Attribute-Conditioned Attention Scaling for Text-to-Image Diffusion Models

Rabia Tahir, Bilal Ahmed Memon
School of Software Engineering, Tashkent University of Information Technology (TUIT), Tashkent, Uzbekistan
School of Business and Economics, Westminster International University, Tashkent, Uzbekistan

***Abstract*—Many text-to-image diffusion models have attained significant success in the generation of images from textual prompts; however, they still face some challenges, such as limited fine-grained control over different semantic attributes. To overcome this issue, this study proposes an Attribute-Conditioned Attention Scaling (ACAS), which modulates the cross-attention layers of the UNet model using attribute-related scaling factors. These scaling factors are assigned to the attention maps that allow selective enhancement of features in the ACAS without retraining. Moreover, this model allows precise control over generated images without retraining of the base model at the inference level, along with preservation. For experiments, 30 diverse prompts along with eight descriptive attributes are used to inspect the multi-attribute controllability of the proposed model. Different evaluation metrics, such as CLIP, LPIPS, and Inception Score (IS), are used to quantitatively evaluate the proposed model. Experimental results prove that the proposed model ACAS obtains competitive results with an LPIPS of 0.75, CLIP score of 0.316, IS of 3.98, and a minimal 8.47 seconds generation time. Furthermore, a comparative analysis of the ACAS model with similar baseline methods is performed, and the comparison shows that the ACAS improves attribute controllability without adding extra computational cost. Overall, this model bridges the gap between fine-grained attribute control and prompt-based guidance in the latest diffusion models.**

***Keywords*—Diffusion models; attribute-control; CLIP; LPIPS; text-to-image generation; cross-attention modulation**

I. Introduction

In today's world, numerous people want to create images inspired by their visual imagination. With the advent of text-to-image models, it is possible to generate visually stunning images by just providing a prompt [5]. Diffusion models have emerged as a powerful method in the domain of text-to-image generation, and they are inspired by non-equilibrium thermodynamics. The rising interest in diffusion models proves their versatility and potential in various image-related tasks, especially text-to-image generation [1]. Latest models such as DALLE-2 [2], Stable Diffusion [10], and Imagen [4] prove the power to generate photorealistic and diverse images by conditioned-based prompts. Despite the huge success of these models, there are still a few challenges, such as limited fine-grained control on semantics attributes, e.g., "younger face", "metallic texture", and "futuristic city". Existing models are aimed at two main directions, i.e., fine-tuning the model (retraining) and embedding-level guidance. The fine-tuning or retraining of the model leads to more computational cost, while embedding-level guidance lacks explicit mechanisms to adjust the attributes' influence during the denoising task. The main motive of the proposed model is to modulate the influence of attributes on the attention layers with the help of a scalar weight. This allows suppression, increase, and balance of the attribute strength in the image generation. Further, the proposed model, ACAS, can deal with multiple attributes simultaneously with the help of normalized weights. Hence, the proposed model can integrate diverse attributes like ("younger", "hat", "smiling") in one-step generation. To address these issues, this study proposes a lightweight Attribute-Conditioned Attention Scaling (ACAS) for multi-attribute-based text-to-image generation in a well-controllable manner with no retraining overhead. During the inference stage, the proposed model injects attribute-specific scaling coefficients into the cross-attention blocks of the UNet model. The main motive is to allow all attributes to be balanced and enhanced while generation with the help of a scalar control parameter. The proposed model offers three significant benefits: fine-grained control, multi-attribute conditioning and no retraining overhead. Experiments are performed and a comparison is conducted for the proposed model with baseline models and the results show the superiority of the proposed model. The results are evaluated with both quantitative and qualitative metrics, including CLIP [15], LPIPS [16], Inception Score (IS) [17] and generation time. Overall, the contributions of this study are:

- A novel ACAS model is proposed for controllable attribute-guided text-to-image generation without retraining.
- This model injects the semantic attributes into the denoising to preserve the textual prompt semantics and to enhance the fidelity of attributes.
- Extensive experiments are performed with a comprehensive evaluation to show the efficiency of the proposed model.
- The proposed model, ACAS, bridges the gap between fine-tuning and prompt-level control. It offers a flexible and practical solution for real-time text-to-image generation.

