DynamicImportance: Interactive Assignment of Word Significance in a Generalist Vision Model

Meri Topuzyan[1], Levon Khachatryan[1,2]
[1]Yerevan State University (YSU) [2]Picsart AI Resarch (PAIR)

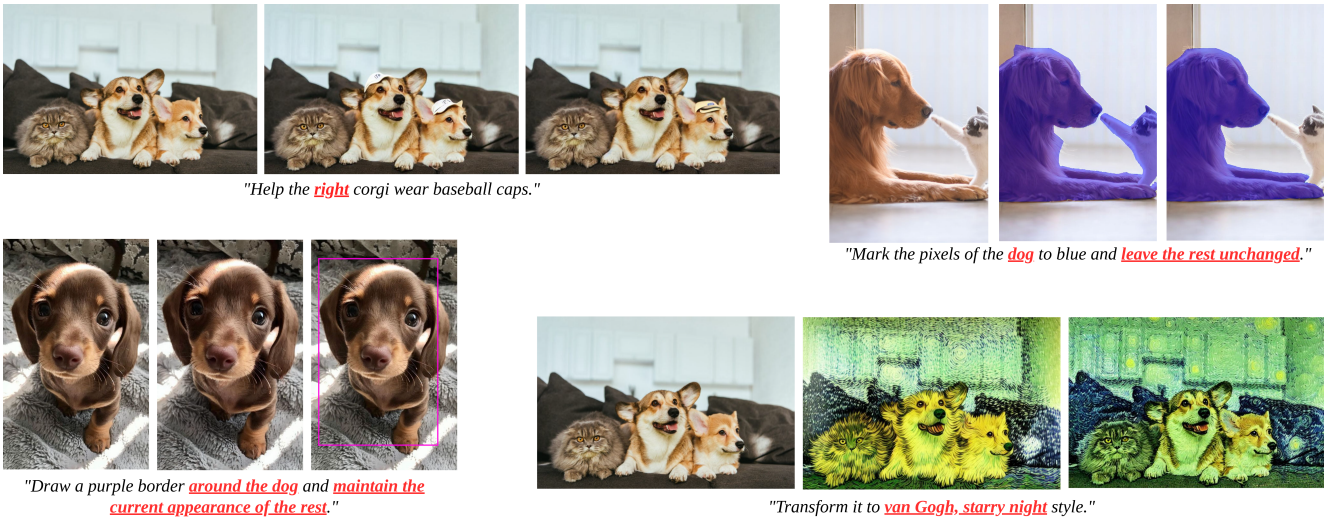


Figure 1. Comparative Visualization of Image Editing Techniques.

Abstract

Recent advancements in Deep Learning have markedly expanded the potential for creating a generic model for various vision tasks. At the forefront, InstructDiffusion presented a unified framework for aligning various computer vision tasks with human instructions (textual prompts). The model can handle a variety of vision tasks, including understanding tasks (such as segmentation and keypoint detection) and generative tasks (such as editing and enhancement). It is built upon the diffusion process and is trained to predict pixels according to user instructions, such as "encircling the man's left shoulder in red." (keypoint detection) or "applying a blue mask to the left car." (segmentation). However, not all words in the given textual prompt hold equal importance. Certain indicator words (e.g. "left shoulder" in the previous example) require more emphasis than others. To address this limitation, we propose DynamicImportance, a post-hoc method inspired by Attend-and-Excite. This technique conducts On-the-Fly optimization to direct the model in refining the cross-attention units based on the specified weights assigned to each word. The mechanism selectively enhances relevant features and suppresses less informative ones, thereby refining the model's focus during the image editing process. By dynamically adjusting to the content specified in text instructions, it allows for more nuanced and context-aware modifications of images. This strategic enhancement in attention processing not only improves the model's interpretability of textual instructions but also significantly boosts the precision and coherence of the resultant image edits. Our novel attention-augmented InstructDiffusion model, therefore, represents a symbiosis of interpretative and generative capabilities, paving the way for more intuitive and human-like editing performances in AI-driven systems.

1. Introduction

The convergence of natural language processing and computer vision in text-driven image manipulation signifies a major evolution in artificial intelligence technologies.

