DL Project Proposal: Team GKMS
Image Generation via Stable Diffusion Using Latent Space Manipulation

Garrett Devaney Kshitij Pathania Moises Andrade Sneha Maheshwari
Georgia Institute of Technology
{gdevaney3, kpathania3, mandrade, smaheshwari63}@gatech.edu

1. Introduction

1.1. Objective

Our goal is to enhance pretrained stable diffusion models to generate images that maintain strong associations with both the provided prompts and specific objects or features in the input images. By achieving this, we aim to empower end-users with greater control over the image generation process. For instance, given an input image depicting a dog and a cat playing in a park and a prompt describing the scene, we intend to produce a new image of a park while retaining the original dog and cat in active states.

1.2. Background

Diffusion models (Rombach et al., 2022 [10]; Saharia et al., 2022; Ramesh et al., 2022 [9]) offer a novel approach to high-resolution image generation by leveraging text prompts and existing photos. However, existing architectures often struggle to maintain strong ties between the generated images and the input references. Specifically, when utilizing stable diffusion for image generation with text prompts and reference images, the influence of the reference image is frequently diminished during the process. Moreover, the lack of model transparency impedes explainability and control over the generative decisions.

2. Motivation

The ability to generate images closely aligned with user input has significant implications across various domains. It facilitates creative content generation, aids in visual storytelling, and enhances user experiences in applications such as image editing and virtual environments. By enabling users to exert more influence over the generative process, our approach promises greater user satisfaction and utility.

Figure 1: Example of generated image using our approach when reference image is above original image and provided prompt is: Lady(Subject) on a serene sandy crest(scene).

3. Related Work

Recent works propose methods for image editing/refinement by directly manipulating cross-attention mechanisms or encouraging them to follow certain patterns[1]. For instance, [6], [11], and [7] propose training-free methods that prompt attention blocks to focus on specific areas given additional textual or visual inputs. [13] and [1] augment pretrained models with image features to promote preservation of recurring elements, such as human skeletons and outlines. [3] enforces cross-attention to attend to all subject tokens in the prompt to drive image generation that fully represent the described subjects. Our proposal intersects with these works, as we indirectly promote Stable Diffusion's attention mechanisms to preserve contextual fidelity by infusing information on noised representations of the original images.

4. Data

For dataset creation, we utilized the COCO (Common Objects in Context) dataset https://cocodataset.

[1]For a thoroughly survey on diffusion models and their variants, see [2].

org/, a widely recognized benchmark for object detection, segmentation, and captioning tasks in computer vision. The COCO dataset encompasses a diverse array of images with annotated objects across various categories, providing an ideal resource for training and evaluating image generation algorithms.

We curated our dataset by selecting specific object categories from the COCO dataset that align with our objectives. These categories include airplane, automobile, and animal, covering a broad spectrum of objects with each category having 10 images. Furthermore, we subdivided each category into specific types or classes, such as aeroplane, bus, car, and various animals like dog, cat, elephant, and horse.

Additionally, for generating images we considered the contextual scenes for the images to set diverse context for the images on which we want to test them. Scenes such as "Runway," "Cloud," "Highway," "City," "Beach," "Forest," and "Desert" were identified and associated with the corresponding images. When considering the scenes associated with each image, the dataset expands significantly. With multiple scenes per image, the effective size of the generated images grows accordingly.

