This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Conditional Balance: Improving Multi-Conditioning Trade-Offs in Image Generation

Nadav Z. Cohen[1], Oron Nir[1,2], Ariel Shamir[1]
[1]Reichman University, [2]Microsoft Corporation

https://nadavc220.github.io/conditional-balance.github.io/

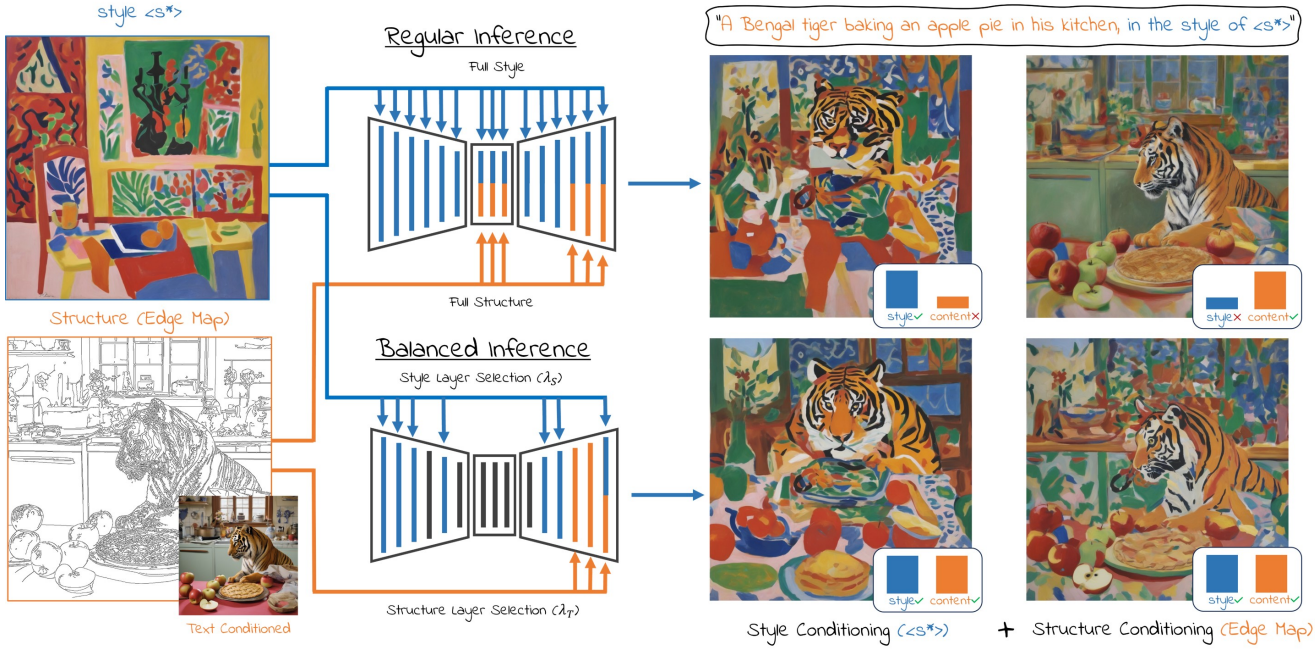


Figure 1. **Balanced Conditioning Image Generation.** *We analyze the sensitivity of model layers to various aspects in conditional inputs. This allows to limit the inputs only to specific layers during inference, thereby balancing the different conditions and preventing content and style from overshadowing each other. As a result, the generative model reduces artifacts and gains artistic freedom when combining complex conditional inputs. As can be seen, by selecting only highly sensitive layers of style (λ_S) and structure (λ_T) we get better color and texture control and better geometric style control.*

Abstract

Balancing content fidelity and artistic style is a pivotal challenge in image generation. While traditional style transfer methods and modern Denoising Diffusion Probabilistic Models (DDPMs) strive to achieve this balance, they often struggle to do so without sacrificing either style, content, or sometimes both. This work addresses this challenge by analyzing the ability of DDPMs to maintain content and style equilibrium. We introduce a novel method to identify sensitivities within the DDPM attention layers, identifying specific layers that correspond to different stylistic aspects. By directing conditional inputs only to these sensitive layers, our approach enables fine-grained control over style and content, significantly reducing issues arising from over-constrained inputs. Our findings demonstrate that this method enhances recent stylization techniques by better aligning style and content, ultimately improving the quality of generated visual content.

1. Introduction

To master different aspects of paintings such as color and light, a notable approach practiced by artists involves creating a collection of paintings of the same subject under varying conditions. A well known example of this practice is Claude Monet's series of paintings of Rouen Cathedral [40].

Today, modern Denoising Diffusion Probabilistic Mod-

els (DDPMs) allow creating high-quality images of any subject in various styles by iteratively refining random samples. To direct these models to output an image with a desirable content at inference, conditional inputs were developed, starting from descriptive text prompts and continuing with images that condition the output to align with content and style information. This process, while powerful, tends to lose its conditional constraining ability as the number or complexity of constraints increase. As a result, in addition to losing the ability to control the generated content, the model's attempt to satisfy all conditionals leads to more undesirable issues such as image artifacts.

In this work, we investigate the style-content trade-offs of different conditional inputs separately and in combination in generative models. Our experiments reveal that many issues in conditional generation arise from over-conditioning and the combination of conditionals that were underrepresented during model training. We analyze a diffusion generation process (using SDXL's architecture [26]) and isolate different aspects of its generative capabilities.