The rest of the study is organized as follows: Section II describes the related work on controllable text-to-image generation and diffusion-based methods. Section III explains the proposed Attribute-Conditioned Attention Scaling (ACAS) model. Section IV illustrates the data, experimental setup, quantitative and qualitative results, discussion, and ablation study. Finally, Section V concludes the study and highlights some future research directions.

II. Related Work

Generative Adversarial Networks (GANs) have been used for image generation and they achieved strong performance on the benchmark datasets. However, these models even suffer from training instability and limited controllability limitations [3]. Diffusion-based text-to-image methods [6, 7, 8, 10] have shown significant progress in the past few years by generating highly realistic images with the help of a text prompt. Diffusion-based models avoid the hectic process of training a huge text-to-image model from the beginning and take benefit of already existing pre-trained models for attribute manipulation and image editing tasks [21]. Despite their huge success, some models [4, 9, 23] lead to high computational cost, and many of these models depend on denoising diffusion probabilistic models (DDPM).

Vanilla stable [10] diffusion framework proposed by Romach et al., produces images from textual prompts with the help of a latent diffusion manner. Instead of generating high-quality images, this model has low control over fine-grained attributes like environment, style type, and mood. Moreover, Prompt-to-prompt [11] and ControlNet [12] show improved guidance for generated images, but they still need extra control with specific attribute adjustment. Uni-ControlNet [18], the extended version of ControlNet, allows various control models in a unified framework simultaneously and leads to composite local and global controls. Furthermore, DynamicControl [19] improves adaptive selection and ordering of multiple control signals in order to handle various conditions. It improves the fine-grained attribute adjustment in distinct text-to-image generation tasks. These models either need additional input or are less efficient to control various attributes simultaneously. DiffEdit [14] and Attend-and-Excite [13] models perform attribute-level manipulation as well, but they require complicated pipelines for unseen attributes. Latest research explores the semantic and fine-grained control in a more significant manner. DilightNet [20] proposed fine-grained lightning control with the help of diffusion process guidance. This method produces images with respect to a given text prompt and lightning. By keeping these limitations in mind, Attribute-Conditioned Attention Scaling (ACAS) is designed to control multiple attributes over generated images in a precise way without retraining the base model. To check the efficiency of the proposed model, ACAS, state-of-the-art evaluation metrics are used, i.e., CLIP score [15], LPIPS [16], Inception score (IS) [17], and generation time. These metrics are used to evaluate the perceptual diversity, semantic alignment and generated content quality. Besides, the comparison of the proposed model, ACAS, with related models is performed, and the results show the superiority of the model in various factors.

III. Methodology

A. Overview

The proposed model, Attribute-Conditioned Attention Scaling (ACAS), improves the traditional text-to-image diffusion methods in order to obtain precise control over visual attributes in the produced image without retraining. As compared to the vanilla diffusion model that globally interrupts the prompt, this model applies a forward hook added at all cross-attention layers of the UNet without retraining. A prompt P and a set of descriptive attributes $A = \{a_1, a_2, a_3, \ldots, a_n\}$ are assigned to ACAS as an input. Then, the proposed model modulates the UNet cross-attention map M during the forward pass, as shown in Eq. (1) below:

$$\tilde{M} = \gamma\,(a_i).M \tag{1}$$

where, $\gamma(a_i)$ is the attention scaling factor for attributes a_i, M is UNet based cross attention map and $\tilde{M}$ is the modulated attention map that is used for forward pass after scaling. $\gamma(a_i)$ allows controllability of attribute influence without the need of retraining the base model. Fig. 1 shows the systematic view of the proposed model ACAS.

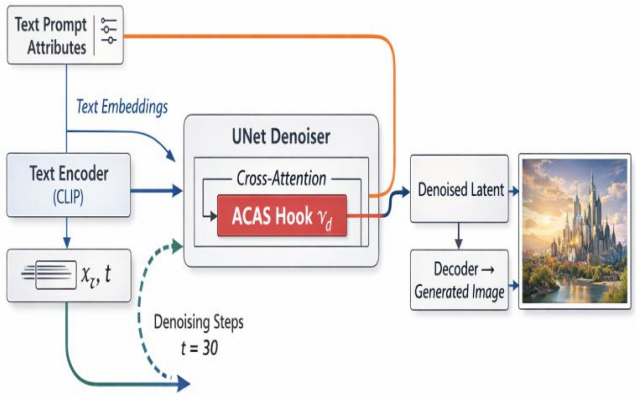

Fig. 1. Systematic view of the proposed model ACAS.