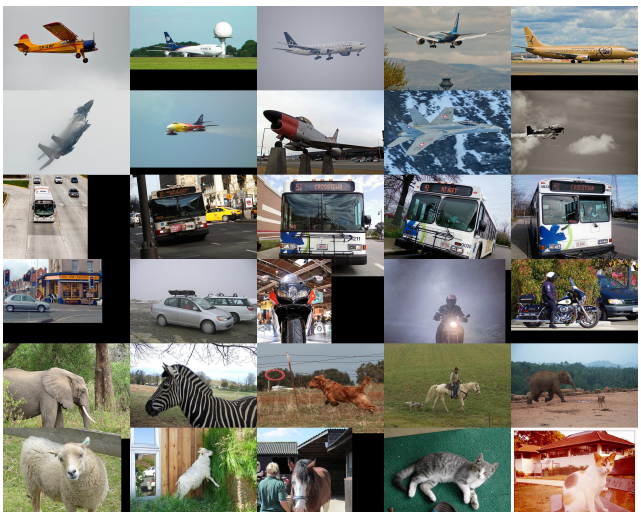


Figure 2: Sample of reference images in curated Dataset.

5. Approach

Our approach entails enhancing stable diffusion models for text-to-image synthesis by integrating dynamic masking techniques. The key innovation lies in leveraging cumulative attention scores derived from textual prompts to selectively preserve contextual information while generating visually coherent images. This novel approach aims to address the challenge of maintaining strong ties between generated images and input references, thus empowering users with greater control over the generative process.

The pipeline incorporates several crucial hyperparameters:

1. **Prompt (Text)**: Defines the subject details of the output image. In all our experiments we fix the prompt in a particular format:

   $$< \text{Subject} > on < \text{Scene} >$$

   We do this to compare all the experiments fairly so that we can reduce any bias on the basis of prompt.

2. **Input Image**: Provides the subject which we want to retain in the generated image. This is the original image we will give our model as the reference image to generate new images.

3. **Noise Strength**: Determines the level of noise added to the latent vectors during generation, we experimented by keeping noise at different values such as $0.6, 0.7, 0.8, 0.9, 1.0$ and we figured that 0.7 works best for our experiments, hence we use strength $= 0.7$.

4. **Classifier-Free Guidance Scale**: Influences the strength of guidance provided to the diffusion model.

5. **Inference Step Count**: Specifies the number of steps taken during the inference process.

6. **Seed Specifier**: Facilitates reproducibility of results by specifying a seed value.

5.1. Baseline Model

For baseline model we utilized the foundational codebase for the diffusion model available in this link. This codebase integrates the authentic weights of the stable diffusion model, sourced from Hugging Face's repository. We developed an image generation pipeline that incorporates several key parameters: a textual prompt defining the subject details of the output, an input image serving as the background, constants for noise strength, classifier-free guidance scale, and inference step count. We kept our noise strength to 1, cfg scale to be 8 and kept our number of inferences steps to be 50.

For experiments, our implementation begins with the prompt analysis, we utilize the spacy language processing library to extract words correlating with the subject matter. The input image is then encoded, and its latents with and without noise are calculated for specified timesteps. During the denoising step, cumulative attention maps for the subjects present in the prompt are computed, informing the construction of a dynamic mask. This mask identifies regions in the latent vector to be replaced with the original image to preserve context.

Experimentation involved two main phases.