Drawing inspiration from Monet's many series of paintings, our analysis uses a collections of images where the content subject is fixed, and a specific stylistic aspect is varied. Using such collections we examine and rank the sensitivity of each layer at each timestep. Later, we direct different conditionals only to specific, sensitive layers, allowing to better balance different conditions (see Fig. 1). We show that even without pure disentanglement this approach reduces artifacts caused by over-conditioning and enhances the model's overall output quality, stylistic freedom, and consistency without the need of additional training.

We summarize our contributions as follows:

1. We develop a novel analysis method inspired by classic art and simple modern statistics which reveals sensitivities in the generation process of diffusion models.
2. We analyze SDXL and use our findings to reduce conditional inputs in the generation process. This leads to balance, artifact reduction, and general image quality improvement for text, image, and style conditioning inputs.
3. Using our findings also allows a flexible way to control content and style aspects of the generated image, leading to a more stable and creative method to generate content.

Our complex-conditioning evaluation set and code are available through our project page.

2. Related Work

Content-Style Applications. Seminal style transfer methods [10, 15, 19, 21, 22], explore different ways to transfer a style of a stylistic image to a given "content image." These works demonstrate the natural trade-off between artistic style and content preservation, and explore ways to combine them in a visually pleasing way. Later, inspired by the success of Generative Adversarial Networks [16], more methods were developed for styled generation [23, 45] enabling generating images from noise samples. Although these methods show improvement in combining content and style, they lack flexibility as they can not generate combinations outside of their optimized domain.

Diffusion Models and Conditioning. DDPMs [9, 26, 28, 30, 33] achieved a significant leap in generating novel images conditioned on text prompts by refining over text-image paired datasets [34]. While these models combine content and style effectively, their sole reliance on text conditioning comes with shortcomings, such as failure to align to complex text conditions [24, 39, 42] and an inferior ability to generate content and style combinations which were underrepresented during training.

To overcome these limitations, various conditioning methods were developed, influencing the generated image to resemble information from another image. ControlNet [43] conditions the output with structure information using image maps such as Canny, depth, and pose. Gal et al. [13], and others [2, 14, 32], condition the generated image to preserve unique (personalized) properties of an object.

Image-based style conditioning is a thoroughly researched area [6, 8, 12, 17, 20, 31, 36, 38, 41], usually demonstrated on SDXL [26] for its artistic superiority. IP-Adapter [41] and InstantStyle [38] inject the style using a pretrained dedicated Cross-Attention layer [37] for the conditioning image. B-LoRA [12] and ZipLoRA [35] uses LoRA [18] to fine-tune residual weights to match content and style of conditioning images. StyleAligned method [17] generates a style image in parallel to generating the output image and injects the style information using AdaIN [19] between Self-Attention layers [37] while Jeong et al. [20] extend this idea by replacing the attention feature directly.

Even though recent methods address content and style combinations, they still remain a challenge as methods which excel in style, often sacrificing content fidelity [17, 20], while others which preserve content fidelity may often show inaccurate style [12, 38].

DDPM Model Analysis. Recent works [12, 20, 38, 44] analyze the diffusion process to pinpoint parts responsible for generating various visual aspects. ProSpect [44] examine how varying text conditions at different timesteps impact aspects like material, artistic style, and content alignment. Our findings align with theirs, revealing that each timestep serves a unique conditioning function. However, we extend this analysis by investigating the model's internal layers, allowing for the use of multiple conditional inputs without interference. While ProSpect focuses exclusively on text conditioning, our study incorporates both text and

