Compared to Weighted Sum

The success of SNB hinges upon its ability to create a high-quality mask based on the noise generated by the two models, which is then used to determine content coverage and retention. As shown in Figure 9, the experimental results of SNB and the fusion method that directly averages the two noises differ significantly. When the weighted sum of two noises is used, the resulting image content appears cluttered and lacks the desired semantic alignment. In contrast, our SNB method outperforms the direct averaging method by achieving precise semantic alignment, resulting in more accurate and realistic image content.

It is worth noting that, CDM [18] is belong to the case of weighted sum, which treats two noises equally in different regions, while we trust different models in different regions by designing a saliency-aware fusion mechanism. We further show the comparison results with CDM in Fig. 10. CDM fails to generate accurate car images in the case of complex prompts (left) and tends to produce conceptually mixed images or exhibits unstable performance (right). In contrast, our proposed method can accurately generate specified images of cars as well as vivid depictions

Figure 10: Comparisons with CDM. In the case of complex prompts (left), CDM fails to generate accurate car images. Moreover, CDM tends to produce conceptually mixed images (mid) or exhibits unstable performance (right).

	Text-Image Similarity		Image Similarity
	App. 1 ↑	App. 3 ↑	App. 2 ↑
Ours	**31.18**	**32.76**	**0.8401**
Baseline	29.56	31.01	0.7582

Table 1: The CLIP-based Quantitative Metrics.

	App. 1	App. 2	App. 3
Ours better	81.4%	84.7%	79.3%

Table 2: Human evaluations.

of lions wearing crowns. Overall, our method demonstrates significantly better performance compared to CDM.

5.5. Quantitative Metrics

We conducted quantitative evaluations for three corresponding applications. Specifically, we generated 5k test samples for each application to calculate the CLIP-based score. For application 1 and application 3, we measured the similarity between the generated images and the corresponding prompts in terms of Text-Image Similarity; for application 2, we calculated the similarity between the recontextualized images and the input images, which is referred to as Image Similarity. The evaluation results are presented in Table 1. Stable diffusion serves as the baseline for application 1 and application 3, while Dreambooth [28] serves as the baseline for application 2. Based on the results, it can be concluded that our method has demonstrated improved performance across all three applications.

In addition, we also conducted 500 human evaluations for each application and presented the results in Table 2. In application 1 and application 3, we compare our model to stable diffusion, and in application 2, we compare it to Dreambooth. The data indicate that for all three applications, the majority of the participants regarded the images generated by our method as superior in quality.

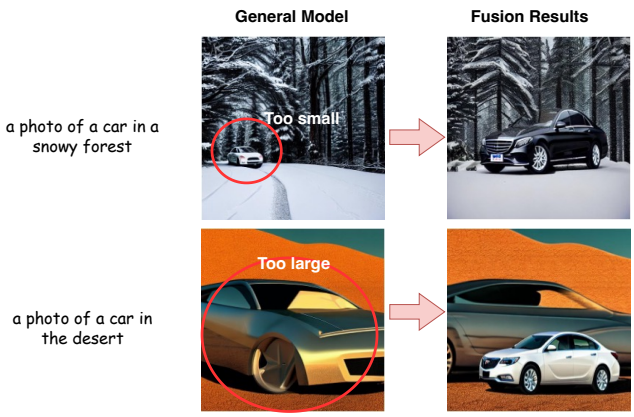

Figure 11: Failure cases. Misalignments may appear due to the various car size.

6. Conclusion

In this work, we study the problem of integrating pre-trained text-guided diffusion models to achieve more controllable generation. We propose a simple yet effective Saliency-aware Noise Blending (SNB) to preserve the strengths of each individual model. Extensive experiments on three challenging applications (*i.e.*, a general model + a cartoon model, a fine-grained car model, and a DreamBooth model) show that SNB can significantly empower pre-trained diffusion models. With the rapid development of pre-trained large generative models, we believe our work is of great value.

Limitations. Although our method has demonstrated significant performance improvements, it may not be applicable in certain scenarios. The advantage of SNB is its ability to form automatic masks based on saliency response values; however, when the difference in object size between the two merged images is too large, semantic alignment can be difficult to achieve. As shown in Figure 11, when the general model produces cars that are either too large or too small, the results display misaligned semantic positions. In the future, we will keep on studying the integration of different large generative models and extend our method to more general settings.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China: No. 91948303-1, 611803375, No. 61803375, No. 12002380, No. 62106278, No. 62101575, No. 61906210; the Postgraduate Scientific Research Innovation Project of Hunan Province: QL20210018 and the National Key R&D Program of China (2021ZD0140301).

References

[1] Fine-tuned stable diffusion model trained on screenshots from a popular animation studio. https://huggingface.co/nitrosocke/mo-di-diffusion. 2022.

[2] Merging two diffusion models. https://stable-diffusion-art.com/models/#merging_two_models. 2022.

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. eDIFF-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[5] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[7] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021.

[8] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, 2014.

[11] D Gopika and B Azhagusundari. An analysis on ensemble methods in classification tasks. 2014.

[12] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic object accuracy for generative text-to-image synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1552–1565, 2020.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[14] Nupur Kumari, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Ensembling off-the-shelf models for gan training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10651–10662, 2022.

[15] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.

[16] Gang Li, Heliang Zheng, Chaoyue Wang, Chang Li, Changwen Zheng, and Dacheng Tao. 3ddesigner: Towards photorealistic 3d object generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2211.14108*, 2022.

[17] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. Object-driven text-to-image synthesis via adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12174–12182, 2019.

[18] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV (17)*, volume 13677 of *Lecture Notes in Computer Science*, pages 423–439. Springer, 2022.

[19] Joao Mendes-Moreira, Carlos Soares, Alípio Mário Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. *Acm computing surveys (csur)*, 45(1):1–40, 2012.

[20] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[21] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in neural information processing systems*, 32, 2019.

[22] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[25] Ye Ren, PN Suganthan, and N Srikanth. Ensemble methods for wind and solar power forecasting—a state-of-the-art review. *Renewable and Sustainable Energy Reviews*, 50:82–91, 2015.

[26] Lior Rokach. Ensemble-based classifiers. *Artificial intelligence review*, 33:1–39, 2010.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed

Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[32] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[33] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[34] Ming Tao, Hao Tang, Songsong Wu, Nicu Sebe, Xiao-Yuan Jing, Fei Wu, and Bingkun Bao. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv preprint arXiv:2008.05865*, 2020.

[35] Sandro Vega-Pons and José Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 25(03):337–372, 2011.

[36] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2487–2495, 2019.

[37] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *Cornell University - arXiv*, 2017.

[38] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3973–3981, 2015.

[39] Pengyi Yang, Yee Hwa Yang, Bing B Zhou, and Albert Y Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010.

[40] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021.

[41] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021.

[42] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6199–6208, 2018.

[43] Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, Wanrong Huang, and Wenjing Yang. Null-text guidance in diffusion models is secretly a cartoon-style creator. *arXiv preprint arXiv:2305.06710*, 2023.

[44] Ying Zhao, Jun Gao, and Xuezhi Yang. A survey of neural network ensembles. In *2005 international conference on neural networks and brain*, volume 1, pages 438–442. IEEE, 2005.

[45] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5810, 2019.