This synergy facilitates the direct transformation of textual instructions into visual changes, optimizing the interaction between human inputs and automated outputs. InstructDiffusion [[6]], an innovation by Microsoft Research Asia revolutionizes image manipulation by unifying editing, segmentation, and keypoint detection under a single instruction-driven framework. Built upon the diffusion process [[7]], it translates textual instructions into precise pixel modifications, enhancing the naturalness and intuitiveness of human-computer interactions. We propose a new method, DynamicImportance, which aims to enhance the attention layers within the InstructDiffusion model, leveraging its full potential across a broad spectrum of computer vision tasks. We aim to enhance the model's contextual processing to improve accuracy in executing text-based commands across diverse scenarios. The refinement enhances basic vector-based attention with the introduction of Attend-and-Excite mechanisms [[1]], which dynamically adjust the model's focus on relevant image areas based on textual cues. Advancing the attention mechanisms in InstructDiffusion is expected to significantly improve the model's performance, facilitating a seamless and dynamic interaction between textual instructions and visual content.

Our research enriches InstructDiffusion's ability to process and respond to complex editing instructions, narrowing the gap between high-level intentions and pixel-level manipulations. By doing so, we aim to refine the paradigms of text-driven image manipulation, enhancing the model's capacity to identify and emphasize critical visual elements accurately, pushing forward towards a more integrated and intelligent system in AI-driven computer vision.

2. Related Work

The landscape of text-conditional image editing is characterized by a diverse array of methodologies and models, each contributing to the evolution of this interdisciplinary field. In addition to InstructDiffusion [[6]], which stands as a prominent example of text-guided image manipulation, other models have made significant strides in leveraging natural language descriptions for visual transformations. One such model is CLIP (Contrastive Language-Image Pre-training) [[8]], which represents a breakthrough in pre-training large-scale vision-language models on extensive datasets. By learning rich representations of both textual and visual data, CLIP enables seamless understanding and manipulation of images based on textual prompts. Its success underscores the potential of text-conditional image editing in facilitating intuitive interactions between users and visual content.

Similarly, DALL-E [[9]] introduces a transformer-based [[14]] architecture specifically designed for text-to-image synthesis, showcasing the power of generative models in bridging the semantic gap between textual descriptions and visual representations. With its ability to generate diverse and contextually relevant images from textual input, DALL-E demonstrates the feasibility of fine-grained control over image synthesis processes. These models collectively highlight the growing interest and investment in developing sophisticated techniques for text-guided image manipulation, driving forward the capabilities and applications of this emerging field.

Furthermore, diffusion-based approaches such as Guided Diffusion [[2]] and Taming Transformers [[5]] have emerged as a compelling paradigm for text-conditional image editing, offering a principled framework for generating high-quality images while accommodating textual guidance. For example, Taming Transformers leverages diffusion models to synthesize images conditioned on textual descriptions, showcasing the potential of diffusion-based methods in achieving accurate and contextually coherent image edits. This approach underscores the importance of probabilistic modeling in capturing the uncertainty inherent in image synthesis tasks, thereby enabling more faithful adherence to textual instructions.

Moreover, recent advancements in self-attention mechanisms, exemplified by models like Vision Transformer (ViT) [[3]], have demonstrated the efficacy of attention-based architectures in various vision tasks, including image classification and object detection. By allowing models to dynamically attend to relevant features within the input data, self-attention mechanisms enhance the model's ability to capture long-range dependencies and contextual information, thereby improving performance across a range of visual tasks. Integrating such attention mechanisms into text-conditional image editing frameworks holds the promise of further enhancing the interpretability and fidelity of generated images, facilitating more nuanced and contextually relevant edits.

The broader research scope includes novel techniques like Spatial Transformer Networks that perform geometric transformations on images based on textual commands, and methodologies like Attention-Guided Image Generation that prioritize relevant areas during image synthesis. These innovations underscore the deepening integration of natural language and visual content, facilitating sophisticated visual media manipulation.

3. Preliminaries

3.1. Foundations of Diffusion Models

Diffusion models [[7]] represent a significant breakthrough in generative machine learning, particularly in the realm of image synthesis. These models are grounded in a concept inspired by the physical process of diffusion, where they initially introduce randomness into structured data and then methodically reverse this process. This reversal is not