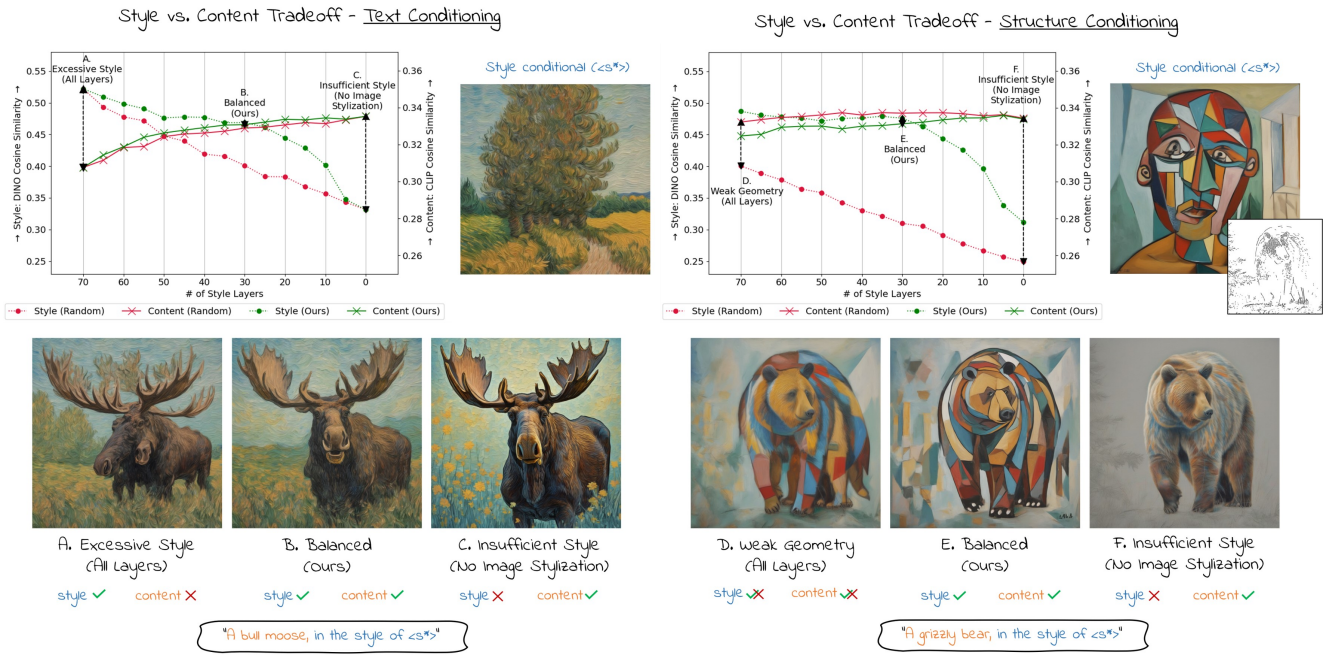


Figure 2. **Preliminary Experiment Results.** *We present outcomes of selecting sub-set of layers to apply conditioning. We show the content/style tradeoff both with text conditioning (left) and with edge-map conditioning (right). As can be observed in both cases using too much or not enough style conditioning generates an imbalanced output. Selecting layers randomly does not assist much while using our layer reduction strategy enables achieving good quality conditional combinations with less than 50% of the layers. Note that the content and style graphs are presented on different scales as Clip and Dino similarity values are not aligned.*

image conditioning. Style conditioning works [12, 38], examine internal layers to pinpoint those sensitive to content and style. B-LoRA tests prompt stylization effects in different attention blocks and employs LoRA [18] to fine-tune style sensitive weights for specific images. InstantStyle analyzes SDXL attention layers for content alignment and style sensitivity and concludes the same style-sensitive blocks. In contrast, our approach thoroughly evaluates each self-attention layer independently, ranking its sensitivity for each timestep. Additionally, our method can analyze other attention architectures, like the Joint-Attention layer used by the more recent SD3 model family [11]. This methodology maximizes style and content conditioning while minimizing interference between layers, thus enhancing stylistic freedom and reducing content artifacts.

3. Content-Style Tradeoff

In this section, we explore the issue of over-constrained conditional image generation. To illustrate the problem, we examine two recent style conditioning methods: StyleAligned [17] and B-LoRA [12].

To investigate the relationship between content and style, we first choose 10 different artistic style conditionals (see supplemental file for details) and generate images with textual prompts either with or without edge conditional maps (using Canny edges of a real photograph). The textual prompts were divided into two categories: Easy and Complex. A prompt is considered complex based not only on its length and details but also on how far its content deviates from the typical subjects of the style. For example, generating a 'fantasy' theme is harder than a 'rural-life' theme for an artistic style that follows Van Gogh, since his works often depict European rural life in the late 19^{th} century. This division allows us to isolate the impact of textual conditional alone and better understand its limitations.

Next, we evaluate the style and content of all generated images. We follow recent works [17, 20, 38] and evaluate *style* using the cosine similarity of the Dino embeddings [5] between the style image and the generated image. For *content* we calculate the cosine similarity between Clip embeddings [27] of the text prompt and the generated image.

StyleAligned method applies stylization across all SDXL self-attention layers, often resulting in images that are misaligned with the text prompt and exhibit noticeable artifacts. To alleviate this, we gradually reduce the number of layers for stylization randomly (from 70 to 0). In Fig. 2 (left) we show both a plot of quantitative results of averaging across all generated images and a specific qualitative example. As can be seen, reducing the number of layers significantly improves text alignment. However, while text compliance improves, style fidelity decreases more or

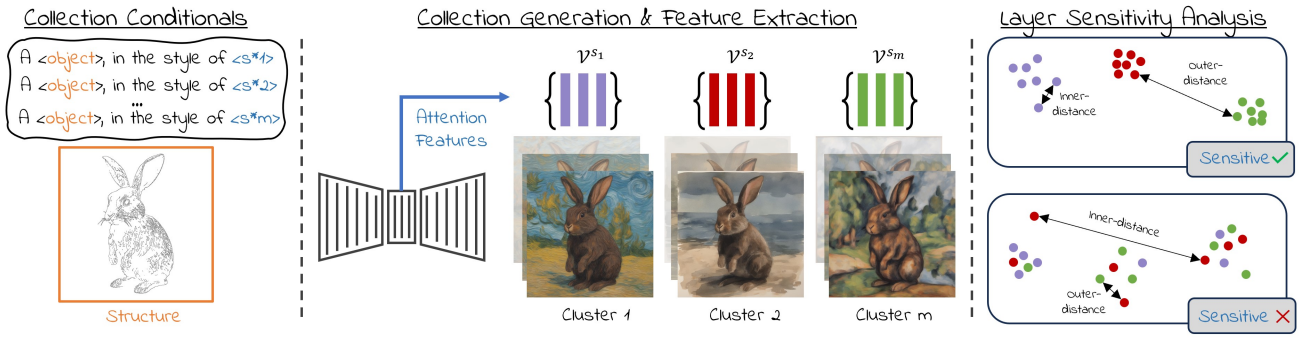


Figure 3. **Method Overview.** *We generate a collection of images isolating a specific aspect of interest. Then, during generation, we extract self-attention features for each layer, at each timestep to represent the image. We calculate each layer's clustering score of all images sharing the same artistic aspect. The score measures both inner and outer cluster distances between all images in this representation. Finally, we rank each layer sensitivity to this aspect by its clustering score and use this information in our balancing strategy.*