B. Cross-Attention in Traditional Diffusion Models

UNet in the traditional text-to-image diffusion models uses cross-attention to combine latent image features f_x with text features f_t, as explained in Eq. (2) below:

$$\text{Att(K, V, Q)} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{2}$$

where, K, V and Q are key, value and query metrics that are derived from text and image immersing.

C. Scaling by Attributes

To modulate attention weights, the proposed model ACAS presents a scalar $\gamma(a_i)$, as shown below in Eq. (3):

$$\widetilde{Att}\text{(K, V, Q; } a_i) = \gamma(a_i).\,\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3}$$

where, $\gamma(a_i) < 1$ subdues the features influence and $\gamma(a_i) > 1$ improves it. During denoising in UNet, it is imposed in each cross-attention block.

D. Multi-Attribute Scaling

In real-time, users give more than one attributes in a textual prompt to generate an image. For example, "an old man is smiling and wearing hat". Therefore, a single scalar is not enough to address this issue. The proposed model introduces a weighted multi-attribute formulation. Suppose, there is a set of N attributes $A = \{a_1, a_2, a_3, \ldots, a_n\}$, where a scalar weight is assigned to each attribute and this weight is independent. Hence, this multi-attribute attention is defined in Eq. (4):

$$\widetilde{Att}\text{(K, V, Q; } A) = \frac{1}{\sum_{i=1}^{N}\gamma(a_i).}\sum_{i=1}^{N}\gamma(a_i).\,\text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where, $\sum_{i=1}^{N} \gamma(a_i)$ is a normalization term that handles the magnitude of attention when more attributes are added. It allows the attributes to be scaled independently and balanced with weight selection.

E. Integration into UNet Denoising Process

Stable diffusion addresses the iterative refinement for a latent noise vector z_t during the denoising step. In the proposed ACAS, the modified attention $\widetilde{Att}(K, V, Q; A)$ is used at all cross-attention blocks of the base UNet model. This modifies attention and applies during all denoising step $t \in \{1, 2, \ldots, T\}$. Suppose, $\epsilon_\theta(z_t, t, P)$ shows a denoising UNet with condition oriented prompt P. Then, with ACAS, the conditioning prompt is shown in Eq. (5):

$$\epsilon_\theta(z_t, t, P, A) = \epsilon_\theta(z_t, t, P, \widetilde{Att}(K, V, Q; A)) \quad (5)$$

Afterward, the latent space becomes:

$$z_{t-1} = \text{Denoise}(z_t, \epsilon_\theta(z_t, t, P, A)) \quad (6)$$

This step in Eq. (6) has no training overhead and it does not change the architecture of UNet.

F. Weight Interpretation and Optimization

The attribute weightage $\gamma(a_i)$ gives a balanced control to influence the attributes. $\gamma(a_i)$ controls the degree of attention given to the attribute by the model in the cross attention map. Higher value of $\gamma(a_i)$ means a stronger influence, while low value of $\gamma(a_i)$ leads to lower influence. This degree of control over multiple attributes is without retraining the model. It is described below:

$\gamma(a_i)<1$ decrease the influence

$\gamma(a_i)>1$ increase the influence

$\gamma(a_i)=1$ same like vanilla diffusion model

Moreover, the proposed ACAS allows learnable weights for strong semantic alignment. $\gamma(a_i)$ is optimized with a CLIP-based loss, as shown in Eq. (7):

$$L_{CLIP} = 1\text{-}COS(f_T(P, A), f_I(A)) \quad (7)$$

where, $f_T(.)$ and $f_I(.)$ are CLIP-based text and image encoders and COS is used for cosine similarity. This loss function helps to generate an image which is aligned with both the attributes and prompt.

