Study of Text-guided Image Inpainting Model Performance - Final Report

Yuhang Qian Tanmay Jain Jiayi Zhang Angelos Guan Jimin Yoon

I. Introduction

The rapid advancement of technologies in recent years has led to the emergence of text-guided image inpainting as a thriving subdomain within the broader field of text-guided image generation. This final report undertakes a comprehensive investigation of the predominant factors contributing to the performance of various pipelines in the text-guided image inpainting research domain. Initially, an extensive review of the literature was conducted, followed by an in-depth investigation of several cutting-edge model pipelines to understand how comparable performances were achieved despite their distinct model architectures. Subsequently, controlled experiments were performed on the two unique inputs of the generative model, specific to this domain, in the baseline pipeline, to further study the potential for enhancing the model's performance. Additionally, analogous experiments were carried out on different pipelines to assess the generalizability of our findings across the different model architectures.

II. Previous Work

In recent years, there has been a significant improvement in the quality and versatility of image generation models that are based on text prompts [1], [2], [3], [4]. This has especially been made possible with the introduction of novel deep learning architectures [5], families of generative models such as diffusion [6], [2]; and finally, large-scale image-text paired datasets. However, not all architectures enable natural-language conditioned image inpainting, which is the focus of our project. Therefore, this section examines various related methods that primarily address this task.

A. CLIPSeg with Stable Diffusion

CLIPSeg[7] is an image segmentation model which extends the well-known CLIP transformer for referring expression, zero-shot, and one-shot segmentation tasks by integrating a lightweight transformer-based decoder. The CLIPSeg model uses the original visual transformer-based (ViT-B/16) CLIP[5] model as the backbone and extends it with a small, parameter-efficient transformer decoder. While CLIP itself was frozen, the decoder was trained with 1,122,305 parameters on the PhraseCut[8] dataset. Compared to the original CLIP transformer, CLIPSeg enables different image sizes as inputs by interpolating the positional embedding. Stable Diffusion[6] is an image generation model that uses latent diffusion to iteratively generates images over multiple timesteps from text prompts.

B. DIFFEDIT

DIFFedit[9] is an end-to-end diffusion-model-based method for semantic image editing using natural language prompts. Contrary to other diffusion-based methods, given a textual query, this method infers the relevant regions to be edited rather than requiring a user-generated mask or mask generated by a segmentation method such as CLIPSeg. It achieves this using a text-conditioned diffusion model, which estimates the difference between the noise conditioned to the query text and the reference text. This difference can be used to infer a mask that identifies what parts of the image need to be changed to match the query.

C. Muse

Muse[10] is a text-to-image Transformer model that achieves state-of-the-art image generation performance while being significantly more efficient than diffusion or autoregressive models. The model is trained on a masked modeling task in discrete token space, where it is given the text embedding extracted from a pre-trained large language model (LLM) and is trained to predict randomly masked image tokens. This approach enables fine-grained language understanding and translates to high-fidelity image generation with an understanding of visual concepts such as objects, their spatial relationships, pose, cardinality, and more.

III. Baseline Model Implementation

We implemented CLIPSeg[7] with Stable Diffusion[11] for our experiments. CLIPSeg is open-source, and its code base is readily available on GitHub. The entire pipeline is implemented with PyTorch with the additional CLIP visual transformer imported as a package. The pre-trained weights are provided in a PTH format. Multiple Jupyter Notebooks are also provided to set up the model and reproduce some of the results mentioned in the paper. We are able to initialize the model and achieve reasonable performance with sample input images and textual prompts. A sample output mask can be seen in Figure 1.

Stable-Diffusion-Inpainting with Stable-Diffusion-v-1-2 [11] was implemented as the generation model. The set of inputs to this model includes our upstream outputs such as the original image, mask image, and replacement prompt. Before feeding the inputs to the model, the images are preprocessed to dimension 512x512. This dimension matches the resolution of the training dataset of the inpainting on "laion-aesthetics v2 5+". The pipeline for the implementation was loaded

from the diffusers library[12], which has been implemented in PyTorch. The output of the stable diffusion model for the replacement text prompt "Cow" has been displayed in Figure 1.

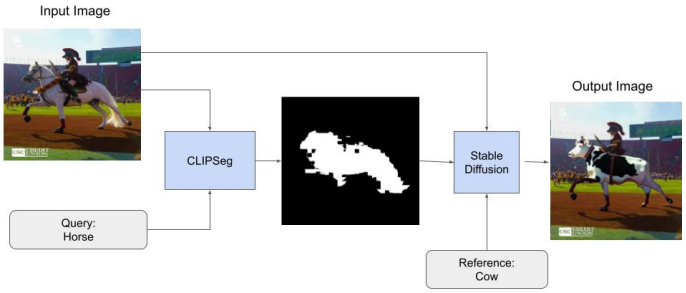

Fig. 1. Flow diagram of the CLIPSeg + Stable Diffusion pipeline with results

IV. Model Metric and Performance Comparison

We achieved good results with sample inputs from both pipelines. However, we recognize that it's unfair to benchmark these models on the same dataset to measure their performance as they are originally trained on different datasets. Given the time and resource constraints and the fact that our models utilize some of the most commonly-used model architectures nowadays, we decided to find models from other publications with similar architectures that are trained and benchmarked on the same dataset for both segmentation and generation tasks and refer to their performance as a fair estimate of our models' performance.

We researched and explored the state-of-the-art model performance for image segmentation and generation. We chose to use the COCO (Common Objects in Contexts) dataset published by Microsoft, as it's one of the most popular datasets researchers use for object detection, image segmentation, and captioning research. Since each image in COCO includes a caption, it perfectly matches our models. COCO includes multiple classes so that it is enough for us to test model performance. We obtained a current leaderboard of all published research papers with their models' performance benchmarked on the MSCOCO dataset and documented it below.

A. Image Segmentation

IOU (Intersection over Union) is used in computer vision and image processing to measure the accuracy of object detection algorithms. It calculates the overlap between the predicted bounding box and the ground-truth bounding box. It produces a score ranging from 0 to 1, where a higher score indicates better object localization. A high IOU score is desirable in object detection tasks, indicating accurate object localization.

B. Image Inpainting and Generation

1) FID (Fréchet Inception Distance): FID is a metric used to evaluate the quality of generative models in computer vision. It measures the similarity between the distribution of real images and the distribution of generated images using features extracted from a pre-trained neural network (Inception).

2) CLIP Score: CLIP Score [13] is a metric used to evaluate the accuracy of the CLIP model developed by OpenAI, which can understand both text and images. The score measures the similarity between text and image representations in the CLIP model. The higher the score, the more accurate the model's prediction is.

While FID is more commonly used to measure the performance of the generative models. We realized that CLIP Score is more suitable to measure the model's performance for our experiments since FID calculates the similarity between the original images and newly generated images, which doesn't fit our use case because we are modifying the context of the images, which would result in a decreased FID score.

Model	IOU
UNIEXT-H[14]	82.19
PolyFormerL[15]	75.96
VPD[16]	73.25
CLIPSeg[17]	69.52

TABLE I
Segmentation Model Performance Comparison

Model	FID	type
Parti[18]	3.22	Transformer
eDiff-I[19]	6.95	Diffusion
GLIGEN[20]	5.61	GAN and attention
Muse-3B[21]	7.88	Transformer

TABLE II
Inpainting Model Performance

As shown in I and II above, we evaluated various models on the leaderboard that employed different approaches and training datasets. We found that despite their differences, many were able to achieve comparable results on COCO. Most top-performing models utilize transformer, diffusion, or GAN architectures in their pipelines. This has inspired us to study further and analyze the specific factors contributing to the model's performance for image inpainting and generation.

V. Experimentation and Results

A. Dataset Curation

Since text-guided image inpainting is a specific subdomain of text-guided image generation that requires a replacement object prompt and an input mask in addition to the other inputs, no such datasets currently support the benchmarking of this subdomain, as confirmed by [22]. Therefore, following the same approach, we curated our own dataset by expanding the COCO dataset with the missing components mentioned above.

Considering the nature of our experiments and the resource constraints, we randomly selected 25 images from the COCO dataset, which scales into around 500 generated images per experiment.