less linearly as fewer layers are stylized. Hence, we find that using all layers for stylization is an indication of *Style Over-Conditioning* phenomenon.

We also observe the opposite effect which we term *Content Over-Conditioning*. This phenomenon is caused, for example, by using challenging (Complex) text prompts or by using conditional maps like edge and depth maps. In this case, even when using all layers for stylization there is reduction in style consistency, such as appears in the geometric aspects of the target style. In Fig. 2 (right) we show both a plot of quantitative results of averaging across all generated images using an edge-map condition, and a specific qualitative example. This issue appears with or without image-based style conditioning (as illustrated in Fig. 4), which suggests that the output is over-constrained by the content map rather than lacking more style conditioning.

We used a similar analysis for B-LoRA method (see supplemental file) and found that the basic method often suffers from under-stylization, resulting in images that lack style. In addition, both methods suffer from a significant style loss when conditioned with challenging content inputs.

These experiments suggest there is a content-style trade-off that can possibly be alleviated by decreasing the strength of the different conditions for better balance between them. This can be done by limiting the influence of style and content conditionals by applying them only to a sub-set of the layers. However, simply selecting random layers does not provide stable results and generates inconsistent outputs. This motivated us to identify a smaller subset of style-sensitive layers that could achieve effective stylization while still maintaining good content alignment of various types (text, shape, geometry).

Fig. 2 shows our method's results of balancing content and style by judiciously selecting layers compared to the random layer selection. As can be seen, the potential in style alignment increases dramatically faster than that of the random selection and is significantly higher in all number of layers without sacrificing content information. Additionally, style levels are preserved when using an edge-map condition, unlike the random layer selection which drops dramatically compared to text-only conditioning.

In the following section, we expand our approach, which includes a novel method for analyzing style sensitive layers in diffusion models, and a way to apply this knowledge to balance the use of conditionals at inference time without the need for any additional training.

4. Method

Hertz et al. [17] demonstrate that style information is embedded in attention layers during generation and applies it for stylization. Building on this insight, we analyze attention layer data to assess style sensitivity at each timestep. Since this information is encoded in the network, our approach implicitly extracts it using analysis of a collection of images created in varied artistic styles to interpret these sensitivities (see Fig. 3).

Image Collection. We first choose m distinct styles such as Claude Monet and Winslow Homer, and for each style $s\,(1 \leq s \leq m)$ we create a cluster C_s by generating n images where a single stylistic aspect is repeated multiple times in different images by using various random seeds. This results in a collection of $m \times n$ images.

To isolate a particular artistic aspect within each cluster, we impose the condition that, apart from the aspect being examined, other aspects (such as the subject) must either remain constant across the collection or vary consistently throughout. This ensures that the analysis is focused solely on the chosen aspect, as will be further discussed Sec. 5.

Image mapping. Hertz et al. [17] align a series of generated images to a single reference style image by using Adaptive Instance Normalization (AdaIN) [19] between the self-attention layers in the denoising UNet. For each layer, the Key and Query features are extracted from the style image and projected to an $[S, H, D]$ shaped tensor where S is the number of pixels, H is the number of attention heads and D is the dimension of each head. They calculate the mean and std of S, resulting in two mean and std vectors of shape $H \times D$ which are used to normalize the features of the output images. We use the same method and calculate the vector of the means and standard deviations (μ, σ) for each image in our collection from each layer before applying AdaIN. we treat this vector $\mathcal{V}$ of (μ, σ) as the image representation in each layer for each timestep.

Our key idea for evaluating the sensitivity of each layer at each timestep is to measure how well the structure of the space created by mapping all images, using their representation $\mathcal{V}$, aligns with the real clusters in the collection of images. A Good clustering score means this layer (in this timestep) has high sensitivity to the aspect represented by the collection. We rank all layers based on the clustering score and use this ranking to select the top K layers for this aspect conditional injection (see Fig. 3).

Clustering Score. The representation of each image i from style s, $\mathcal{V}_i^s$ is defined as a multi-dimensional vector of (μ, σ). We treat this representation as simple multi-dimension Gaussian distributions (where the covariance matrix has entries only on the main diagonal), and measure distances in feature space using Jensen-Shannon Divergence (JSD) metric, which is calculated as follows:

$$JSD(\mathcal{V}_1, \mathcal{V}_2) = \frac{1}{2}(D_{KL}(\mathcal{V}_1, M) + D_{KL}(\mathcal{V}_2, M)) \quad (1)$$

Here, D_{KL} is the KL-Divergence, $\mathcal{V}_1$ and $\mathcal{V}_2$ are the two Gaussian distributions being compared, and M is the average Gaussian distribution defined by the means and standard deviations of $\mathcal{V}_1$ and $\mathcal{V}_2$.