IV. Experiments

A. Data

For evaluation, 30 diverse prompts with eight attributes are used to generate 240 images. Fig. 2 shows some of these generated images with their prompt and attribute. For experiments, the prompt-based image generation is used as the base dataset. A total of 30 base prompts are defined that cover various visual domains such as scenes, human figures, animals and objects. For example; these textual prompts are "a futuristic city at sunset", "a surreal dreamscape". These prompts can be extended to different semantic categories that show the model's ability to create the diverse content with the help of multiple attribute as well. Against each prompt, eight semantic or stylistic attributes are used i.e. "low details", "high details", "mysterious", "cartoon", "hot and sunny", "raining", "realistic" and "oil painting". A strength coefficient (c) is assigned to each attribute that scales it in the attention layer of the UNet model while image generation. Table I shows the different value of weightage attribute γ that are assigned to all attributes. For example, $\gamma > 1$ is used for those attributes that require higher stylistic features, such as cartoon and high detail. For $\gamma < 1$, it implies that attribute details are not dominating the text prompt, such as "Low detail". The γ values are selected regarding the expected visual dominance of each attribute. It is ideal to start the initial estimate like 0.5 and then increase it or decrease it according to the obtained results in order to get the balance fine-grained control of attribute influence. This strength value gives fine-grained control over attributes. Table I shows the best γ values that yielded best in the experiments.

The dataset used for experiments in this work is designed to evaluate attribute-level controllability instead of large-scale image generation. Each attribute-prompt pair is constructed carefully in order to cover diverse semantic categories. These pairs cover diverse semantic categories, which are style, appearance and object-level characteristics. The applicability of the ACAS is not limited to a specific dataset because it works by modifying the cross-attention layer within a diffusion model. If the number of prompts increases, it would not alter the final results significantly. The consistent performance against diverse attribute-prompt pairs shows that the model ACAS generalizes across multiple visual concepts.

TABLE I. Strength Values Used in the Model Against Each Attribute.

Attributes	Strength (γ)
Low Detail	0.75
Realistic	1.25
High Detail	1.35
Cartoon	2
Hot and Sunny	1.15
Rainy	1.50
Mysterious	1.10
Oil Painting	1.05

B. Model Setup

For baseline method, Stable Diffusion v1-5 from Runway ML is used for image generation. Experiments are performed on Google Colab with T4 GPU and mixed precision (torch.float16) to optimize inference speed and memory. Mixed-precision helps in reducing memory without degrading the semantic quality. For comparison, all images are generated with fixed diffusion steps, i.e., 30 for ACAS, baseline and vanilla diffusion model. Moreover, CLIP (ViT-B/32) is used for quantitative evaluation of semantic alignment among the textual prompts and generated content. Table II shows details of parameters used in the proposed model, ACAS.

Contrastive Language Image Pre-training (CLIP) is a model proposed by OpenAI that is used to compare the images and text in the likewise semantic space. It calculates how the image matches with a text prompt. Vit-B/32 is a version of CLIP that

uses Vision Transformer as a backbone in the model. It deals with the images in patches of 32x32 pixels. So, this is used as an automatic metric to calculate the score for how well the output images matches the text description. A forward hook is implemented on all cross-attention layers of UNet model in order to modulate the influence of attributes on the generated image. With the help of attribute strength factor γ, the scaling of attention outputs can be controlled. Each text prompt is concatenated with the attribute, such as "a retro television set + high detail". After concatenation, pass it to the stable diffusion pipeline. The generation time is calculated for each image and stores the image in the corresponding directory.

TABLE II. PARAMETERS AND THEIR VALUES USED IN ACAS

Parameter	Value / Setting
Base model	stable-diffusion-v1-5
Diffusion Steps	30
Image Size	512 × 512
Device	Cuda (GPU)
Attention Hook	Custom Cross Attention hook
Mixed Precision	torch.autocast("cuda")
CLIP Model	ViT-B/32
LPIPS Model	Alex
Number of Prompts	30