a mere act of noise cancellation but a sophisticated reconstruction phase where the model learns to unveil the original data from its noise-encumbered state. At the heart of diffusion models is a step-by-step procedure that incrementally adds Gaussian noise to an image over a series of stages, progressively moving the data towards a completely noisy distribution. This gradual process is key to the model's ability to synthesize images, as it learns the complex statistical relationships between different stages of noise addition and reduction. In image synthesis, this equates to a model that can take a completely random noise pattern and, through a learned reverse process, reconstruct an image that exhibits high fidelity to the original dataset's characteristics. The operational mechanics of diffusion models involve a delicate balance between noise addition and image reconstruction. This balance is maintained through a training regimen that leverages a large corpus of images, enabling the model to effectively 'understand' and replicate the underlying distribution of real-world visual data. Through this training, diffusion models become adept at generating images that are not just visually compelling but also rich in detail and variation, mirroring the diversity found in natural visual scenes. One of the pivotal strengths of diffusion models lies in their capacity for controlled generation. Unlike some generative models that might produce arbitrary or unsolicited results, diffusion models can be conditioned to generate images that adhere to specific guidelines or frameworks, including textual descriptions. This attribute is particularly useful in tasks where the generation needs to be aligned with certain semantic or stylistic criteria, making diffusion models versatile tools for a wide array of applications in art, design, and media production. Moreover, the iterative denoising process intrinsic to diffusion models offers a unique advantage: it facilitates a fine-grained analysis of the generation process at each step, providing insights into how the model perceives and constructs visual information. This not only serves as a fascinating study into the model's operational dynamics but also allows for iterative improvements and optimizations in the model's architecture and training methodology. In summary, diffusion models have emerged as a formidable class of generative models, known for their robustness, versatility, and high-quality output. Their development and refinement continue to be a major focus in the field of artificial intelligence, reflecting their crucial role in advancing the capabilities of generative technology and their potential to shape the future landscape of digital image creation.

3.2. Stable Diffusion

Stable Diffusion [[10]] is a diffusion model operating in the latent space of an autoencoder $\mathcal{D}(\mathcal{E}(\cdot))$, namely VQ-GAN [[4]] or VQ-VAE [[13]], where $\mathcal{E}$ and $\mathcal{D}$ are the corresponding encoder and decoder, respectively. More precisely if $x_0 \in \mathbb{R}^{h\times w\times c}$ is the latent tensor of an input image Im to the autoencoder, i.e. $x_0 = \mathcal{E}(Im)$, diffusion forward process iteratively adds Gaussian noise to the signal x_0:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),\ t = 1, .., T \quad (1)$$

where $q(x_t|x_{t-1})$ is the conditional density of x_t given x_{t-1}, and $\{\beta_t\}_{t=1}^T$ are hyperparameters. T is chosen to be as large that the forward process completely destroys the initial signal x_0 resulting in $x_T \sim \mathcal{N}(0, I)$. The goal of SD is then to learn a backward process

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

for $t = T, \ldots, 1$, which allows to generate a valid signal x_0 from the standard Gaussian noise x_T. To get the final image generated from x_T it remains to pass x_0 to the decoder of the initially chosen autoencoder: $Im = \mathcal{D}(x_0)$.

After learning the abovementioned backward diffusion process DDPM [[7]] one can apply a deterministic sampling process, called DDIM [[12]]:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(x_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta^t(x_t), \quad t = T, \ldots, 1, \quad (3)$$

where $\alpha_t = \prod_{i=1}^t (1-\beta_i)$ and

$$\epsilon_\theta^t(x_t) = \frac{\sqrt{1-\alpha_t}}{\beta_t}x_t + \frac{(1-\beta_t)(1-\alpha_t)}{\beta_t}\mu_\theta(x_t, t). \quad (4)$$

To get a text-to-image synthesis framework, SD guides the diffusion processes with a textual prompt τ. Particularly for DDIM sampling, we get:

$$x_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(x_t, \tau)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\epsilon_\theta^t(x_t, \tau), \quad t = T, \ldots, 1. \quad (5)$$

It is worth noting that in SD, the function $\epsilon_\theta^t(x_t, \tau)$ is modeled as a neural network with a UNet-like [[11]] architecture composed of convolutional and (self- and cross-) attentional blocks. x_T is called the latent code of the signal x_0 and there is a method [[2]] to apply a deterministic forward process to reconstruct the latent code x_T given a signal x_0. This method is known as DDIM inversion. Sometimes for simplicity, we will call $x_t, t = 1, \ldots, T$ also the *latent codes* of the initial signal x_0.

3.3. Attend-and-Excite

Attend-and-Excite [[1]] represents a transformative approach to refining the image synthesis process in diffusion models, specifically designed to enhance the fidelity and accuracy of text-driven image generation. This mechanism

modulates the attention distributions dynamically, ensuring a more semantically faithful translation of textual instructions into visual outputs. Integrating Attend-and-Excite into the InstructDiffusion framework addresses challenges typically associated with subject token neglect—where specific descriptive elements of the text are underrepresented or omitted in the generated images.

Mechanism Overview: Attend-and-Excite dynamically adjusts attention weights during the model's denoising phase. By recalibrating these weights, it ensures that each subject token from the text prompt exerts significant influence over corresponding image patches, essentially guiding the synthesis process to better represent all semantic tokens. This process involves manipulating the noised latent code at each timestep, directing it toward a more semantically accurate generation.