Next, for a given layer l at timestep t, we calculate the "inner distance" score by computing the average distance between all pairs of Gaussians within the same cluster C_s (paintings with the same artistic style):

$$D_{l,t}^{in} = \frac{1}{m}\sum_{s=1}^{m}\frac{1}{\binom{n}{2}}\sum_{i,j \in C_s} JSD(\mathcal{V}_i^s, \mathcal{V}_j^s) \quad (2)$$

where $\mathcal{V}_i^s$ and $\mathcal{V}_j^s$ represent Gaussians for two different paintings within the same cluster C_s.

In addition, we calculate the "outer distance," which is the average distance between Gaussians from different clusters (paintings with different artistic styles):

$$D_{l,t}^{out} = \frac{1}{\binom{m}{2}n^2}\sum_{s_1 \neq s_2} JSD(\mathcal{V}_i^{s_1}, \mathcal{V}_j^{s_2}) \quad (3)$$

where $\mathcal{V}_i^{s_1}$ and $\mathcal{V}_j^{s_2}$ are Gaussians from two different clusters C_{s_1} and C_{s_2}.

Our objective is to identify layers that bring similar styles closer together and push different styles further apart. This means we aim to minimize $D_{l,t}^{in}$ and maximize $D_{l,t}^{out}$. Therefore, to evaluate the clustering, we calculate the ratio of inner to outer distances as follows (similar to Dunn index [4]):

$$G_{l,t} = \frac{D_{l,t}^{in}}{D_{l,t}^{out}} = \frac{\binom{m}{2}n^2 \sum JSD(\mathcal{V}_i^s, \mathcal{V}_j^s)}{\binom{n}{2}m \sum_{s_1 \neq s_2} JSD(\mathcal{V}_i^{s_1}, \mathcal{V}_j^{s_2})} \quad (4)$$

After computing $G_{l,t}$ for each layer l and timestep t, we rank the layers from lowest to highest. This ranking indicates the sensitivity of each layer in each timestep to the artistic aspect being analyzed in the image collection. Then, to balance different conditions, we apply conditioning only on the K most sensitive layers instead of all the layers. This allows for better balancing between multiple conditions and creates more stable results as we will demonstrate in Sec. 6.

5. Analysis

Using our method we conduct an analysis of SDXL layers attempting to identify layers which are sensitive to different conditioning aspects. We concentrate on two popular aspects: style and structure. Each analysis was repeated 5 times using different objects to ensure robustness, and the final layer grading for each layer at each time step was averaged neglecting the best and worst grade to prevent outliers.

We conduct a similar analysis on the SD3.5-Large architecture [11]. Details can be found in the supplemental file.

5.1. Style Sensitivity

To analyze style sensitivity we create a collection featuring 10 artistically varied styles but using a single object for content which is also constrained by a canny edge-image. For each style we generate 5 images using text-prompts of the format "<*content prompt*>, <*style prompt*>" which creates a collection with 50 images. A sample of the collection can be seen in Fig. 3, while the entire collection is presented in the supplemental file.

Using this data we apply our method and find that the dependency on timestep is small so we rank layers for any timestep for style sensitivity. Fig. 2 shows results of applying style conditioning on decreasing subsets of layers from 70 to 0 based on this ranking. As can be observed, compared to a random layer choice, our ranking yields subsets with stronger style consistency. This results in better stylized output generation, while preserving content alignment and lowering the conditional burden of the style image. To better balance the different conditions we found that in most cases conditioning only on less than half of the layers (K=30) yields best results.

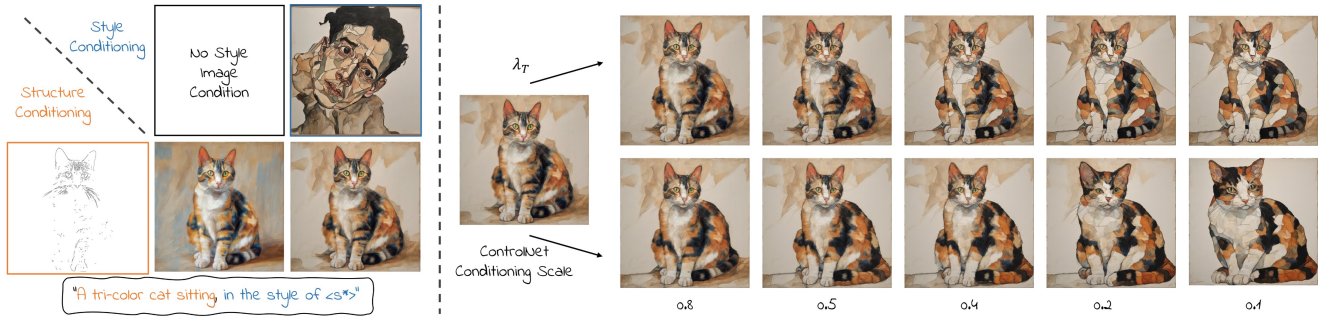


Figure 4. **Geometric Style Control**. *In our experiments we found that geometric style is often lost when using structure conditional maps, whether using style conditionals or not (left). By limiting the structure conditional inputs only to non-sensitive geometric style layers we achieve a geometric style interpolation which combines geometric style gradually, using a timestep limiting parameter λ_T. Note that this is not possible to achieve by changing the default ControlNet scale parameter. Our method keeps the structure dictated by the structure conditional while gradually injecting geometric style to the image. (Please zoom in to view images better.)*