C. Qualitative Results

This section shows the qualitative results of the proposed model. Fig. 2 shows the generated images by different textual prompts with eight different attributes. These attributes are “cartoon”, “high details”, “low details”, “mysterious”, “oil painting”, “photorealistic”, “Rainy” and “hot and sunny”. This figure shows results against four prompts which are “are "a futuristic city at sunset", “a tropical beach with palm trees", "a medieval castle on a hill" and “a cup of coffee”. In Fig. 3, the proposed model, ACAS, generates appealing images that describes both text prompt and attributes. For example, “a futuristic city at sunset” with cartoon attribute reflects an image that represents the futuristic city scape with cartoon aesthetic. This demonstrates the capability of ACAS to disentangle and recombine the attribute and textual information. Fig. 3 shows the visual comparison of the proposed model with the baseline stable diffusion v1.5 model [10] and the vanilla diffusion model v1.4 [10]. This clearly depicts the superiority of the proposed model over baseline models. The first row shows "a medieval castle on a hill" with the attribute “Oil painting”. The ACAS model generates a better image that satisfies both the text prompt and the attribute. The second row shows "a white dove flying” with the attribute “rain”. Vanilla and base model failed to generate a realistic image with a complete dove flying in rain, but ACAS generates an image that portrays the actual text prompt and attribute. The last row shows "a robot holding a flower" with attribute “hot and sunny”. The other two models do not satisfy the “hot and sunny” attribute and create an ordinary image of robot holding a flower. However, the proposed model generates a realistic image of a robot holding a flower in hand while the background shows the hot and sunny weather fulfilling the attribute condition. This is possible because the proposed model efficiently combines the attribute-guided attention with the text prompts and precisely controls the visual features of the generated image. This makes it a promising model for generating stylist images that reflects both textual prompts and attributes.

D. Quantitative Results

This section demonstrates the quantitative results for the proposed model, ACAS, and with related models stable diffusion v1.5 [10] and vanilla diffusion v1.4 [22]. State-of-the-art evaluation metrics CLIP [15], LPIPS [16], Inception score (IS) [17] are used along with the generated time for each image. A total of 30 experiments are performed for each model to generate 30 different images with respect to the attribute and text prompt and an average is calculated for all experiments. Table III summarizes these results for all three models. To obtain fair results, the same parameter values are used for all three models. The CLIP score is used to calculate the semantic consistency among generated images and text prompts. To calculate this score, OpenAI CLIP ViT-B/32 is used for calculating cosine similarity between image features and the encoded text. The second measure is LPIPS, which calculates the perceptual diversity among generated images with the help of an AlexNet Backbone. Pairwise LPIPS distances are measured for each attribute and then averaged. The Inception Score (IS) is used to inspect the quality and diversity of the generated images. The higher value of IS shows better results and is consistent to the textual prompt and attribute. The results in Table III show that ACAS obtained comparable results as compared to Stable Diffusion v.15 and vanilla diffusion v1.4. These results show that the proposed model, ACAS, preserves the overall quality of images without degradation. Furthermore, it is noted that these standard metrics are designed to compute the perceptual similarity, global image quality and distribution of realism. For fine-grained attribute control, they are not sensitive. Thereby, the lack of significant improvements in these standard metrics is not surprising and does not contradict the novelty and contribution of the model. The main motive of ACAS lies in allowing interpretable attribute-specific control by using attention scaling in the cross-attention layers instead of optimizing these standard metrics. Therefore, introducing controllable image generation without retraining with comparable performance to the strong baseline methods is a significant contribution. The generation time remains same for all models means that utilizing of attention hook in ACAS. Fig. 4 shows that the evolution of different metrics over time with respect to three models does not increase the computational cost additionally.

Fig. 2. This figure shows four different prompts with eight attributes. These prompts are "a futuristic city at sunset", "a tropical beach with palm trees", "a medieval castle on a hill" and "a cup of coffee". Attributes are given at the top of each column.

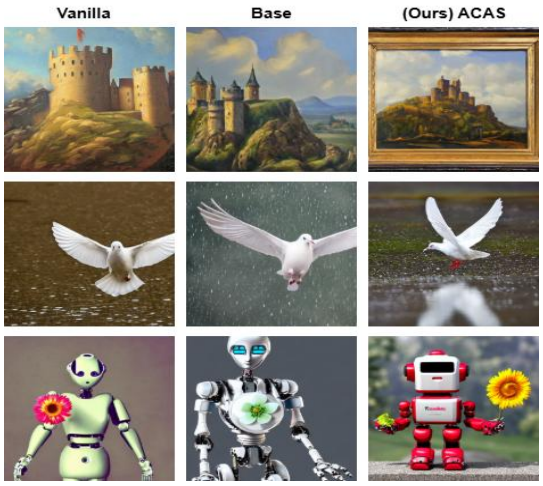

Fig. 3. Comparison of the ACAS with baseline models; first row shows "a medieval castle on a hill" + Oil painting, second row shows "a white dove flying" +rain, and the last row shows "a robot holding a flower" + hot and sunny.