B. Metric Selection

As suggested by [22], human evaluation via blind surveys is known to be the best and most accurate way to quantify the generative model's performance in this field. As explained in the section above, no metric currently measures the model performance for our specific domain. However, CLIP score, as mentioned previously, which measures the similarity between the generated image and the input prompt, usually provides good confluence to the human evaluation results quantitatively. CLIP score has been shown to correlate positively with human evaluation results by [22]. Therefore, we performed both human evaluation and CLIP score calculation for our experiments.

C. Generative Models Implementation

Beyond the baseline model pipeline (discussed in section III), we also tried to implement other generative models to perform extensive experiments on them. The idea was to compare generative models belonging to different architecture families - stable diffusion, masked modeling, and GANs. However, we could not finally include them in our experiments due to computational restrictions and imperfections in the open-source implementation of these methods. Firstly, we implemented TDANet [23], which uses a Text-Guided Dual Attention Inpainting Network to fill the corrupted/masked region. The method was trained and evaluated on the Caltech-UCSD Birds 200 and the COCO datasets. It was implemented in Pytorch, and the code and pre-trained weights are available in their official GitHub repository. We, however, did not use it since the implementation pipeline with the given pre-trained weights does not function properly.

Unfortunately, the official code for OpenAI's DALLE-2[1] is not publicly available. Hence, we tried an open-source implementation of DALLE-2, available on Huggingface know nas mini-DALLE. However, since the inpainting module of the pipeline had errors and produced highly incorrect results with the given pre-trained weights as shown in Figure 2, we could not use it for our experiments. We suspect it did not function as desired because the publicly available pre-trained weights were trained specifically for a different task. Finally, we planned to implement MUSE, a masked-modeling-based approach. But, the pre-trained weights are not available in their official implementation, and training the model from scratch was simply impossible given the limited computational resources. As a result, we decided to use the official DALLE-2 API from OpenAI with a fee, which produces stable and high-quality results as shown in the sections below.

D. Experiments Details

1) Input Mask Size: This experiment investigates how changes in the input mask size w.r.t. the target object size affect

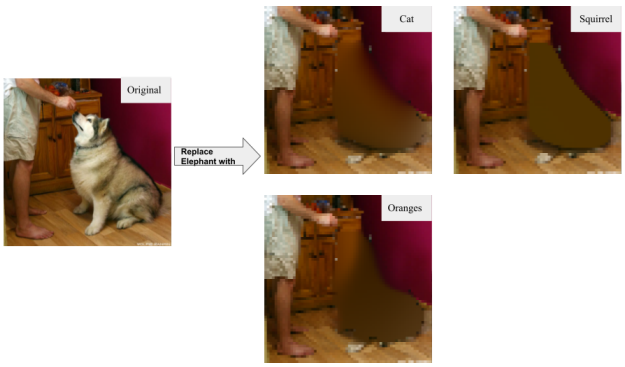

Fig. 2. Results of generated images with mini-DALLE

the generated image quality. The binary mask image is first generated by a probability threshold based on the segmentation model and then fed into the downstream image generation model as an input to create the final image output. Initially, we processed the binary mask by adding and subtracting a variable proportion of pixels vertically and horizontally around the edge to alter its size relative to the target object. Later, we improved the method by switching to adjusting the probability threshold used in generating the binary mask to maintain the mask's shape better.

The experiment was strictly controlled - we froze the upstream image segmentation model CLIPSeg, the original image, prompt inputs, and the downstream image generation model Stable Diffusion. We adjusted the mask size by changing the default probability threshold by ±15% and ±30%. Five sample images were generated for the five (±15%, ±30%, and default) mask sizes for the 25 baseline images curated in the dataset. The best image for each mask size and each input image was then selected from the 625 generated images to form the final results of 125 images. One set of examples is shown in Figure 3 (full outputs are included in Appendix).

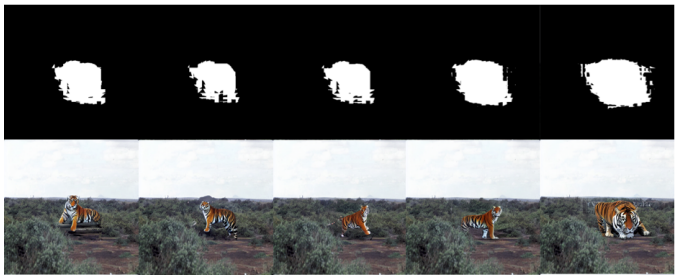

Fig. 3. Stable Diffusion generated image samples using different input mask sizes

Qualitative and quantitative analyses were performed for evaluation as suggested by [22]. Human evaluation was done via a blind survey via Google Forms. Ten multiple-choice questions were formulated with three unlabeled options: an image generated by the default mask, one by the enlarged mask, and one by the shrunk image. Given the original unmodified image and the input prompt, survey takers are asked to select

the unlabeled option that best fits the input prompt and has the highest image quality. Quantitative analysis was performed by calculating the CLIP scores of the generated images from the five different mask sizes for each original image. CLIP score measures the similarity between the generated image and the given text prompt. For the results of the experiments, 40

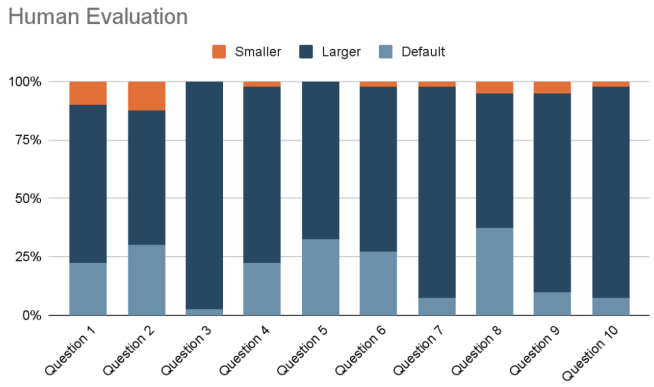

Fig. 4. Blind survey results show that larger binary masks produce better results consistently

human evaluation responses were received via Google Forms. The results shown in Figure 4 show that more than 50% of the surveyed population think that the images generated by binary masks larger than the defaults produced consistently better results across all ten images. In comparison, less than 10% think the smaller masks produced better results than the defaults for all images. The calculated CLIP score also shows a slight improvement among images generated by larger masks and a slight drop for smaller ones. However, the changes are insignificant and on average within 1% of the default measurements across all 125 images (shown in Appendix).

Although CLIP score is only meant to be considered as additional confluence, we don't think it accurately measures the scope of changes for our specific tasks - it measures the similarity between the entire image and its caption, dampening its effectiveness in measuring the change of one object inside the image. This is likely why we only saw insignificant changes reflected in the calculated CLIP scores across the board. To

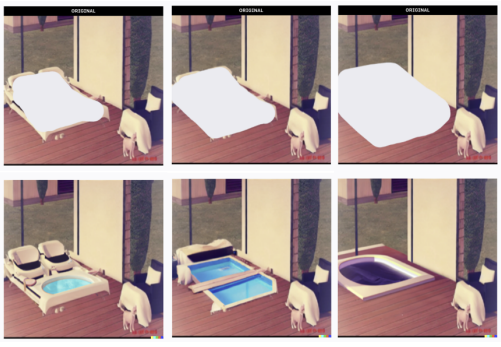

Fig. 5. DALLE-2 generated image samples using different input mask sizes, the larger mask generated better results than the smaller mask

examine how well these findings generalize across the overall research field, we also performed the same experiments in OpenAI's latest DALLE-2 image modification pipeline. Since the model does not allow image mask modification via altering probability thresholds, we manually adjusted the mask size using its pixel selection tool. As shown in Figure 5, the results and findings align with our original pipeline results.

Upon detailed investigation of the input mask size and its correlation to the target object, we conclude that, in general, the generative model's performance improves when the input mask size is big enough to cover the entire object's edges and contains some background pixels surrounding the objects to provide the model with more contextual information. However, it's also worth noting that when the mask size gets too large, the overloaded contextual information could pollute the original object and lead to loss of information, such as the original gesture and composition of the object, shown in Figure 6.

Fig. 6. An overly large binary mask on the right leads to the generated object losing its original composition in both Stable Diffusion (row 1) and DALLE-2 (row 2)

2) Input Mask Quality: In this set of experiments, we aim to study the effect of adding noise to the segmentation mask on the generated image quality. The metric we used was the change rate of CLIP score of the generated image from the noisy segmentation mask against the generated image from the original segmentation mask from the segmentation pipeline.