Mathematical Formulation: The method involves a generative semantic nursing process where the noised latent code at each timestep t is adjusted based on the gradient of a loss objective designed to maximize the attention values for each subject token. The transformation applied is defined by:

$$\mathbf{z}_t' = \mathbf{z}_t - \alpha_t \cdot \nabla_{\mathbf{z}_t} L \tag{6}$$

where α_t is a scaling factor and L represents the loss, calculated to enhance the representation of neglected tokens.

Implementation Details: The Attend-and-Excite mechanism does not require changes to the foundational architecture of the pre-trained diffusion model but introduces adaptive adjustments on the fly during inference. This ensures that the attention maps are recalibrated in real-time, enhancing the accuracy of image synthesis based on textual descriptions.

Empirical Validation: In practice, the integration of Attend-and-Excite has demonstrated significant improvements in the model's ability to adhere to complex textual descriptions, particularly in scenarios involving multiple subjects or detailed scenes. The method effectively reduces instances of subject token neglect, ensuring that each element of the text is accurately and vividly represented in the generated images. Beyond its primary function of addressing token neglect, Attend-and-Excite also enhances the overall coherence and contextual relevance of the generated images. By ensuring that all textual elements are considered during the image synthesis process, the model can produce outputs that are not only more visually appealing but also semantically consistent with the input instructions.

The Attend-and-Excite approach thus marks a significant advancement in the field of AI-driven image editing, particularly within the framework of diffusion models like InstructDiffusion. By enhancing the interaction between textual inputs and visual outputs, Attend-and-Excite contributes to the broader goal of creating more intelligent, responsive, and capable image synthesis systems. This method not only enriches the model's capacity for detailed and accurate image generation but also sets a new standard for the integration of advanced attention mechanisms in generative models.

3.4. InstructDiffusion: A Pioneering Model

InstructDiffusion [[6]] marks a notable leap forward in the field of text-driven image editing, uniquely combining advanced diffusion model techniques with a nuanced understanding of textual instructions. This model innovatively interprets all computer vision tasks through the lens of image generation, emphasizing three primary output formats: 3-channel RGB images, binary masks, and keypoints. This strategic focus allows for a diverse application range, from basic image enhancements to complex scene reconstructions.

Text-Guided Image Synthesis: InstructDiffusion harnesses textual descriptions to meticulously steer the image synthesis process. This ensures that the resultant images are not only visually appealing but also align precisely with the intents specified in the text. It utilizes a sophisticated denoising process that methodically integrates textual cues at each stage, thereby facilitating the creation of images that are true to the provided instructions. This function is critical as it enables the model to implement wide-ranging modifications, from nuanced color adjustments to complete compositional transformations based on textual input.

$$\mathbf{I}_{\text{final}} = \text{Denoise}(\mathbf{I}_{\text{noisy}}, \mathbf{C}; \theta) \tag{7}$$

Here, $\mathbf{I}_{\text{final}}$ is the final image, $\mathbf{I}_{\text{noisy}}$ is the intermediate image with noise, $\mathbf{C}$ represents the text commands, and θ denotes the model parameters.

Comprehensive Training and Versatility: InstructDiffusion has been trained on a diverse dataset of image-text pairs, enabling it to master a wide array of editing tasks. This extensive training allows the model to accurately translate different textual descriptions into corresponding visual changes, providing the flexibility to handle various image modification requests effectively. The robust training ensures that the model can be used for aesthetic enhancements or specific modifications according to user needs.

Recognizing Limitations and Potential for Improvement: While InstructDiffusion is a powerful tool, its success in text-to-image translation depends on the effectiveness of its attention mechanisms in processing and integrating textual instructions. Although the model performs well overall, there is potential for improvement in capturing more detailed and nuanced aspects of the instructions. Enhancing these attention processes would result in more accurate and contextually appropriate image edits, thus improving the model's performance.