TABLE III. SUMMARY OF CLIP, IS, LPIPS AND GENERATION TIME

Metrics	Vanilla V1.4	Stable Diffusion v1.5	ACAS (the proposed model)
CLIP	0.315	0.319	0.316
LPIPS	0.74	0.76	0.75
IS	3.94	4.06	3.98
Generated Time	8.4 sec	8.47 sec	8.47 sec

E. User Survey

A user survey is conducted to inspect the semantic alignment of the proposed model. Fig. 5 shows the summary of user survey to evaluate the performance of three models: ACAS, base model and vanilla diffusion. Five different textual prompts integrated with five attributes were used to generate images with all three models. Then, users were asked to choose the best image that matches both prompt and attribute. It is clearly evident in Fig. 5, that ACAS consistently obtained the highest score as compared to the other two models. The vanilla and base models were less selected and received fewer votes. This survey gives an overall insight and significance of the proposed model, ACAS.

F. Discussion and Ablation Study

The proposed model, ACAS, performance is discussed in this section. Three attribute-prompt pairs are used to inspect the effect of attribute strength gamma. These pairs are "a horse running in meadow", "a busy street with people" and "a waterfall in a dense forest", with three attributes "rainy", "mysterious" and "cartoon". Three different γ values, i.e., 1, 2 and 3 are used for the attribute strength.

The results show that when γ =1, the influence of the attribute is subtle and it preserves the main identity of the prompt. When γ increase to 2, then the visibility of attribute is more visible. When γ =3, then the attribute becomes more dominant and its influence is more visible on the generated images. However, increasing γ value more than 3 may lead to generate unrealistic images. As shown in Fig. 5, when γ is 3 then there are artifacts on the horse image in contrast to gamma value 2 that generates a balanced images. This observation shows that ACAS allows gradual and interpretable control over the attribute strength. In contrast to baseline methods that modify latent representation or need fine-tuning, the proposed model ACAS maintains the image quality while preserving the original semantics.