In order to study the effect of noise in the segmentation mask, we are fixing the segmentation pipeline with the CLIPSeg model as a control variable. The segmentation model produces a segmentation mask using the default probability threshold as described in the first experiment. We then pass the edited mask we generate from the noise addition function to the stable diffusion image generation model. For each input image, we generate the original segmentation mask from the segmentation model, then pass the original mask into a noise generation function, where we flip the labels of random pixels in the given mask. The adjustable parameters of the noise generation function include noise mode and noise level. Noise level is the percentage of noise to add with regard to the total pixels in the segmentation area from the original mask. Noise mode defines the way we generate noise. There are three modes in total: (1) in the segmented area, randomly pick segmented pixels and convert their masks to the background. (2) outside the segmentation area, randomly pick background pixels and convert their masks to segmentation. (3) mixture of

(1) and (2). We chose four noise levels to run the experiment on: 0.0001, 0.001, 0.01, 0.1 for each of the three modes. In total, we produced 12 edited masks (Fig.7) and their generated images for each input image (see appendix) in the self-curated dataset and averaged the CLIP score of all input images in the dataset for each combination of noise level and noise mode in Table III. Some qualitative examples are shown in Figure 17. A subset of the qualitative experiment result on the entire curated dataset is shown in the appendix. It consists of noise levels 0.001 and 0.1 for three noise modes for each input. The quality of the image is straightforward for human evaluation as more noise tends to decrease the quality of the image, thus, we only performed CLIP score calculation and result generation and omitted the survey for this experiment.

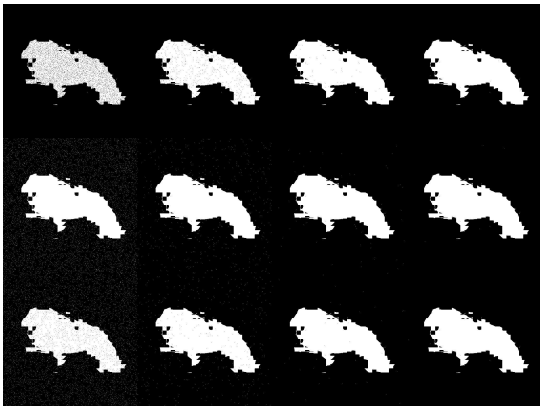

Fig. 7. Complete set of generated input masks with different noise levels and modes for one input image

For the result of this experiment, quantitatively, the change rate of CLIP score values is too low to draw any meaningful conclusions; qualitatively, we can see several trends in the generated output images: (1) In all three experiment modes, when the noise level is over certain threshold specific for each image, adding noise in the segmentation mask will introduce noise in the generated image. (2) Adding more noise in the mask will introduce more noise in the generated image in all three experiment modes. (3) In the experiment mode where we only add noise to the background pixels, the generated results inside the original segmentation area are still affected due to the change in contextual information. In addition, the generation results in the original segmentation area show similar noise patterns to the artificially introduced noise pattern in the original background. The upshot is the model can learn contextual information from pixels outside the replacement object. However, the performance will be compromised if the mask is too noisy.

The limitation of using CLIP score as a metric for measuring the quality of the generation results may come from the fact that it measures the similarity between the entire image and its text caption. In other words, it considers the entire image instead of just the segmented area that the generation model modifies. The quality of the image generation may only be a small portion of what the CLIP score measures, thus, the change in CLIP score between the original mask and the noisy mask is not numerically large enough to draw any meaningful conclusion.

	noise lv = 0.0001	noise lv = 0.001	noise lv = 0.01	noise lv = 0.1
segmented pixels only	-0.0877%	-0.2150%	0.2043%	0.3239%
background pixels only	-0.0166%	-0.0827%	0.2877%	0.1642%
mixed	0.0464%	0.4463%	0.4004%	-0.0181%

TABLE III
CLIPSCORE CHANGE RATE ON CURATED DATASET FOR INPUT MASK QUALITY EXPERIMENT

For future work, with more computing resource, we can try to run the experiment on more noise level to plot the trend of generated quality with regard to the noise level of the mask and compare the difference of such trends in the different noise modes. Furthermore, different noise pattern can be explored, for example, we can try to add noise with strips, circles, and random shapes to see how they will affect the performance of the pipeline.

3) Replacement Object Size Ratio: This experiment aims to study the impact of the size of the replacement object on the output image. The baseline pipeline test prompts targeting objects with a similar size; for example, a bowl of strawberries to a bowl of oranges, horse to zebra, and horse to donkey. This proposed experiment aims to study the result of the in-painting model when the replacing object size is varied.

During the set-up of this experiment, no existing solutions were found to control the size of the replacing object; the existing in-painting pipelines don't provide an API for the replacing object in the in-painting part specifically and only take in a text-based prompt. Therefore, the experiments were grouped by the relative difference in the object size ratio rather than a numeric value. For each pair of objects, the ratio was first calculated using the average volume of each object. For example, elephant replaced with squirrel (0.5% size ratio) was assigned the highest size difference of 4 and elephant replaced with tiger (70% size ratio) was assigned 1. The range of object ratios that were tested falls under the scale of 1 to 4.

To illustrate, an example of the calculations and the size chart are shown in Table IV.

	Tiger	Person	Squirrel	Elephant	Rabbit	Monkey
cubic centimeter	108900	62000	800	160000	1250	32400
size percent	68.06	38.75	0.5	1	0.78	20.25

TABLE IV
OBJECT SIZE RATIO CHART FOR ORIGINAL OBJECT OF ELEPHANT

The example of replacing an elephant with other objects of varying size ratio are shown in the Table V. The first row is the baseline CLIP score of the original image. For each replacing object size ratio, the CLIP score was normalized with the CLIP score of the newly generated image with the object replaced. The experiments were run on 6 original images with the object in the image being replaced by 4 new objects of different size ratios with respect to the original object. The full list of the object prompts is shown in Table VI.

	original-elephant			
CLIP Score of Original Image	23.87			
Object Size Ratio	1	4	2	3
CLIP Score of Replaced Image	22.75	23.65	25.11	24.28
CLIP Score (Normalized)	-1.12	-0.21	1.25	0.41

TABLE V
CLIP SCORE RESULTS FOR REPLACING AN ELEPHANT ACROSS OBJECT SIZE RATIOS

Original Object	Ratio 1	Ratio 2	Ratio 3	Ratio 4
Elephant	Tiger	Monkey	Rabbit	Squirrel
Clock	Airplane	Small Plane	Yellow Rose	Magic Wand
Black and White Bird	Squirrel	Oranges	Pink Tulip	Magic Wand
Laptop Computer	Squirrel	Oranges	Magic Wand	Fighter Jet
Husky	Cat	Squirrel	Oranges	Pink Tulip
Person	Dog	Tiger	Squirrel	Pink Tulip

TABLE VI
LIST OF REPLACING PROMPTS FOR EACH ORIGINAL OBJECT PROMPT

The first column labels the prompt of the object to be replaced in the original image and the subsequent columns list the prompt of the replacing object by the object size ratio. The experiments were run on size ratios across the scale of 1 to 4 as defined above and the results were evaluated with normalized CLIP score and human evaluations ratings. The resulting images are tabulated in Fig. 18. The first column is the original image input to the model and the subsequent columns are in-painted images that were generated with the replacing object prompt falling in the respective object size ratio of 1 to 4 for the respective original image. Qualitatively, the experiment with the smallest size shrinkage, for example, elephant to tiger, generated a new image with quite realistic quality. With the higher shrinkage rates, however, the new images were not very realistic, either the size of the object being unreal or generating a cartoon-style drawing instead of a photo-realistic one.