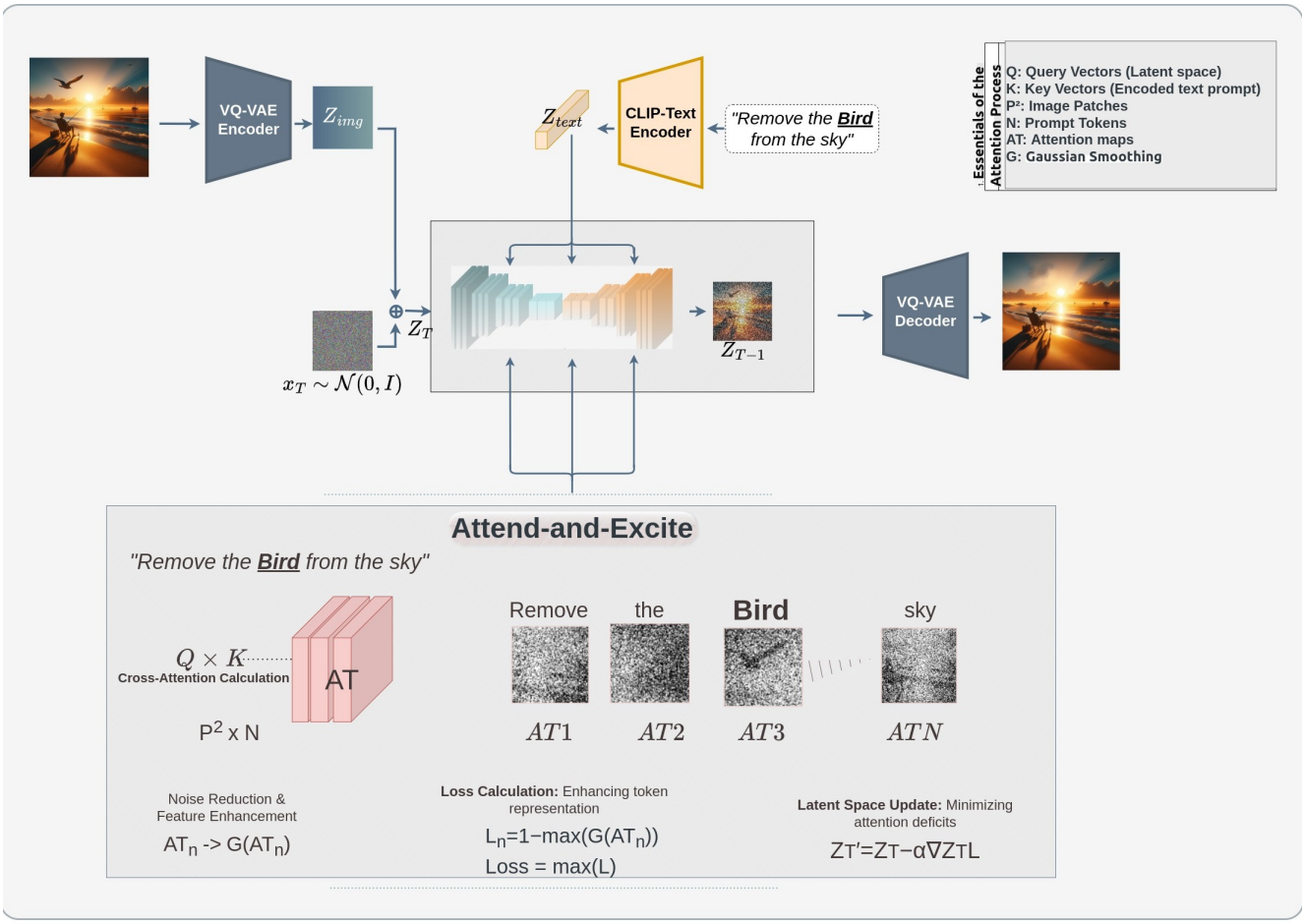


Figure 2. Schematic Overview of the Model Architecture.

4. Method: Integrating Attend-and-Excite into InstructDiffusion

In the evolving landscape of text-driven image editing, our research introduces a groundbreaking enhancement to the InstructDiffusion [[6]] model by embedding the Attend-and-Excite [[1]] mechanism, tailored to drastically improve the translation of complex textual instructions into highly precise and contextually accurate visual outputs. This enhancement is crucial for advanced image editing applications, where the fidelity of textual descriptors to visual representations must be maintained at a high standard. Building upon this framework, we introduce DynamicImportance, a novel method that allows for the dynamic adjustment of attention weights at the inference stage. This method enables the model to assign variable importance to different tokens based on their relevance and impact on the final image output. By dynamically prioritizing crucial elements of the text, DynamicImportance ensures that the generated images adhere more closely to the nuanced specifics of the input instructions. The Attend-and-Excite integration directly addresses the often encountered issue of subject token neglect—where crucial textual elements are not sufficiently visualized in the generated images. By refining the attention distribution within the InstructDiffusion architecture, DynamicImportance ensures an enriched visual correspondence to textual content, thereby more accurately bridging the divide between textual intentions and their visual outcomes.