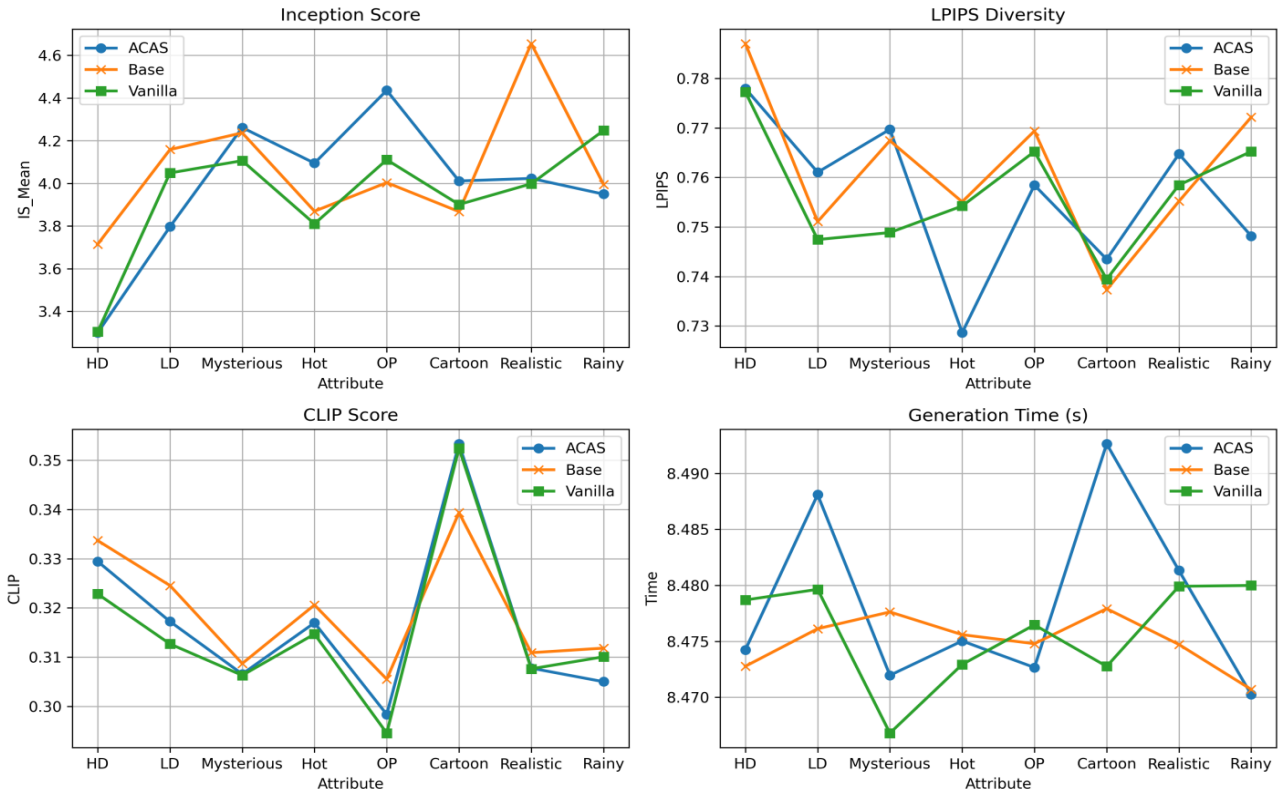

Fig. 4. Evolution of CLIP, LPIPS, and IS over time, along with generation time for each image (ACAS: the proposed model, Base: Stable diffusion v1-5, Vanilla).

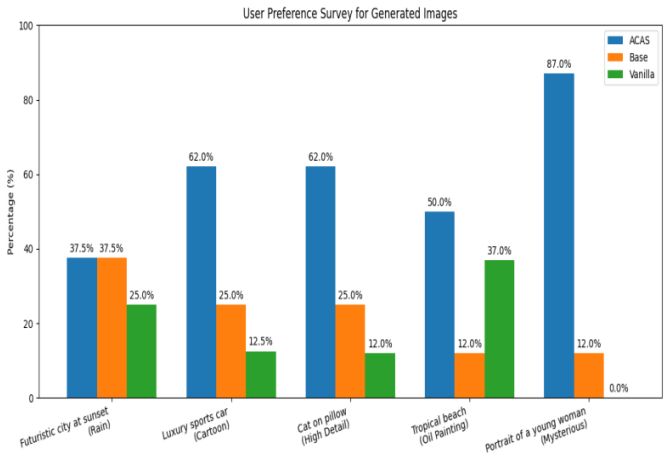

Fig. 5. User survey for ACAS, base model and vanilla model on generated images.

Moreover, this analysis shows that the standard quantities metric values are comparable with baseline models because the global image distribution is mainly unchanged during adjustment of attribute-specific features. In a nutshell, the ablation study shows that the ACAS provides interpretable and stable control over attributes. Furthermore, this discussion shows that the proposed model ACAS obtains controllability over attribute along with comparable performance to the similar existing models. In this section, the significance of attribute strength γ in the proposed model, ACAS, is briefly discussed. For this purpose, an ablation study is performed, focusing on attribute strength γ on the resultant images.

Fig. 6. Ablation study with different attribute strength γ ={1, 2, 3}. The first column shows generated images with $\gamma = 1$, the second column shows $\gamma = 2$ and the last column shows $\gamma = 3$.

Fig. 6 shows three different textual prompts with three different attributes. These prompts are "a horse running meadow", "a waterfall in a dense forest" and "a busy street with people" while the attributes are "rainy", "cartoon" and "mysterious". For each prompt-attribute pair, the attributes strength values are set to 1, 2 and 3. The generated images are obtained by just changing these values without any extra modification in the model. Fig. 6 shows that when γ =1, the effect of that specific attribute is subtle and not very dominant. When γ increases to 2 then the attribute influence becomes more

visible and prominent. At γ =3, the generated image strongly influences the textual prompt by intensified attribute details such as rain or cartoon textures. However, a higher value of γ does not guarantee the best results always. For example, when γ =3 then the image generated against the prompt "a waterfall in a dense forest" looks unrealistic and best results can be seen at γ =2. Therefore, it is concluded that ACAS allows interpretable control on various attributes without modifying the baseline model.