Figure 5. **Collection Example.** *Collection examples used for structure sensitivity analysis. The object depicted in the collection is constant (cat), where each cluster holds a different style for style conditioning (top). At the bottom, we use "black and white ink" style to normalize color balance for structure conditioning.*

5.2. Structure Sensitivity

Conditioning structure in image generation is usually addressed by providing edge maps or similar maps with ControlNet [43]. Structure is also linked to aspects of geometric style of an artist's work, like contours and brush effects. Therefore, to analyze structure sensitivity we use Canny edge-maps conditioning. We select 10 artists with strong geometric styles and create five images for each artist, using a single object prompt for consistency. To further isolate geometric style from color, we apply an "Ink Drawing" style. A sample of the collection can be seen in Fig. 5, while the entire collection is presented in the supplemental file.

Applying our analysis method to this collection shows sensitivity in various Up layers at all timesteps. Reducing ControlNet input in these layers reveals geometric style changes as more timesteps are used. We conducted an ablation study (see supplemental file) to better understand these effects and discovered that the Up layers seem to handle the fine details from control maps, while Middle layers maintain overall structure alignment. Therefore, using our analysis we can rank the layers and timesteps according to structure sensitivity and apply conditioning (this time ControlNet input) only to a subset of the layers, allowing more balanced control with other conditions. Note that this kind of geometric style control cannot be achieved by simply reducing the scale parameter of ControlNet (see Fig. 4).

6. Results

Instead of choosing the number of layers K, we define two interpolation parameters: λ_S and λ_T, that control the percentage of layers used for style and structure respectively. This allows users to control different content-style balancing ratio, so we encourage tuning λ_S and λ_T. However, to be robust in our comparisons we set the layer subset size to 30 ($\lambda_S = 0.43$) for both text and content image conditioned generations and limit the content control over 850 timesteps ($\lambda_T = 0.15$) which induces geometric style and freedom. We find these values robust for various styles, including non-painting ones (see supplemental file.) We present results in Fig. 6 and in the supplemental file.

6.1. Evaluation Details

To evaluate our results we expand the evaluation set from Sec. 3. For style conditionals, we use 32 different styles varying in origin (Europe, North America, South America, Asia), material (Oil, Watercolor, Digital, etc...) and style (Realism, Impressionism, Expressionism, Cubism, Anime, Pixel Art, etc...). For content conditionals, we use 10 "Easy" and "Complex" prompts each, and condition them with "text only," Canny and Depth maps. For generating Depth maps we employ MiDaS [29] following [17]. We use 4 random seeds for generation of "text only" images, and two of these seeds for each image conditioning method, which uses 2 different content conditional images, one for each seed. Combining all this results in 5120 images for each method. We evaluate the style and content of each image according to the distance measures we defined in Sec. 3, and average across all images for a given method.



Figure 6. **Results.** *Example results generated by balanced-StyleAligned over various styles (top). Prompts used from top to bottom: "A robot wearing a fedora holding a flower," "A portrait of a woman," and "A family of panda bears wearing kimonos and sharing some tea."*