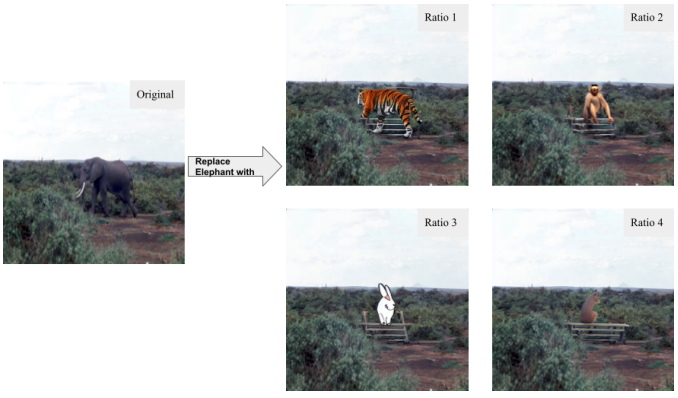

Fig. 8. Result of an end-to-end Pipeline using CLIPSeg and Stable Diffusion: Original Object of Elephant Replaced with Replacing Object Size Ratio Groups 1-4

Another type of shortfall was observed in the case of Object Size Ratio group 3 which scored the lowest in both human evaluation and CLIP scores. Fig 9 shows an example that demonstrates this finding. In this task with object size ratio group 3, rather than replacing the object naturally, the pipeline tends to fill up the original object area even though it misfits the background context. In this aspect, one interesting topic could be explored to provide a way of controlling the object size so that the generated image can fit the general background context with respect to the object sizes.

Fig. 9. Result of an end-to-end Pipeline using CLIPSeg and Stable Diffusion: Original Object of Husky Replaced with a Replacing Object Ratio Group 3

In Table VII, overall results are tabulated for the two metrics considered: Human evaluation ratings and CLIP score. Human evaluation ratings were collected from a survey of 41 responses for a rating of each generated image on a scale of 1 to 5. CLIP scores were normalized with respect to the CLIP score of the original image.

Object Size Ratio	1	2	3	4
Avg. Human Evaluation Ratings (out of 5)	2.56	2.19	1.97	2.32
Avg. CLIP Score (Normalized)	-0.78	-0.73	-1.57	-0.88

TABLE VII
LIST OF REPLACING PROMPTS FOR EACH ORIGINAL OBJECT PROMPT

In terms of the quantitative metrics, human evaluation results showed the highest rating with the smallest size shrinkage, and across the bigger shrinkage, ratings dropped significantly. For the CLIP score, all the replacement experiments reduced the score w.r.t. the original image score, but the object size did not affect the score consistently. This result can be explained by the fact that the CLIP score looks at the individual picture and tries to match it with the caption rather than considering the general context across the set of different images.

Further, we tested replacement tools with DALLE-2 to observe how the experiment can be replicated with other pipelines. Here, the test was run on the OpenAI browser. To avoid the input bias, data sets consisted of randomly pooled two original images and object ratio groups 1 - 3. The DALLE-2 on the browser provides an interface to manually create the mask instead of using an input mask image. The tests were carried out with the original image and a replacement text prompt. The replacement tasks across the object size ratios were done on the same edited version of the original image. Results are tabulated below in Fig. 10.

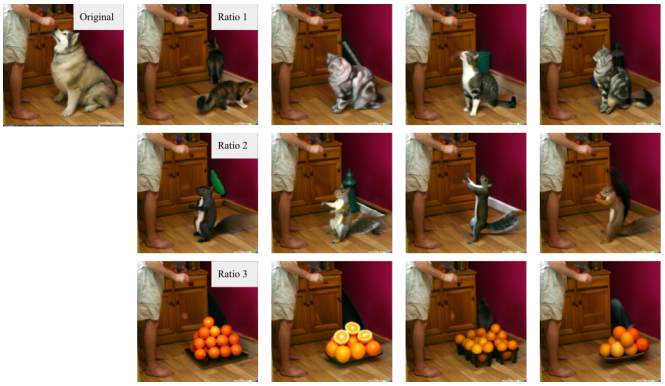

Fig. 10. Results of DALLE-2 on openAI Browser with the Original Object of Husky with Replacing Object Ratio Groups 1 - 3

With the same set of replacement objects, DALLE-2 on the openAI browser showed the better performance out of the two pipelines tested in terms of the qualitative results. Rather than filling up the entire original object space, the replacing objects show appropriate sizing that better matches the background context of the original image. Comparing to the baseline pipeline involving CLIPSeg and Stable Diffusion, this improvement in the performance may be attributed to the generation and in-painting performances within the DALLE-2 tool and the working of the inner modules together. For future work, as discussed above, it was discovered that there is currently no existing solution for controlling the replacing object size. The downfall of this is humans who already have the perception of the objects sizes in the general context of the image may see the general image is unnatural.Future efforts can be made to explore algorithmic ways and one potential path could be to create a program to segment the original object area and in-paint the segments separately.

4) Replacement Prompt Quality: The aim of our study is to examine how the level of contextual details in a replacement prompt affects the quality of the generated image. Prior research has shown that generation models perform better when given more contextual details as input [22]. In light of this, we will be testing various levels of contextual details in the text input, including the object's name and contextual description, while utilizing a fixed input mask image.

To conduct the experiment, we selected 25 images from the COCO dataset and created three different replacement prompts with varying degrees of contextual details for each image. For each generated image, we generated 5 different variants and selected the best one in our survey. Participants were asked to evaluate the quality of the generated images based on ten groups of images generated from these prompts. Our findings indicate that participants were more inclined to choose a simple replacement prompt, suggesting that adding too many contextual details may not necessarily result in better image quality.

To further evaluate the generated images, we developed a CLIP score calculation pipeline to compare the similarity between the replacement prompt and the newly generated image. Our results revealed that both human evaluators and the CLIP score pipeline had a 60% probability of selecting the same image as the best-matched image among the three levels of replacements[Fig 8]. Previous research that created 11.5k model evaluation tasks (240 images * 4 models * 4 samples)[22] also concluded that human evaluation and CLIP score had a 66.8% match probability. Therefore, our findings are consistent with previous research, and we can conclude that human evaluation and CLIP score results are reliable.

We repeated the previous experiment using DALLE-2 to test the finding's generalizability. From the previous 25 images, we selected 5 for evaluation given the limited number of tokens we have for DALLE-2. Subsequently, for each image, we generated three sets of outputs with varying text prompt complexity levels, with four image variations per set. It consistently generated meaningful images even when presented with complex text prompts which Stable Diffusion could not. The model performance is comparably better on the same image with reasonably complicated inputs. However, although OpenAI suggests using the description of the whole image as their input prompt for text-guided inpainting, we found that when dealing with overly complicated text prompts, DALLE-2 also suffers the same fate as Stable Diffusion, where it gets confused by the complex contextual information embedded in the text prompt, and resulting in degraded performance, as seen in Stable Diffusion. An example is shown in Figure 11.

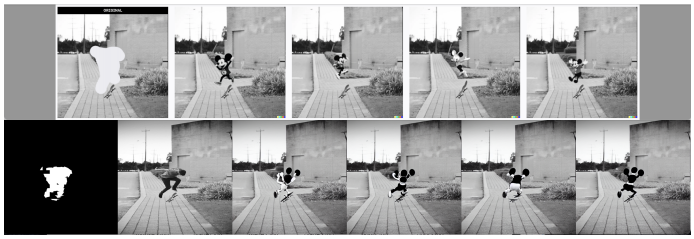

Fig. 11. DALLE-2 generated images when using a fairly complex prompt (up) VS a highly complex prompt in Stable Diffusion model(down)

With the current state of text-guided image inpainting models, we found that inpainting with highly complex input prompts is still a challenging task that even the most state-of-the-art models still struggle with. We believe that improving the performance of such tasks will positively impact the model's utility because it enables the model to handle more complex tasks. One potential approach is to increase the number of parameters in the transformer layers to strengthen the model's understanding of complex mappings between textual and visual representations of images.

VI. Conclusion

Our research project aimed to examine the latest approaches utilized in the field of text-guided image inpainting. Through our comprehensive experimentation, we sought to uncover the techniques employed in recent studies and gain insights into the underlying mechanisms responsible for their effectiveness to improve their performance further. A total of four experiments were specially designed to understand how the nature

of the mask and the text prompts affect the performance of the generative model in various image inpainting pipelines.

In our first set of experiments, we examined the impact of input mask size on the quality of images generated by the image generation module. Our findings revealed that the generative model's performance improves when the input mask size is sufficient to encompass the edges of the entire object while also containing some background pixels to provide additional contextual information. However, we observed that excessively large mask sizes can result in an overabundance of contextual information that can potentially degrade the original object's composition and gesture, leading to a loss of essential details.