Figure 3: Initial altered stable diffusion architecture

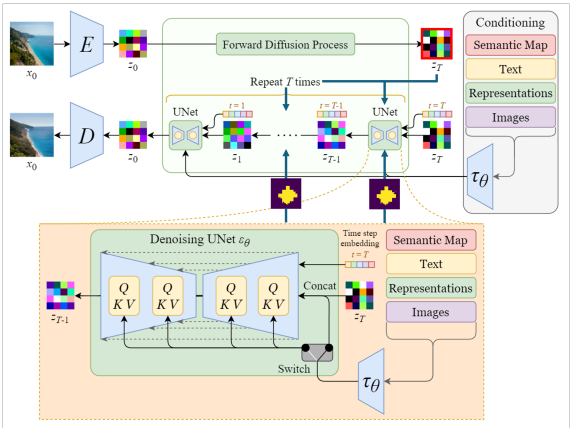


Figure 4: Aggregated attention mask and z_τ

Figure 5: Final altered stable diffusion architecture

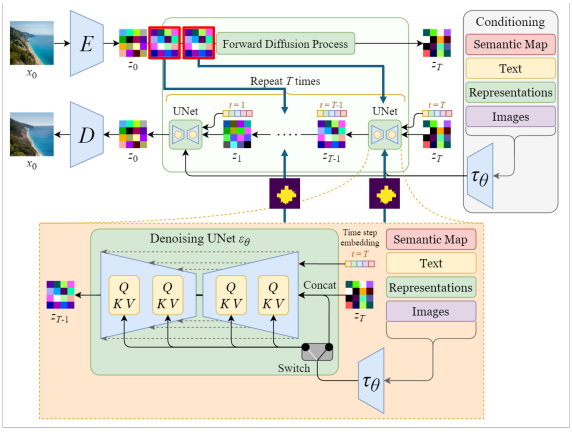


Figure 6: Scene-based attention mask and cached latent vector

5.2. Experiment 1

In Experiment 1, As shown in Figure 3, we first chose to select random latents in the denoising steps, and ultimately added the noised image representation at three timesteps in early stages of denoising. Also, dynamic masking was applied at selected timesteps. We calculated masks as bounds on cumulative attention map and experimented by setting the threshold on masking as:

$$\text{masked attention map} > \mu(\text{cumulative attention latents})$$

Figure 4 displays the masks guided by attention scores derived from the prompt and the image representation after forward diffusion. Our results demonstrated the effectiveness of our approach in preserving context and blending images compared to the baseline model. We use the following hyperparameters for this experiment:

- strength $= 0.7$, threshold $=$ mean

5.3. Experiment 2

In Experiment 2, we aimed to improve upon the results of Experiment 1 by visualizing the noised latent vectors and corresponding masks at each manipulated timestep. Ablation studies were conducted to determine the optimal threshold for the dynamic mask, balancing context preservation and mask coverage. Additionally, instead of aggregating cross attention maps, we only focused on the cross-attention map building the background scene (e.g. "beach" attention map given prompt "dog on beach"). We crafted a new formula and generated a hyperparameter to guide thresholding:

$$\text{masked attention map} > e^{\mu(\log(\text{subject attention latents}))+\lambda}$$

During our ablation studies, we noticed the noised latent produced by forward diffusion (z_τ) is not a good representation of the original image context. Therefore, we cached specific latent vectors early in the noising process to retain key features of the original input image (Figure 5). The cached latent vectors then replaced only the pixels defined by the mask as the subject at select denoising timesteps. Figure 6 displays the advancements in mask thresholding to better silhouette our subject and the context retained from the original image by choosing latent vectors in earlier stages of forward diffusion. This technique forced stable diffusion to account for subject features while also giving the model creative power to generate a new scene based on the prompt. We used the following hyperparameters for this experiment:

- strength $= 0.7$, threshold $\lambda = 0.2$

Throughout development, we encountered challenges such as handling prompts with multiple subjects, determining the threshold for the dynamic mask, managing computational complexity, and conducting comprehensive evaluations. These challenges prompted iterative refinement of our approach and underscored the need for careful optimization and experimentation.

Our approach represents a significant advancement in the field of diffusion models for image generation, offering improved sample quality and greater user control. By systematically integrating dynamic masking techniques with stable diffusion models, we aim to enhance the accuracy and flexibility of text-to-image synthesis, opening avenues for diverse applications across various domains.

6. Experiments and Results

6.1. Experimental Setup

To measure the efficacy of our approach, we compared images generated with a baseline Stable Diffusion model to the ones generated with the two methods described in section 5.