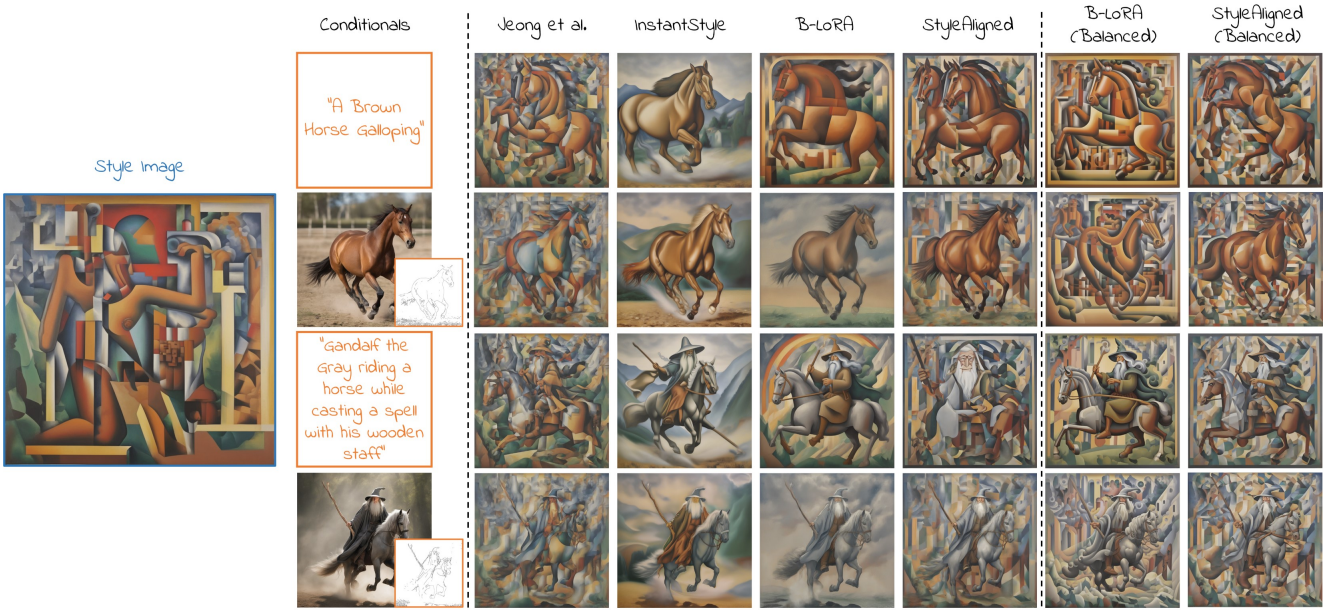


Figure 7. **Qualitative Comparison**. *Comparing various conditional combinations: "Easy" vs "Complex" prompts (rows 1,2 vs. 3,4), Text only vs. Text and structure conditioning (rows 1,3 vs. 2,4). As can be seen, both balanced methods achieve consistency over all conditioning combinations while imbalanced methods show inconsistent generation quality and in some examples content and style issues.*

6.2. Comparisons

We apply our balancing strategy to StyleAligned and B-LoRA and compare our method to four style-based approaches: StyleAligned and B-LoRA (without balancing), InstantStyle, and Jeong et al. (we leave their pipeline analysis and balancing for future work) . We also compare to other methods [8, 31] in the supplemental file. As B-LoRA pre-trains residual LoRA weights prior to inference, we replace the trained layers to the ones analyzed by our method. For fairness, we base our decision on the same number of self-attention layers (10) used by B-LoRA, although we find that in some scenarios using a larger number of layers improves results.

By definition, B-LoRA and InstantStyle are not intended to use a style prompt for generation, unlike the rest of the compared methods. For fairness, we evaluate B-LoRA and InstantStyle with style prompts like the other methods. B-LoRA shows a significant improvement in style with a slight decrease in content score. Since this minimally impacts its balance, we use it with style prompts for evaluation. However, for InstantStyle, the style improvement comes with a notable content reduction, significantly affecting balance, so we retain its original prompt for evaluation.

We report qualitative and quantitative results in Fig. 7 and Fig. 8, respectively. As demonstrated in Fig. 7 our layer balancing strategy preserving both style and structure (right columns) and is more consistent over various conditioning combinations. Meanwhile, other methods shows inconsistent output quality, where some combinations show satisfy-

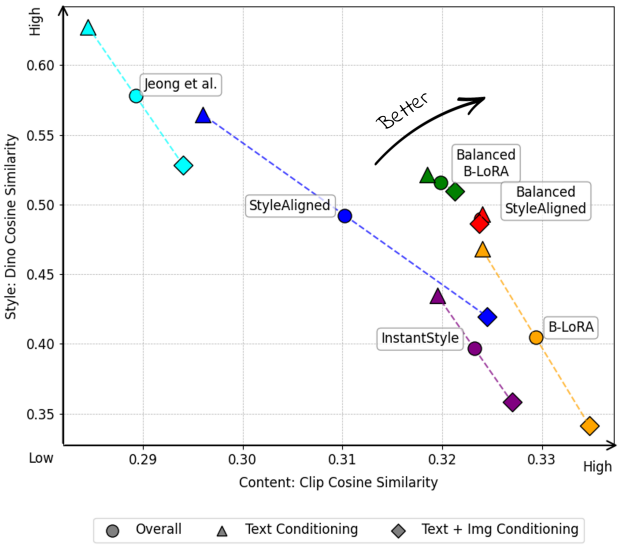


Figure 8. **Quantitative Comparison.** *We compare recent methods to the balanced versions of StyleAligned and B-LoRA. The plot shows the content and style tradeoff improvement (middle circle) as well as the balancing effect over content image conditionals (triangle vs. rhombus).*

Observed (Expected)	Balanced	Imbalanced	Total votes	χ^2_{stat}	$p_{val} <$
1. Multi-choice	386 (210)	244 (420)	630	35.1	0.001
2. A/B B-LoRA	195 (126)	57 (126)	252	30.0	0.001
3. A/B StyleAligned	185 (126)	67 (126)	252	21.9	0.001

Table 1. **User Study.** *Multi-choice 1/6 methods: 2 balanced vs. [12, 17, 20, 38] and two A/B tests: SA/B-LoRA vs. im/balanced.*

ing results while others show apparent issues such as general and geometric style loss, and content artifacts.

This balancing effect is also apparent quantitatively in Fig. 8. While other methods show varying content and style quality on different conditioning combinations (triangle vs. rhombus), balanced methods show consistent high scores in both style and content similarity. We further evaluate these results using an additional style-representation using GRAM matrices [15] (See supplemental file.)

6.3. User Study

We measure the balance effect on qualitative preference. Hypothesis: people find balanced methods better than imbalanced counterparts due to content/style conditioning and aesthetics. We evaluate our balanced methods, StyleAligned (SA) and B-LoRA, with three question setups:

1. A multi-choice comparison to all other imbalanced methods [12, 17, 20, 38] i.e., 6 options and 15 instances.
2. A/B Test: B-LoRA vs. balanced B-LoRA, 6 instances.
3. A/B Test: SA vs. balanced SA, 6 instances.