The second experiment involved studying the impact of the input-mask quality on the final inpainted image generated. The main conclusion from this experiment is that the model can learn contextual information from pixels outside the replacement object, however, if the mask is too noisy, the performance will be compromised. Each generation model also has their own noise threshold where when the noise level is over this amount, the quality of the generated image will drop drastically.

Next, we designed experiments to analyze the impact of the size of the replacement object in the output image. The results showed that smaller size differences between the original object and the replacing object produced more realistic images, while larger size differences resulted in less realistic images. The human evaluation ratings also showed a similar trend. On the other hand, the CLIP score reduction was inconsistent across different size ratios.

In our final experiment, we investigated the influence of replacement prompts on original images. The results showed that both human participants and the clip score pipeline favored simple replacement prompts. Notably, our stable diffusion experiment produced contrasting outcomes compared to prior studies, which reported improved performance with complex replacement prompts. Conversely, our evaluation of DALLE-2 aligned with previous findings, as it performed better with more intricate prompts. However, both models struggled with highly complex replacement prompts, indicating the models' limitations in handling such prompts.

VII. Limitations

1) Resource Constraints: Since the experiments are highly dependent on graphical computations, they put a significant load on the GPUs in our cloud development environment. We constantly ran into memory issues with the 15GB of RAM available on Google Colabs and had to scale down the experiments per epoch. Due to platform constraints, we also couldn't utilize parallel computing which would have drastically scaled up and accelerated the experiments. For the same reason, we had to limit the size of our dataset to achieve a reasonable overall runtime across all experiments.

2) Model Availability: While we spent lots of effort looking for alternative generative models to examine the generalizability of our findings across different model architectures, we found that the majority of the open-source models available online suffer from poor reproducibility due to either incompatibility issues caused by outdated packages or unavailability of pre-trained weights not released by the developers due to various concerns. Few of the functional generation models we found can perform text-guided image inpainting because it is a relatively new image generation subdomain. With the above resource constraints, DALLE-2 became our only alternative option without retraining the models, which could easily cost thousands of dollars.

3) Metric Availability: We continued to look for a metric suitable to quantify the model's performance in text-guided image inpainting tasks throughout the duration of the project. We were yet to find a perfect metric due to the specificity and complexity of this subdomain. Although we were able to perform qualitative analysis via human evaluations, which has become the standard in this research field, we still would like to measure the model's performance quantitatively to reach statistically significant conclusions. We discuss the possibility of modifying the CLIP score calculation methodology to achieve this goal in the section below.

VIII. Future Work

1) Analysis of other models: One natural extension of our research is to analyze different models. One possible future work could be testing the experiments on other state-of-the-art segmentation models such as Segment Anything[24] as well as other generation models such as Muse[10]. Furthermore, aside from two-stage pipelines, it is also essential to compare the experimental results of other image modification pipelines such as Palette[25].

2) Specific metric for text-guided image inpainting: Due to the limitations of CLIP score, in order to perform further quantitative analysis, we need to find better metrics for measuring image generation quality. A valuable research direction would be to develop a better metric that aligns with human perception and measure the quality of the image generation results. We also propose a modified CLIP score calculation where we isolate the modified object from the image and calculate the CLIP score purely based on the modified pixels and their caption.

3) Experiments: Apart from the future works that can be done on each of the four experiments as mentioned in previous sections, another future research direction would be to design and perform more experiments on new variables. In our experiments, we treated input masks and input prompts as independent variables. However, there might be some correlations between these inputs that can also affect a model's performance. Changing two or more variables in a correlated way may also be a meaningful experiment to perform to discover hidden relations between these inputs.

4) New Pipeline Architecture: Finally, this research aims to improve future text-guided image modification pipelines' performance. The major future work is to develop new image modification pipelines and architectures with awareness of the conclusions drawn from this project.

IX. Contributions

1) Yuhang Qian: From a high level, organized weekly meetings and led project planning and developments throughout the entire semester. Steered conversations to identify the aims and vision of the group and defined tasks for each team member. Created plans to achieve the goals, established responsibilities and objectives among team members, and reviewed and adjusted plans and responsibilities as necessary. For the report, wrote Introduction, Dataset Curation, Metric Selection, and Limitations sections. Performed Input Mask Size experiments and documented the results and discussion in the according section. Performed Replacement Prompt experiments with DALLE-2 and documented the results and discussion in Jiayi's section.

2) Tanmay Jain: Implemented stable diffusion, TDANet, mini-DALLE for the experiments. Contributed to the designing of the input mask quality experiment. Drafted the related works, generative model implementations, and conclusion sections of the final report.

3) Jiayi Zhang: Introduce different metrics which are used in our project. Achieve stable diffusion model experiment implementation and analysis for Replacement Prompt Quality experiment. Finish DALLE-2 text prompt effect experiment with Angelos and Yuhang's help.

4) Angelos Guan: Designed and implemented the mask quality experiment. Wrote mask quality experiment and future work sections and helped edit other sections (limitation, conclusion) in the report.

5) Jimin Yoon: - Plan and conduct Replacement Object Size Ratio experiment with CLIPSeg+Stable Diffusion pipeline, CLIPSeg+mini-DALLE pipeline, and DALLE-2 pipeline on the OpenAI browser. -V. D. 3) Replacement Object Size Ratio in Experimentation and Results. -VIII. Future Work - Replacement Object Size Ratio part. -VI. Conclusion - Replacement Object Size Ratio part

References

[1] 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.

[2] C. Saharia, W. Chan, S. Saxena, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[3] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[4] J. Yu, Y. Xu, J. Y. Koh, *et al.*, *Scaling autoregressive models for content-rich text-to-image generation*, 2022. arXiv: 2206.10789 [cs.CV].

[5] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2017. DOI: 10.48550/ARXIV.1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, 2022. arXiv: 2112.10752 [cs.CV].

[7] T. Lüddecke and A. S. Ecker, *Image segmentation using text and image prompts*, 2021. DOI: 10.48550/ARXIV.2112.10003. [Online]. Available: https://arxiv.org/abs/2112.10003.

[8] C. Wu, Z. Lin, S. Cohen, T. Bui, and S. Maji, *Phrasecut: Language-based image segmentation in the wild*, 2020. arXiv: 2008.01187 [cs.CV].

[9] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord, *Diffedit: Diffusion-based semantic image editing with mask guidance*, 2022. arXiv: 2210.11427 [cs.CV].

[10] H. Chang, H. Zhang, J. Barber, *et al.*, *Muse: Text-to-image generation via masked generative transformers*, 2023. arXiv: 2301.00704 [cs.CV].

[11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 684–10 695.

[12] P. von Platen, S. Patil, A. Lozhkov, *et al.*, *Diffusers: State-of-the-art diffusion models*, https://github.com/huggingface/diffusers, 2022.

[13] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, *Clipscore: A reference-free evaluation metric for image captioning*, 2022. arXiv: 2104.08718 [cs.CV].

[14] B. Yan, Y. Jiang, J. Wu, *et al.*, *Universal instance perception as object discovery and retrieval*, 2023. arXiv: 2303.06674 [cs.CV].

[15] J. Liu, H. Ding, Z. Cai, *et al.*, *Polyformer: Referring image segmentation as sequential polygon generation*, 2023. arXiv: 2302.07387 [cs.CV].

[16] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, *Unleashing text-to-image diffusion models for visual perception*, 2023. arXiv: 2303.02153 [cs.CV].

[17] Z. Wang, Y. Lu, Q. Li, *et al.*, *Cris: Clip-driven referring image segmentation*, 2022. arXiv: 2111.15174 [cs.CV].

[18] M. Yasunaga, A. Aghajanyan, W. Shi, *et al.*, *Retrieval-augmented multimodal language modeling*, 2022. arXiv: 2211.12561 [cs.CV].

[19] Y. Balaji, S. Nah, X. Huang, *et al.*, *Ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers*, 2023. arXiv: 2211.01324 [cs.CV].

[20] Y. Li, H. Liu, Q. Wu, *et al.*, *Gligen: Open-set grounded text-to-image generation*, 2023. arXiv: 2301.07093 [cs.CV].

[21] H. Chang, H. Zhang, J. Barber, *et al.*, *Muse: Text-to-image generation via masked generative transformers*, 2023. arXiv: 2301.00704 [cs.CV].