For this purpose, we specified different types of scenes for each category. Ten reference images for each category were obtained from the COCO dataset as described in the dataset section 4. The prompts for each category were created using the following format:

$$<\text{Subject}> \textit{ on } <\text{Scene}>$$

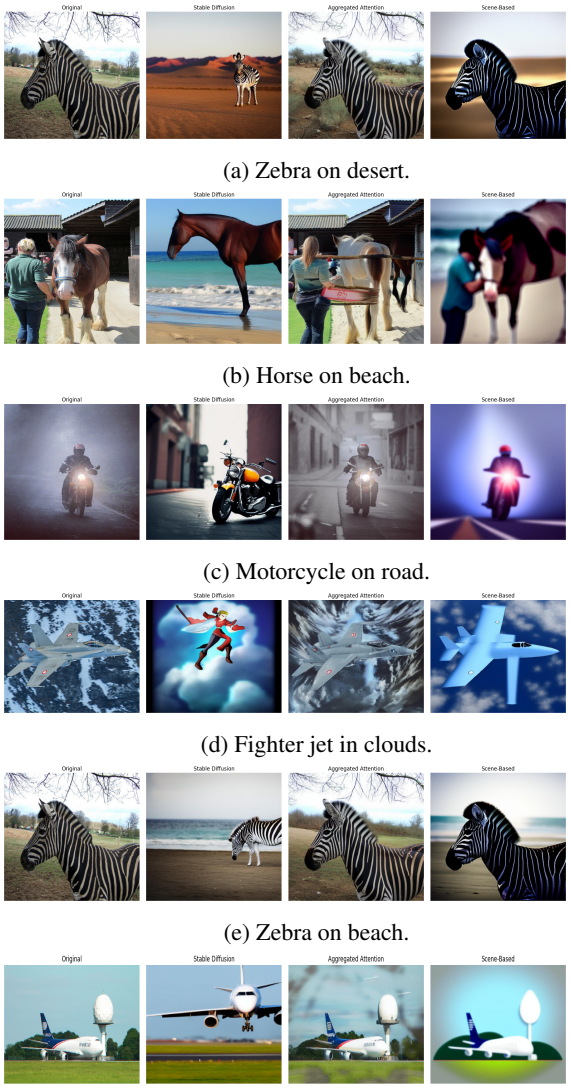
For experiments, we provided the created prompt along with the corresponding reference image. The tuple {prompt, reference} image was fed into the standard Stable Diffusion and the two proposed methods.

For instance, within the *Airplane* category, images were generated with contextual scenes including *Cloud* and *Runway*. Likewise, in the *Animal* category, images were created with diverse settings such as *Beach*, *Desert*, and *Forest*. Each category encompasses a total of 10 images. To thoroughly assess the quality and variation of the generated images, a total of 70 distinct images were generated by each model across all categories and scenes.

These generated images provide a diverse set of contexts for evaluating the performance of our models. We will analyze various metrics such as image quality, coherence with the provided prompts, and diversity of generated scenes to assess the effectiveness of our proposed methods.

6.2. Qualitative Analysis

We performed qualitative assessments where visual inspection confirmed the improved coherence and fidelity of images generated by our model.



(a) Zebra on desert.

(b) Horse on beach.

(c) Motorcycle on road.

(d) Fighter jet in clouds.

(e) Zebra on beach.

(f) Airplane on runway.

Figure 7: Assessing image outputs from diverse models: A comparative study.

Figure 7 contrasts selected images produced by the different model variations and the reference image. As seen in the figures, our methods are generally more effectively than Stable Diffusion to preserve the subjects while changing the context. Nonetheless, the models are not perfect and still struggle in some instances.

For instance, while Scene-Based diffusion tends to be more effective in creating the new scenarios while retaining the subjects, it also creates cartoonish-like figures (see 7, d and f). 'Aggregated Attention', on the other hand, is able to generate more realistic figures, but sometimes fail to produce the complete requested context (see 7, (c) and (e)).

6.3. Quantitative Analysis

To verify quantitatively our qualitative conclusions, we computed four evaluation metrics commonly adopted in the computer vision literature: Frechet Inception Distance (FID), Peak Signal Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) and Inception Score (IS).

Figure 8 and 9 shows the results considering all generated images and a breakdown by our main categories. The appendix provide also the scores broke down by the scenes in each category.