V. Conclusion and Future Work

This study presented a lightweight Attribute-Conditioned Attention Scaling (ACAS) to generate images with textual prompts and attributes. The selective enhancement of the features is made possible by applying the cross-attention layers of the UNet model along with attribute scaling factors. This enhancement of selective features against each attribute is achieved without retraining the base model. A total of four quantitative metrics, semantic alignment (CLIP), LPIPS, Inception Score (IS) and generation time are used. Results show that the descriptive attributes play a great role in generating behavior. Overall, the proposed model is highly responsive to attribute-based prompts and realism. In a nutshell, ACAS bridges the gap between textual prompt guidance and fine-grained attributes and presents a flexible model for controllable image generation with diffusion models. In the future, the aim is to extend this work by handling more complex and multiple attribute conditions and high-resolution image generation. Moreover, larger user studies and the use of more robust metrics could validate and enhance the performance of the model.

Acknowledgment

The authors would like to acknowledge the Tashkent University of Information Technologies (TUIT) and the Westminster International University in Tashkent, Uzbekistan for providing academic support and research facilities. Their institutional resources and environment contributed to the successful completion of this work.

References

[1] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, and S. Chen, "Diffusion model-based image editing: A survey," IEEE Trans. Pattern Anal. Mach. Intell., 2025.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," arXiv preprint arXiv:2204.06125, 2022.

[3] K. Cheng, R. Tahir, M. Li, and Z. Wang, "An analysis of generative adversarial networks and variants for image synthesis on MNIST dataset," Multimedia Tools and Applications, vol. 79, no. 19, pp. 13725–13752, 2020.

[4] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," Adv. Neural Inf. Process. Syst., vol. 35, pp. 36479–36494, 2022.

[5] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 3836–3847, 2023.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Adv. Neural Inf. Process. Syst., vol. 33, pp. 6840–6851, 2020.

[7] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," Adv. Neural Inf. Process. Syst., vol. 34, pp. 8780–8794, 2021.

[8] Y. Li, H. Wang, Q. Jin, J. Hu, P. Chemerys, Y. Fu, and J. Ren, "SnapFusion: Text-to-image diffusion model on mobile devices within two seconds," Adv. Neural Inf. Process. Syst., vol. 36, pp. 20662–20678, 2023.

[9] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, and M. Liu, "eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers," arXiv preprint arXiv:2211.01324, 2022.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 10684–10695, 2022.

[11] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," arXiv preprint arXiv:2208.01626, 2022.

[12] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 3836–3847, 2023.

[13] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-Excite: Attention-based semantic guidance for text-to-image diffusion models," ACM Trans. Graph., 2023.

[14] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, "DiffEdit: Diffusion-based semantic image editing with mask guidance," arXiv preprint arXiv:2210.11427, 2022.

[15] A. Radford et al., "Learning transferable visual models from natural language supervision," arXiv preprint arXiv:2103.00020, 2021.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 586–595, 2018.

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training generative adversarial networks," Adv. Neural Inf. Process. Syst., vol. 29, pp. 2234–2242, 2016.

[18] S. Zhao et al., "Uni-ControlNet: All-in-one control to text-to-image diffusion models,"Adv. Neural Inf. Process. Syst., vol. 36, pp. 11127–11150, 2023.

[19] Q. He et al., "DynamicControl: Adaptive condition selection for improved text-to-image generation," arXiv preprint arXiv:2412.03255, 2024.

[20] C. Zeng et al., "DiLightNet: Fine-grained lighting control for diffusion-based image generation," ACM SIGGRAPH Conf. Papers, pp. 1–12, 2024.

[21] Z. Zhang et al., "SINE: Single image editing with text-to-image diffusion models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6027–6037, 2023.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proc. CVPR, 2022.

[23] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 4195–4205, 2023.