[22] S. Wang, C. Saharia, C. Montgomery, *et al.*, *Imagen editor and editbench: Advancing and evaluating text-guided image inpainting*, 2023. arXiv: 2212.06909 [cs.CV].

[23] L. Zhang, Q. Chen, B. Hu, and S. Jiang, "Text-guided neural image inpainting," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1302–1310.

[24] A. Kirillov, E. Mintun, N. Ravi, *et al.*, *Segment anything*, 2023. arXiv: 2304.02643 [cs.CV].

[25] C. Saharia, W. Chan, H. Chang, *et al.*, "Palette: Image-to-image diffusion models," *CoRR*, vol. abs/2111.05826, 2021. arXiv: 2111 . 05826. [Online]. Available: https://arxiv.org/abs/2111.05826.

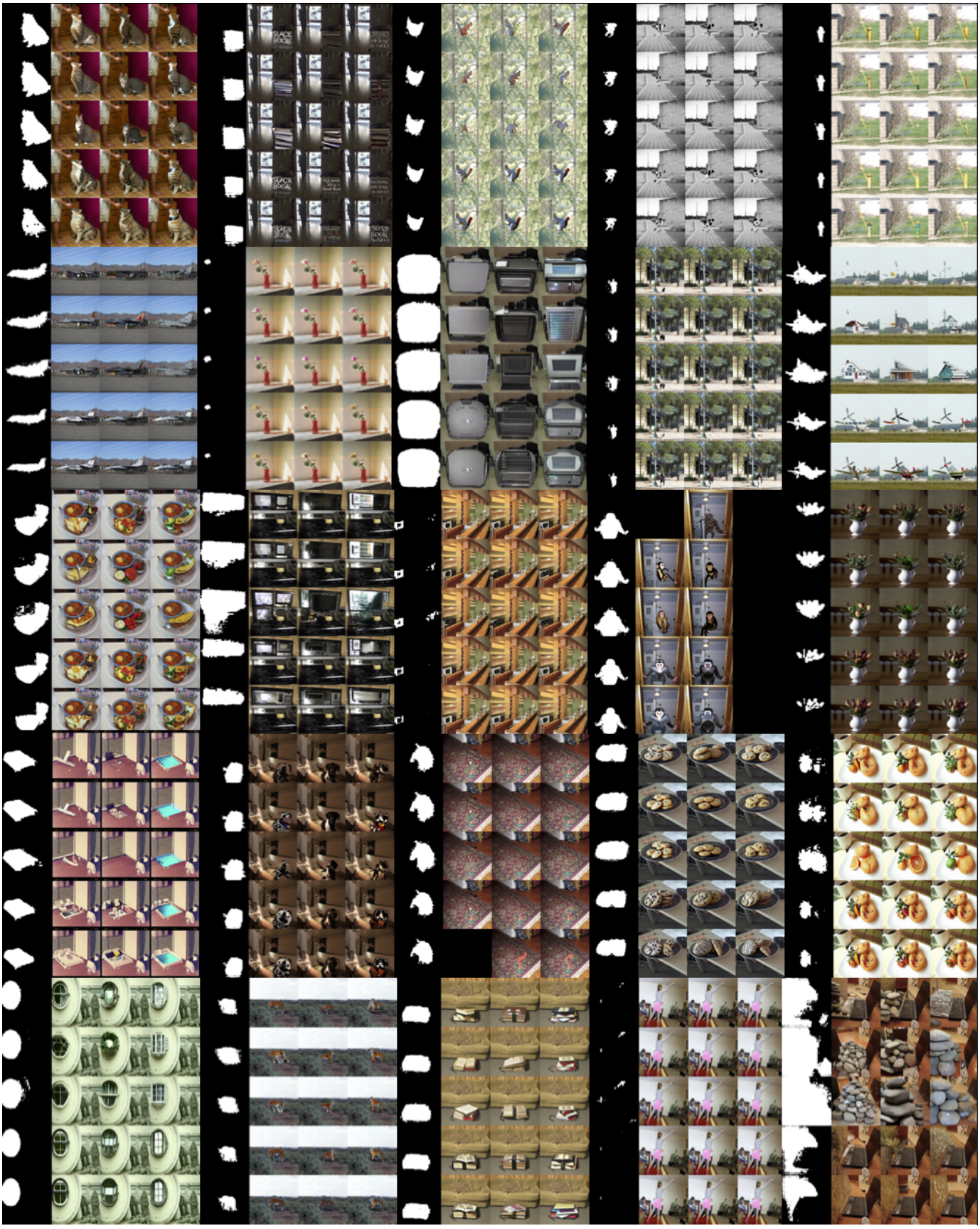

Fig. 12. Stable Diffusion generated images for Input Mask Size Experiments

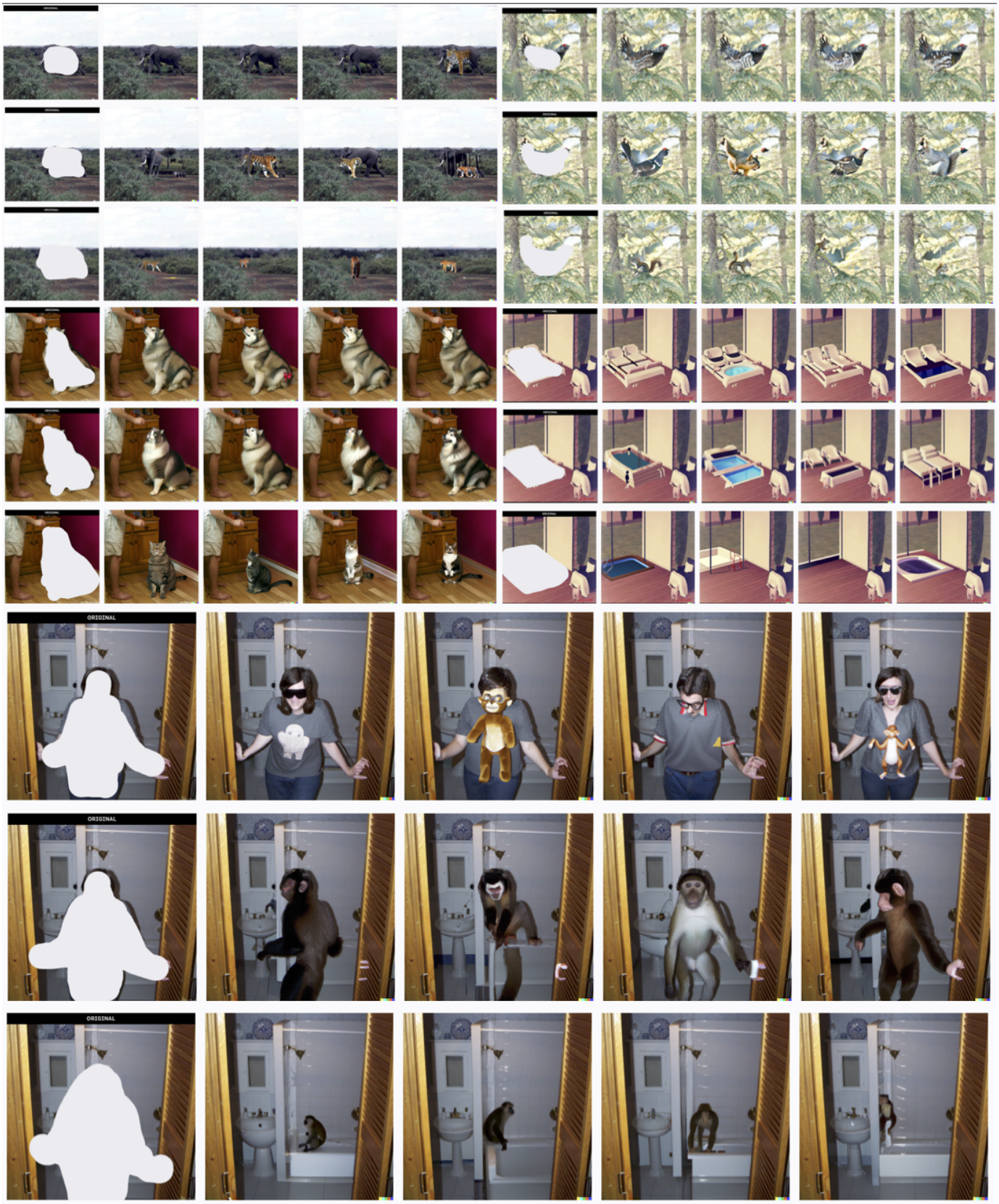

Fig. 13. DALLE-2 generated images for Input Mask Size Experiments

Image	deafult	15low	Improvement	30low	Improvement	15high	Improvement	30high	Improvement
0	22.71652603	22.77383995	0.2523005454	22.70689392	-0.042401336	22.59813118	-0.521183791	22.7772185	0.2671731989
1	20.44716581	20.57681084	0.6340488991	20.41677793	-0.148616579	20.43350538	-0.066808432	20.43055598	-0.081232909
2	21.18029849	21.17663892	-0.017278161	21.0531044	-0.600530189	21.16927592	-0.052041607	21.19100634	0.0505557333
3	21.6062603	21.62085152	0.0675323579	21.5066967	-0.460809030	21.6053791	-0.004078424	21.55321948	-0.245488213
4	21.92153422	21.94713465	0.1167821243	21.95699247	0.1617507539	21.8922081	-0.133776685	21.99278005	0.3250038458
5	25.68531736	25.50287882	-0.710283349	25.96682231	1.095976137	25.755771	0.2742953993	25.60380236	-0.317360277
6	20.92285983	20.89592234	-0.128746667	20.88898532	-0.161901919	20.900177	-0.108411689	20.89786593	-0.119457362
7	20.43386014	20.40546926	-0.138940387	20.42317454	-0.052293608	20.45059586	0.081901865	20.42989477	-0.019405916
8	20.30462774	20.23496819	-0.343072288	20.16062037	-0.709234204	20.10737419	-0.971470876	20.38652484	0.4033420404
9	22.04743067	22.06651751	0.0865717101	22.09265264	0.2051121747	22.11906751	0.3249214695	22.10277494	0.2510236432
10	23.73561859	23.86035474	0.5255230641	23.8383948	0.4330041331	23.89307022	0.663355914	23.88164584	0.6152241053