In line with the qualitative results, our methods demonstrate superior performance than the baseline model for most metrics. Specifically:

- **Fréchet Inception Distance**: As observed in the qualitative evaluation, our Aggregated Attention method achieves FID scores substantially lower than the baseline, indicating produced images more realistic and with higher similarity to the reference images. On the other hand, Scene-Based diffusion performs worse in this metric, reflecting the fact that while it generates more diverse images, eventually them show less realistic traces.
- **Peak Signal Noise Ratio (PSNR)**: Both our methods generated images with higher PSNR ratios than the baseline, reflecting the better fidelity and detail preservation in the reconstructed image compared to the reference.
- **Structural Similarity Index Measure (SSIM)**: Both methods exhibited higher SSIM scores compared to the baseline model, reflecting a higher preservation of structural properties and visual similarities relative to the original image, crucial for seamless blending of the existing background without distorting the scene's integrity.
- **Inception Score (IS)**: All methods achieved a similar and high Inception Score, indicating that it produces images that are both distinct and of high quality across a variety of categories.

	Method		
Metric	Stable Diffusion	Aggregated Attention	Scene-Based
FID	8.56	1.17	15.52
SSIM	0.34	0.44	0.45
IS	1.05	1.06	1.06
PSNR	10.10	14.91	14.30

* Darker = Better score

Figure 8: Evaluation Scores - All images

	Stable Diffusion	Aggregated Attention	Scene-Based
Airplane	5.6	1.2	7.1
Automobile	9.2	1.7	20.1
Animal	17.2	2.0	25.8
All	**8.6**	**1.2**	**15.5**

* Darker = Better score

(a) FID Scores

	Stable Diffusion	Aggregated Attention	Scene-Based
Airplane	0.6	0.7	0.7
Automobile	0.2	0.4	0.4
Animal	0.2	0.3	0.3
All	**0.3**	**0.4**	**0.4**

* Darker = Better score

(b) SSIM Scores

Category	Stable Diffusion	Aggregated Attention	Scene-Based
Airplane	11.07	17.04	16.36
Automobile	8.93	13.20	12.89
Animal	10.23	14.64	13.87
All	10.10	14.91	14.30

* Darker = Better score

(c) PSNR Scores

Figure 9: Evaluation Scores by Category

Limitations: Our task is a middle-ground of image segmentation and generation: while we want to generate the correct background, we want to preserve the subjects. Therefore, we acknowledge that *individually* the above metrics are not perfect representations of performance. This is perhaps clear in the low FID score for the 'Scene-Based' diffusion, which also reflects its ability for more varied images. Two notes on this: 1) While individually the metrics are not representative, having a higher score as a whole, combined with the qualitative evaluation, still gives a good indication that our methods improved on key aspects relative to the baseline Stable Diffusion model.

2) We experimented with another option: utilizing SOTA segmentation models to create "ground truth" labels based on the original images and compute the mIoU for the different models. We were not able of producing all the results, but the codebase [4] contains our partial implementation for reference.

7. Contribution

Contributor	Contribution
Garrett Devaney	Led Experiment 2, conducted the ablation study, analyzed the effects of different model parameters.
Kshitij Pathania	Managed baseline code, experiment generation, dataset preparation, and contributed to documentation.
Moises Andrade	Oversaw the compilation of results. Conducted the quantitative and qualitative evaluation. Was heavily involved in the documentation process and literature review.
Sneha Maheshwari	Led Experiment 1, contributed to the ablation study, and assisted with the project documentation.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 1

[2] Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z. Li. A survey on generative diffusion model, 2023. 1

[3] 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. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1

[4] CS-7643-Project-DL. Diffusion model latent space manipulation. https://github.com/CS-7643-Project-DL/Diffusion-Model-Latent-Space-Manipulation, 2024. Accessed on: 2024-05-01. 5

[5] Yuki Endo. Masked-attention diffusion guidance for spatially controlling text-to-image generation. *The Visual Computer*, pages 1–13, 2023.