Instances were sampled at random while keeping equal conditioning representation. In both setups users were asked: *"which of the images below follows both conditions better: (1) shows content described in 'prompt', and (2) shows the style of 'style image'."* Examples can be found in the supplemental material. The study spanned over 1,134 evaluations by 42 anonymous participants. As detailed in Tab. 1, all experiments concluded with a significant preference for the balanced versions in a χ^2 test for independence.



Figure 9. **Style Transfer.** *A sample of style transfer using B-LoRA with an ungenerated style image (top row) and a content image (left column). Our balancing strategy improves style alignment and consistency for different content inputs.*

6.4. Additional Applications

We experiment with additional applications such as style transfer [6, 15, 19, 22], material-style generation [7], and flexibly-conditioned content editing [1, 3, 25]. We present style transfer results using balanced B-LoRA in Fig. 9. Additional results and implementation details for all applications can be found in the supplemental file.

7. Discussion and Limitations

In this paper we investigated the tradeoff between content and style conditionals and showed that over-conditioning prevents content and style alignment. We further present a novel method for analyzing sensitivity in attention layers of DDPMs by measuring the clustering score of the representation of a collection of images. We use this method to balance multi conditions by limiting conditioning on sensitive layers only, demonstrating improved results.

We found that the main limitations of our method arise from its dependence on the capabilities of the base model. For instance, when generating images with a style unfamiliar to the base model, the result may exhibit an unintended style, as the model lacks sufficient knowledge to properly generate the style reference, leading to a style mismatch. We demonstrate these limitations in the supplemental file.

Acknowledgment

This work was partly supported by Joint NSFC-ISF Research Grant no. 3077/23

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer, 2023. 8

[2] Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. Palp: Prompt aligned personalization of text-to-image models. *arXiv preprint arXiv:2401.06105*, 2024. 2

[3] Shariq Farooq Bhat, Niloy J. Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning, 2023. 8

[4] Tadeusz Caliński and Joachim Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3:1–27, 1974. 5

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 3, 1, 8

[6] Dar-Yen Chen, Hamish Tennent, and Ching-Wen Hsu. Artadapter: Text-to-image style transfer using multi-level style encoder and explicit adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8619–8628, June 2024. 2, 8

[7] Ta-Ying Cheng, Prafull Sharma, Andrew Markham, Niki Trigoni, and Varun Jampani. Zest: Zero-shot material transfer from a single image. *arXiv preprint arXiv:2404.06425*, 2024. 8

[8] Xing Cui, Zekun Li, Pei Pei Li, Huaibo Huang, Xuannan Liu, and Zhaofeng He. Instastyle: Inversion noise of a stylized image is secretly a style adviser. In *ECCV*, 2024. 2, 7

[9] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 2

[10] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 341–346, New York, NY, USA, 2001. Association for Computing Machinery. 2

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 3, 5, 6

[12] Yarden Frenkel, Yael Vinker, Ariel Shamir, and Daniel Cohen-Or. Implicit style-content separation using b-lora, 2024. 2, 3, 8, 1

[13] 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, 2022. 2

[14] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models, 2023. 2

[15] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 8

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[17] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *arXiv preprint arxiv:2312.02133*, 2023. 2, 3, 4, 5, 6, 8, 1

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2, 3

[19] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1510–1519, 2017. 2, 5, 8, 6

[20] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint arXiv:2402.12974*, 2024. 2, 3, 8

[21] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review, 2018. 2

[22] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution, 2016. 2, 8

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. 2

[24] Luping Liu, Chao Du, Tianyu Pang, Zehan Wang, Chongxuan Li, and Dong Xu. Improving long-text alignment for text-to-image diffusion models, 2024. 2

[25] Xiaoyu Liu, Yuxiang Wei, Ming Liu, Xianhui Lin, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Smartcontrol: Enhancing controlnet for handling rough visual conditions. *arXiv preprint arXiv:2404.06451*, 2024. 8

[26] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 2, 6

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 3, 1

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[29] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular

depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 6

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[31] Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Rb-modulation: Training-free personalization of diffusion models using stochastic optimal control. *arXiv preprint arXiv:2405.17401*, 2024. 2, 7

[32] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, June 2023. 2

[33] 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, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 2

[35] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 422–438, Cham, 2025. Springer Nature Switzerland. 2

[36] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, Yuan Hao, Irfan Essa, Michael Rubinstein, and Dilip Krishnan. Styledrop: Text-to-image generation in any style, 2023. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[38] Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024. 2, 3, 8

[39] Luozhou Wang, Guibao Shen, Wenhang Ge, Guangyong Chen, Yijun Li, and Ying cong Chen. Text-anchored score composition: Tackling condition misalignment in text-to-image diffusion models, 2024. 2

[40] Wikipedia contributors. Rouen cathedral (monet_series). https://en.wikipedia.org/wiki/Rouen_Cathedral_(Monet_series), 2024. [Online]. 1

[41] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, 2023. 2

[42] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024. 2

[43] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2, 6

[44] Yuxin Zhang, Weiming Dong, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Tong-Yee Lee, Oliver Deussen, and Changsheng Xu. Prospect: Prompt spectrum for attribute-aware personalization of diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):244:1–244:14, 2023. 2

[45] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2