4.1. Model Architecture

Image Encoding and Latent Space Formulation: The proposed framework begins by converting the input image I into a high-dimensional latent representation. This transformation is accomplished through a Vector Quantized-Variational AutoEncoder (VQ-VAE)[[13]]. The VQ-VAE is specifically chosen for its efficiency in compressing image data into a compact latent space. This approach ensures that the model can effectively handle high-resolution images by reducing their dimensionality, making subsequent processing steps more computationally manageable while retaining

essential features needed for accurate image reconstruction:

$$Z_{\text{img}} = \text{EncoderVQ-VAE}(I) \tag{8}$$

After encoding, the model introduces controlled randomness into the latent space by concatenating Z_{img} with Gaussian noise x_T. It is designed to enhance the model's generalization capabilities and prevent overfitting by providing a robust representation that encapsulates both the deterministic aspects of the image and stochastic elements:

$$Z = [Z_{\text{img}}; x_T], \quad x_T \sim \mathcal{N}(0, I) \tag{9}$$

Textual Instruction Encoding: Concurrently with image encoding, the textual instruction undergoes processing to extract its semantic essence. The CLIP-Text encoder[[8]] is employed to convert the text "Remove the Bird from the sky" into a semantic latent vector Z_{text}. This encoder leverages a large-scale pre-trained model that understands a wide range of human languages and their corresponding visual representations, enabling it to provide a rich semantic embedding. By using CLIP-Text, the model benefits from its ability to correlate textual and visual information effectively, ensuring that the textual instructions are accurately translated into a form that the image processing components can utilize for precise modifications:

$$Z_{\text{text}} = \text{Encoder}_{\text{CLIP-Text}}(\text{"Remove the Bird from the sky"}) \tag{10}$$

Attend-and-Excite Mechanism: At the core of our model is the Attend-and-Excite module, which orchestrates the alignment of the image content with the semantic directives encoded in Z_{text}. This module uses a sophisticated cross-attention mechanism to focus the model's "attention" on specific parts of the image that relate to the textual instructions. The cross-attention mechanism operates by dynamically adjusting the focus on image regions based on the relevance to the textual input, ensuring that the areas requiring modification are prioritized in the image editing process:

$$A = \text{softmax}\left(\frac{Q(Z)K(Z_{\text{text}})^T}{\sqrt{d_k}}\right) \tag{11}$$

Here, Q and K are learned functions that project the latent representations into a shared attention space, facilitating the selective enhancement of relevant image features. This projection into a common space allows the model to compute the attention weights that determine the influence of each part of the image based on the textual context, making the editing process more focused and accurate:

Latent Space Update and Image Reconstruction: To ensure that the final image reflects the modifications specified by the textual instructions, the model iteratively updates the latent representation Z. This process minimizes the attention deficits through a gradient-based optimization approach, refining the image features in line with the semantic goals:

$$Z_{T'} = Z_T - \alpha \nabla_{Z_T} L \tag{12}$$

where L is a carefully designed loss function that maximizes the impact of focused attention, thereby ensuring precise manipulation according to the textual description. This loss function is structured to penalize deviations from the intended modifications, thus driving the optimization process towards achieving the exact visual outcome described by the text:

$$L = 1 - \max(G(AT_n)) \tag{13}$$

Finally, the updated latent space Z_T is decoded using the VQ-VAE decoder to reconstruct the image. The decoding step translates the refined latent representations back into pixel space, producing the final image I with the desired edits incorporated accurately:

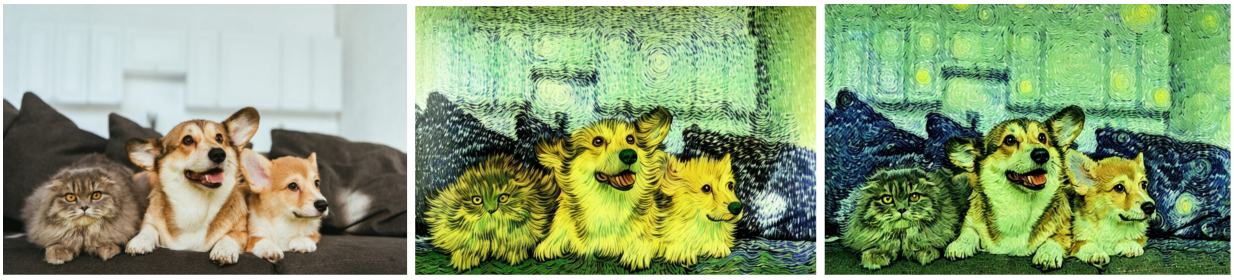
$$I' = \text{DecoderVQ-VAE}(Z_{T'}) \tag{14}$$

By integrating the Attend-and-Excite module into InstructDiffusion, the model is empowered to perform a wide range of complex computer vision tasks. These include intricate image editing, such as adding, removing, or modifying elements within an image based on user-highlighted instructions, as well as segmentation, keypoint detection, and other tasks with enhanced precision. This methodology is anticipated to demonstrate improved performance on benchmarks where the alignment between the instruction and the output image is crucial. DynamicImportance further enhances the InstructDiffusion framework, enabling sophisticated interaction between textual instructions and image content, thereby advancing the capabilities of AI-driven image manipulation and other computer vision technologies.