11	21.15716426	21.03574053	-0.573913024	21.12719472	-0.141651938	21.20059077	0.205256778	21.19913673	0.1983842339
12	23.55136426	23.45380084	-0.414258064	23.43238004	-0.505211592	23.32528559	-0.959938739	23.31199265	-1.016381111
13	22.04207357	22.06083171	0.0851015126	21.99391301	-0.218493749	22.01437314	-0.125670682	21.97674561	-0.296378478
14	22.57131577	22.74881935	0.7864122219	22.80891864	1.052676203	22.48307165	-0.390956917	22.46875	-0.454075607
15	21.77666728	21.7051843	-0.328254894	21.75873693	-0.082337432	21.81336276	0.1685082473	21.75190544	-0.113708105
16	21.60218112	21.57411385	-0.129927949	21.65695953	0.2535781765	21.6581103	0.2589052636	21.66362762	0.2844458504
17	23.9569238	23.95084	-0.025394772	23.97026443	0.055685914	23.97467295	0.074087769	24.11967595	0.6793532961
18	20.53974024	20.48552704	-0.263942997	20.59149297	0.2519638789	20.48395856	-0.271579297	20.4911588	-0.236524126
19	21.30798912	21.34249624	0.1619444964	21.33623568	0.1325632444	21.21587944	-0.432277676	21.24524689	-0.294454032
20	23.89566803	23.77898407	-0.488305913	23.81076113	-0.355323382	23.92991829	0.1433325043	23.94752439	0.217011547
21	19.63294284	19.62898509	-0.020158712	19.59941037	-0.170796914	19.60644786	-0.134951647	19.56163661	-0.363196843
22	18.75268555	18.51820183	-1.25040074	18.51981036	-1.241823138	18.81361008	0.3248842939	18.95900408	1.100207951
23	24.72804387	24.71931712	-0.035290926	24.71994972	-0.032732681	24.72232183	-0.023139905	24.71621704	-0.047827612
24	20.98707136	20.78793271	-0.948863412	20.87285423	-0.544226111	20.87285423	-0.544226111	20.96655273	-0.097767908
AVG			-0.1240326132		-0.07304292776		-0.08884255914		0.0275253995
MAX			0.7864122219		1.095976137		0.663355914		1.100207951
MIN			-1.25040074		-1.241823138		-0.9714708769		-1.016381111
MEDIAN			-0.03529092646		-0.08233743247		-0.05204160723		-0.04782761285

Fig. 14. Calculated CLIP Score % change for Input Mask Size Experiments

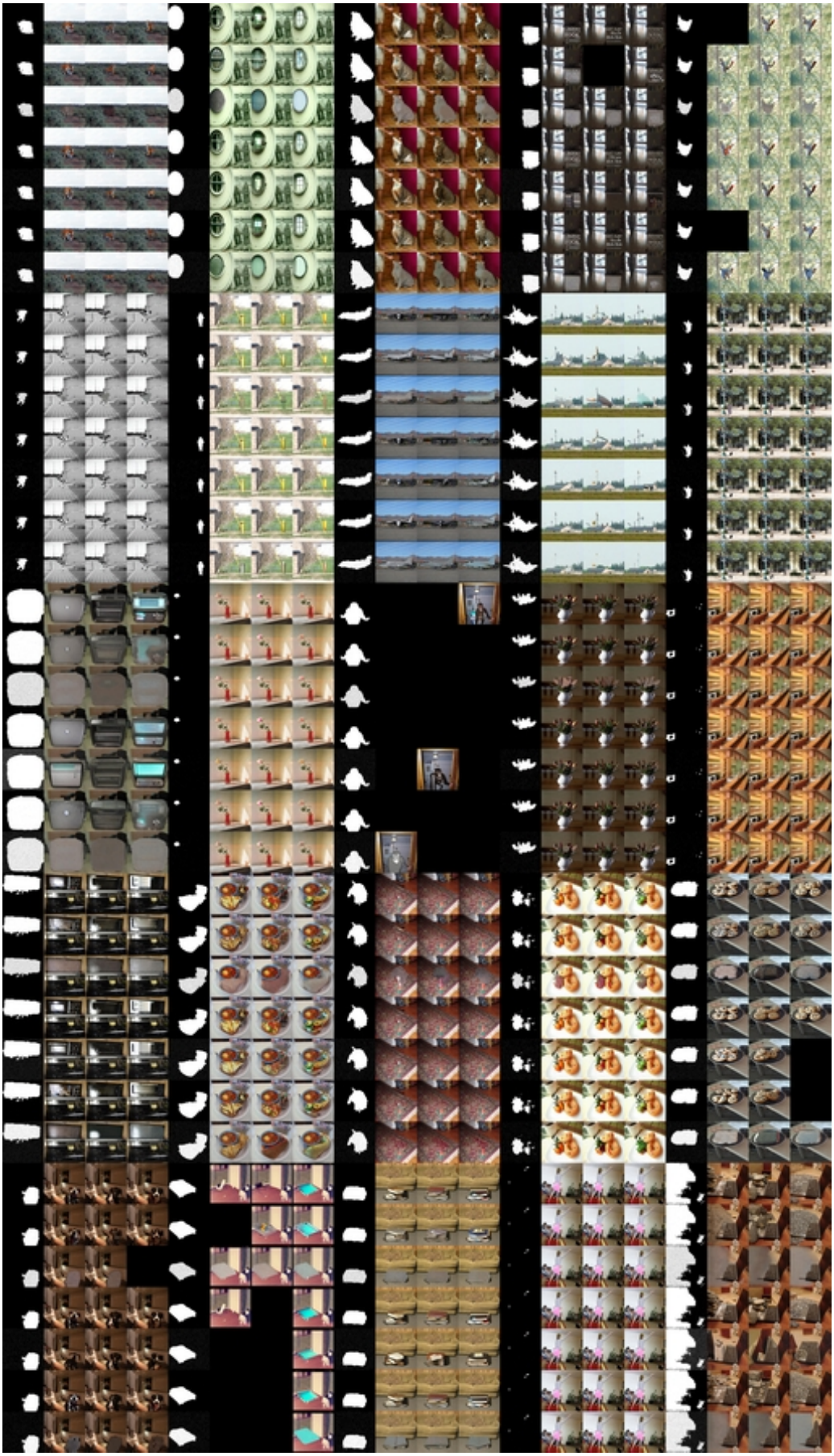

Fig. 15. Stable Diffusion generated images for Input Mask Quality Experiments with high and low noise for segmented pixels, background pixels and mixed pixels