[6] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1

[7] Wan-Duo Kurt Ma, J. P. Lewis, Avisek Lahiri, Thomas Leung, and W. Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance, 2023. 1

[8] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023.

[9] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 1

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[11] 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, 2023. 1

[12] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

8. Miscellaneous

8.1. Images Generated by Implemented Models

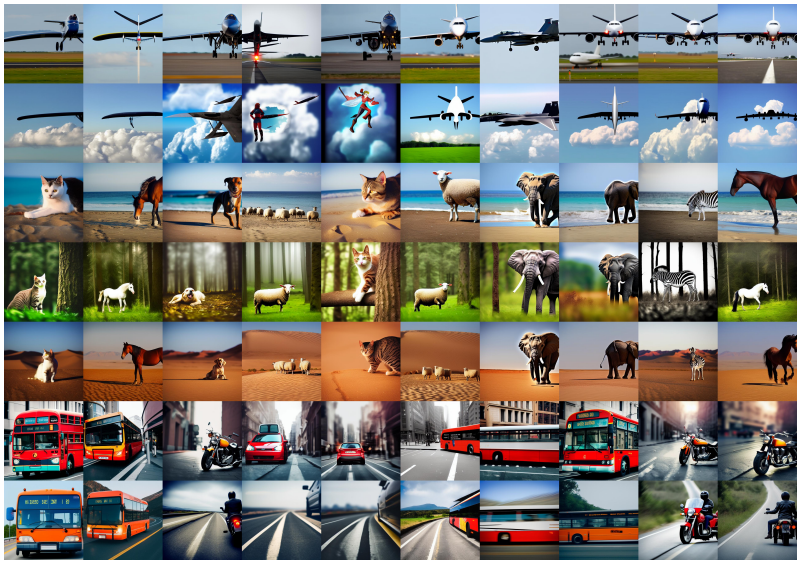


Figure 10: Image generated by Baseline Stable Diffusion Model

Figure 11: Image generated by Model based on Experiment 1

Figure 12: Image generated by Model based on Experiment 2

8.2. Breakdown of Scores by Scenes

		Method		
Scene	**Metric**	Stable Diffusion	Aggregated Attention	Scene-Based
Cloud	FID	8.24	1.92	5.93
	SSIM	0.60	0.70	0.74
	IS	1.00	1.01	1.01
	PSNR	10.90	15.59	16.57
Runway	FID	5.61	0.90	9.17
	SSIM	0.58	0.76	0.75
	IS	1.00	1.00	1.00
	PSNR	11.24	18.48	16.15

Figure 13: Scores - Airplane Category

		Method		
Scene	**Metric**	Stable Diffusion	Aggregated Attention	Scene-Based
Beach	FID	22.29	2.05	29.79
	SSIM	0.25	0.30	0.30
	IS	1.00	1.01	1.01
	PSNR	10.31	15.23	13.89
Forest	FID	12.88	2.81	25.27
	SSIM	0.20	0.27	0.29
	IS	1.00	1.00	1.00
	PSNR	9.58	13.99	13.66
Desert	FID	27.18	2.69	23.49
	SSIM	0.26	0.29	0.30
	IS	1.00	1.00	1.00
	PSNR	10.81	14.70	14.05

Figure 14: Scores Breakdown - Animal Category

		Method		
Scene	**Metric**	Stable Diffusion	Aggregated Attention	Scene-Based
City	FID	6.95	1.78	22.72
	SSIM	0.22	0.38	0.37
	IS	1.00	1.01	1.01
	PSNR	8.57	13.28	12.69
Highway	FID	14.76	2.30	18.46
	SSIM	0.27	0.40	0.38
	IS	1.00	1.00	1.00
	PSNR	9.29	13.12	13.08

Figure 15: Scores Breakdown - Automobile Category