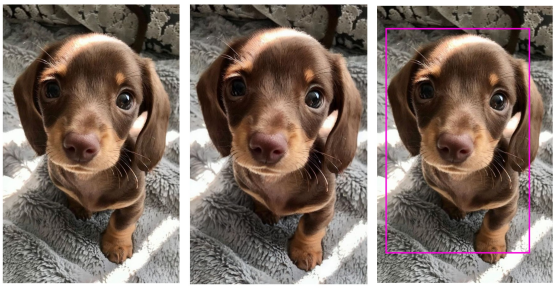
5. Experiments

This section evaluates the performance of the InstructDiffusion and DynamicImportance models across various computer vision tasks designed to assess each model's ability in object addition, color changes, object detection, segmentation and artistic style applications. All figures referenced from left to right depict the original image, followed by the edits made by InstructDiffusion, and concluding with those made by DynamicImportance. Our analysis is centered on how precisely and effectively each model executes the given instructions.

The models were evaluated on tasks depicted in the provided Figure 3, which included applying styles to group images featuring pets, drawing precise borders around individual animals, and transforming natural landscape scenes into artistic renditions. Both InstructDiffusion and DynamicImportance were tasked with applying Van Gogh's



"Transform it to van Gogh, starry night style."



"Draw a purple border around the dog and maintain the current appearance of the rest."

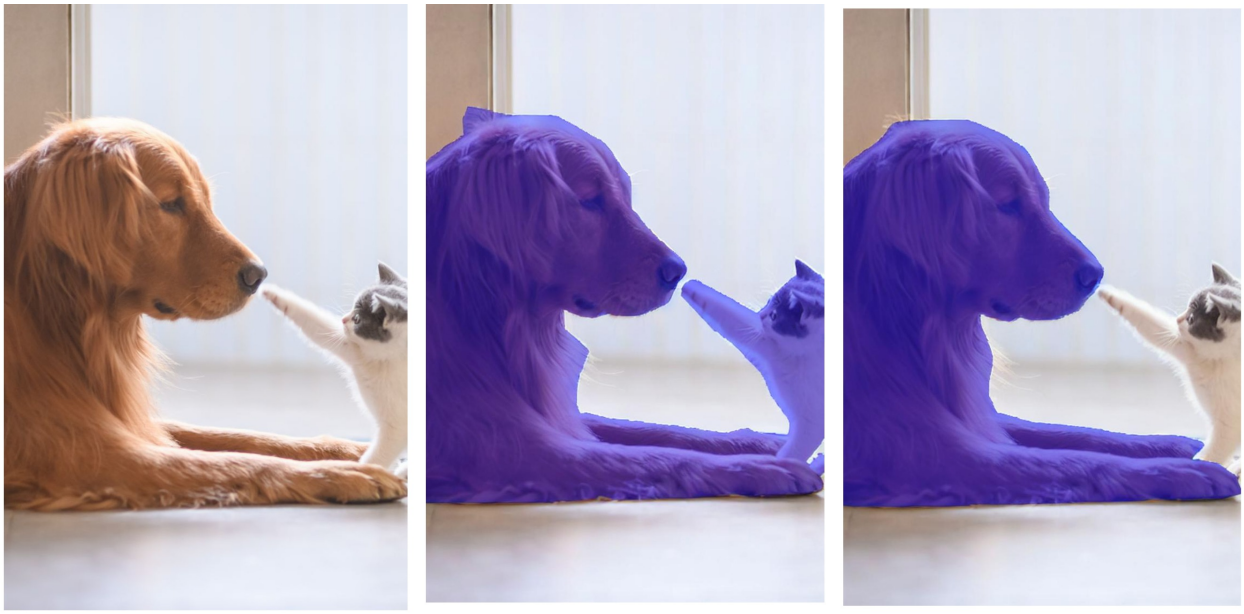
"Make the image look like an oil painting."