		generated image	clipscore	clipscore change rate		generated image	clipscore	clipscore change rate
2	original mask		21.973767		mode = 1, noise_level = 0.01		21.961174	0.000573
3	mode = 0, noise_level = 0.0001		21.958827	0.00068	mode = 1, noise_level = 0.1		21.979589	-0.000265
4	mode = 0, noise_level = 0.001		21.965988	0.000354	mode = 2, noise_level = 0.0001		21.973589	0.000008
5	mode = 0, noise_level = 0.01		21.879673	0.004282	mode = 2, noise_level = 0.001		21.957211	0.000753
6	mode = 0, noise_level = 0.1		21.971684	0.000095	mode = 2, noise_level = 0.01		21.925394	0.002201
7	mode = 1, noise_level = 0.0001		21.973509	0.000012	mode = 2, noise_level = 0.1		21.969875	0.000177
8	mode = 1, noise_level = 0.001		21.974722	-0.000043				

Fig. 16. Full set of input Mask Quality Experiments on one input

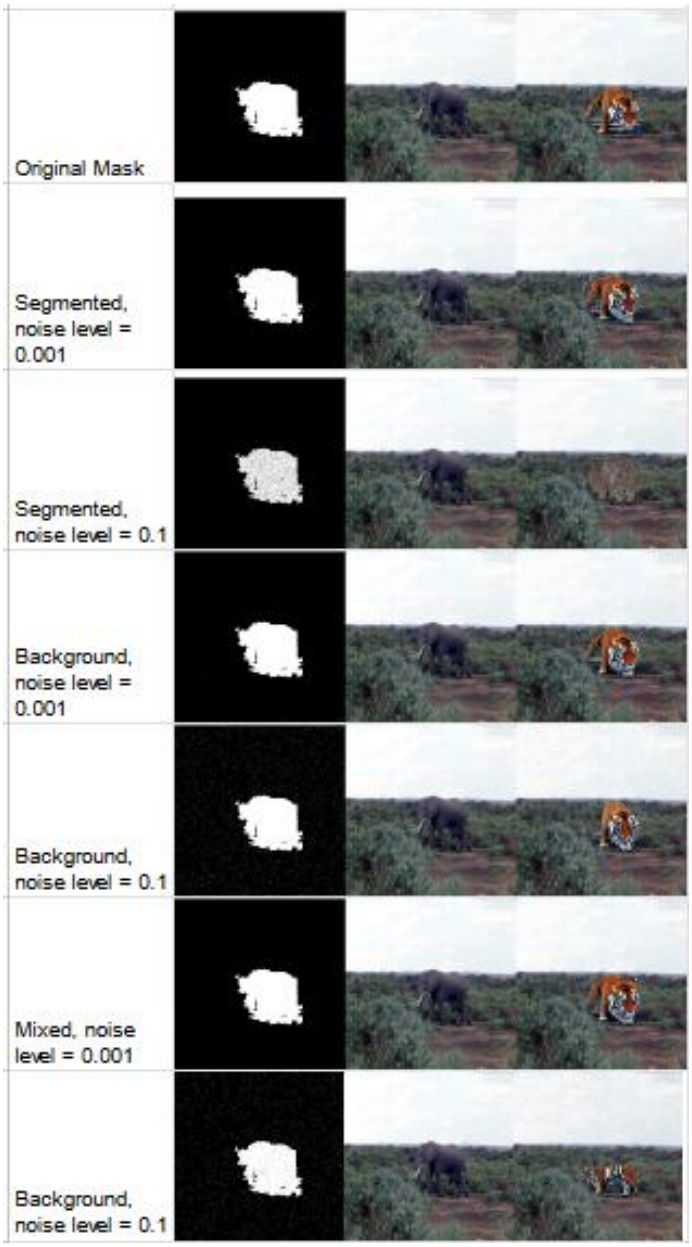

Fig. 17. Qualitative Example For Input Mask Quality Experiment on one input

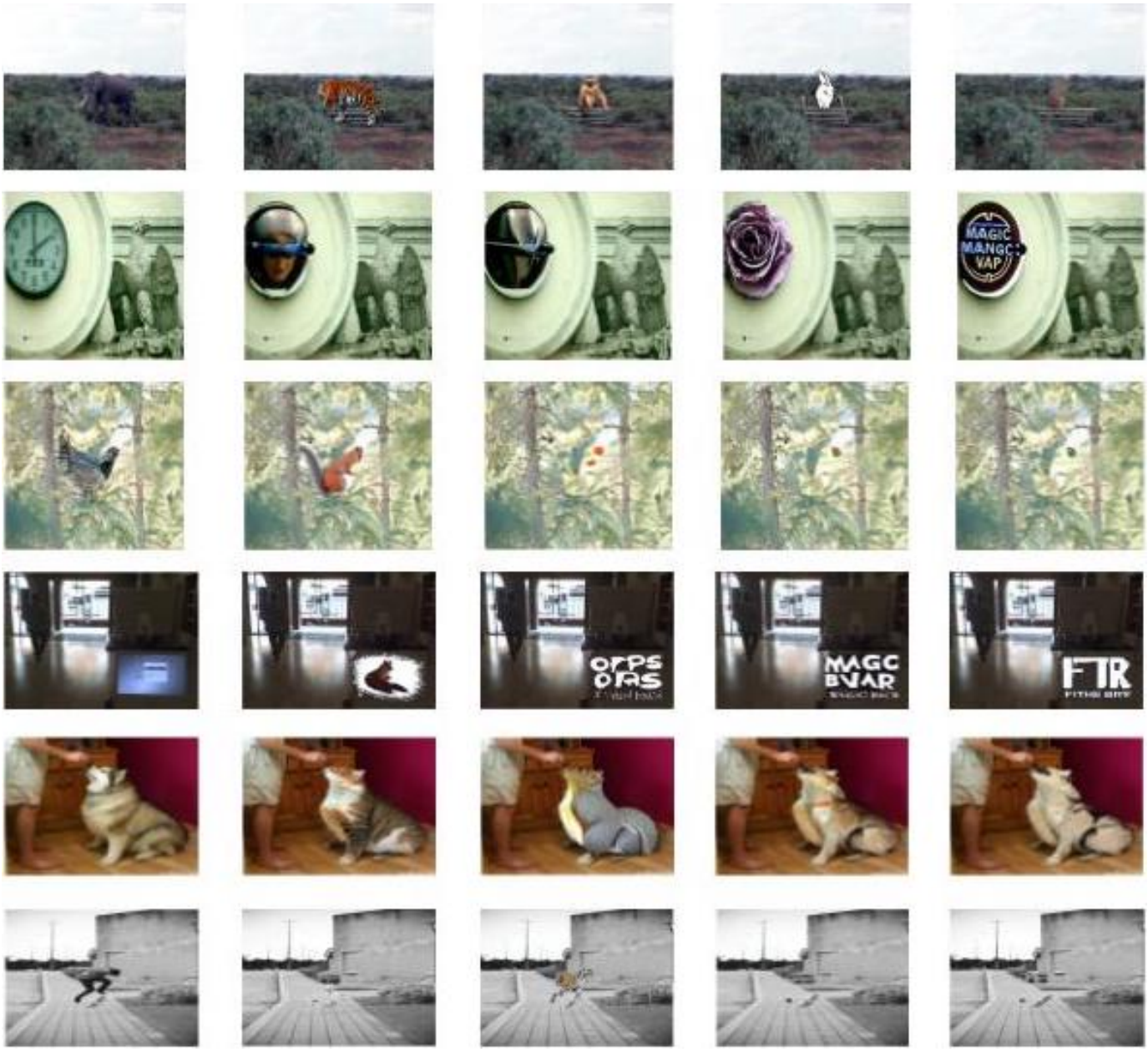

Fig. 18. Result of a end-to-end Pipeline using CLIPSeg and Stable Diffusion with Varying Replacing Object Size Ratio w.r.t. the Original Object Size

Fig. 19. Text prompt experiment survey responses