Figure 3

iconic swirling style to a photograph featuring dogs and a cat. InstructDiffusion captured the essence of the style, but it tended to slightly affect the natural appearance of the animals. DynamicImportance, leveraging its ability to focus on highlighted tokens, preserved more of the original features and colors of the animals while still incorporating the distinctive brush strokes and color palettes of Van Gogh.InstructDiffusion and DynamicImportance were instructed to draw a purple border around the dog without altering the rest of the image. InstructDiffusion, however, failed to execute the task, showing no changes to the original image,in contrast, DynamicImportance successfully completed the task, applying the purple border neatly around the dog.In the task of transforming landscape scenes into oil painting-style renditions, both InstructDiffusion and DynamicImportance performed well, successfully applying the oil painting effect. However, DynamicImportance exhibited superior handling of color blending and vibrancy, enhancing the overall aesthetic appeal of the artistic rendition while maintaining the structure of the original landscape.

In the evaluation of the models' ability to perform precise color modification and segmentation tasks (Figure 4) , we encountered specific challenges with the InstructDiffusion model. The first task required marking a dog's pixels blue. InstructDiffusion, however, inadvertently extended the blue mask to include not only the dog but also a nearby cat, demonstrating a limitation in its ability to precisely target and isolate the intended object within a complex scene. Conversely, DynamicImportance incorporates highlighted importance tokens in the prompt (in this case, the dog), displayed a more accurate performance. It successfully restricted the blue segmentation mask solely to the dog, unaffected by the proximity of the cat. This improved result from DynamicImportance underscores the effectiveness of using targeted importance signals in the model's prompt to enhance object-specific segmentation tasks. In the task involving selective color transformation of furniture, specifically changing the color of only the right chair from brown to white, both InstructDiffusion and DynamicImportance faced challenges in maintaining the integrity of the scene. InstructDiffusion, while tasked with changing only the right chair, inadvertently altered the colors of the left chair and the table as well.In the task of selectively transforming the color of the right chair from brown to white, DynamicImportance showed an improved ability to accurately target the specific object due to the utilization of highlighted importance tokens in the prompt. However, despite this focused targeting, the model still inadvertently altered the chair's shape slightly. This alteration is a limitation inherited from its foundational architecture, InstructDiffusion.

The models were tasked with adding baseball caps to corgis and sunglasses to a running cat (Figure 5). InstructDiffusion's placement of a cap on the middle corgi was slightly misaligned from the intended rightmost corgi. Conversely, DynamicImportance correctly identified the target corgi. For the sunglasses task, InstructDiffusion positioned the sunglasses above the cat's head as well, whereas DynamicImportance aligned the sunglasses precisely only with the cat's eyes, demonstrating superior accuracy in object-



"Mark the pixels of the dog to blue and leave the rest unchanged."

"Change the color of the chair from brown to white."

Figure 4

specific placement.

6. Conclusion

This research has successfully introduced DynamicImportance, a methodical enhancement for the InstructDiffusion model which dynamically assigns importance to the words within textual prompts across a spectrum of computer vision tasks. This approach refines the cross-attention mechanisms within the model, ensuring that critical textual elements are emphasized. The effectiveness of DynamicImportance in enhancing the precision and contextual relevance of output images represents a significant advance in text-driven image editing technologies. Our methodology builds on the foundational concepts of Attend-and-Excite by optimizing attention distributions to selectively amplify features that are crucial for a given task, while suppressing irrelevant information. This selective attention mechanism is crucial for tasks that require high fidelity and nuanced interpretation of textual instructions to generate visually coherent images. With DynamicImportance we have demonstrated that it is possible to significantly boost the model's performance, making it not only more accurate but also more responsive to the subtleties of human language. The experimental results underscore the robustness of our approach. Through those we have shown that DynamicImportance improves the model's ability to interpret complex instructions, leading to outputs that are not only more accurate but also artistically and contextually aligned with the input prompts. This enhancement in performance is particularly evident in the model's ability to handle intricate image manipulation tasks that traditional

"Help the right corgi wear baseball caps."

"Add sunglasses to the cat and maintain the rest unchanged."

Figure 5

diffusion models often struggle with. Furthermore, this research has explored the implications of enhanced attention mechanisms in AI-driven systems, suggesting that similar approaches could be beneficial across various applications in computer vision and beyond. The ability of DynamicImportance to refine model outputs based on textual analysis could be extended to other areas such as automated content generation, real-time video editing, and other interactive tools, where precision and context sensitivity are paramount.

References

[1] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 2, 3, 5

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 2

[4] 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. 3

[5] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021. 2

[6] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, and Baining Guo. Instructdiffusion: A generalist modeling interface for vision tasks. *CoRR*, abs/2309.03895, 2023. 2, 4, 5

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[8] 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. 2, 6

[9] 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

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference,*

Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015. 3

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3

[13] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3, 5

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2