Why do I need to speak to my drone?

Jose Martinez Carranza*, Delia Irazu Hernandez-Farias, Leticia Oyuki Rojas-Perez, and Aldrich Alfredo Cabrera Ponce†
Instituto Nacional de Astrofisica Optica y Electronica (INAOE), Puebla, Mexico
† Benemerita Universidad Autonoma de Puebla (BUAP), Puebla, Mexico

ABSTRACT

Seeking to enhance the autonomous performance of a delivery drone, which is envisioned for delivering parcels to homes, offices, and other urban and rural locations, we present a study where we evaluate the use of state-of-the-art Artificial Intelligence (AI) tools, specifically language and generative models such as Diffusion Models, CLIP, and ChatGPT, to address the challenge of recognising unfamiliar places based on textual descriptions provided by customers. This concept draws inspiration from contemporary e-commerce practices, where customers typically provide textual descriptions of delivery destinations. Such descriptions serve as additional guidance to assist the delivery person in reaching the intended location. Our objective is to investigate how to process such textual descriptions to be comparable with the observations made by a delivery drone. Based on our preliminary assessment, these AI models can leverage the autonomous behaviour of a drone that is required to navigate to unfamiliar destinations, relying on images captured by its onboard camera and the textual description provided by the customer to accomplish this goal.

1 INTRODUCTION

When the Amazon Primer Air program was launched in December 2013, the expectations for drone delivery rapidly grew. Today, the technology is ready, and drones can arrive at a location autonomously using a GPS coordinate to guide them toward the destination. Unfortunately, in USA, as much as in most of the countries of the Commonwealth, the regulations have inhibited the widespread use of this technology [1], and that has made it difficult to carry out a realistic evaluation of how effective this technology would be in real life delivery scenarios, especially in developing countries where urban organisation tends to be less ordered, let alone rural areas.

Poorly designed and maintained neighbourhoods could complicate the transportation of goods, especially in the stage known as the last-mile delivery, where parcels are taken from a distribution centre to the last destination provided by the customer (home or business address) [2]. The last-mile delivery is expected to be the most challenging and costly stage of the logistics behind the delivery of goods in e-commerce [3].

Among the different problems in last-mile delivery, an inaccurate or wrong delivery address is a recurrent one, especially when the customer is away from home to assist the courier [4]. One can expect this problem to worsen if the courier is unfamiliar with the area or has never been to the delivery place. For these reasons, it is a common practice to request the customer with a textual description of the delivery place, soliciting references or landmarks that could help the courier recognise the delivery location more quickly. Under this scenario, even a description of the appearance of the place and objects around it, such as trees, cars, and lampposts, are expected to be provided by the customer if they consider that these are distinctive landmarks that would help to recognise their place.

If, in the near future, companies plan to utilise autonomous drones to assist in the last-mile delivery stage, then we can anticipate that these delivery drones will encounter the same aforementioned issues, in particular, the recognition of the places where it has never been before. In this respect, a delivery drone could also exploit the customer's textual description the same way a delivery person uses it to find a delivery place. However, to achieve this, the textual information has to be represented so that it can be compared with observations of the scene made by the delivery drone. As it has been widely investigated, onboard cameras are low-cost sensors that can produce rich visual information [5], and it has become common now for a drone to have at least one onboard camera.

Therefore, a broader question is, what is the shared space where textual and visual information can be mapped such that these can be compared in order to recognise that what is being described is similar to what is being visually observed? Motivated by the latter, in this work, we explore the use of state-of-the-art language and generative models to address this question. First, we assume that a customer provides a textual description of a delivery place (denoted as *target*). Then, we propose to use a generative model known as the Stable Diffusion model [6], to generate images from a textual description that can be compared directly against the target. Figure 1 shows some examples of images generated by the Diffusion Model from textual descriptions provided by a subject. However, like any other generative model, an image gener-

*Email address: carranza@inaoep.mx

Figure 1: Target images and their synthetic images generated with Stable Diffusion using prompts provided by human subjects in this study. Each column shows the best synthetic image generated with a prompt provided by subject 1 in this study, in total, 10 prompts. The best image has the highest similarity score that compares the image embeddings of the target and synthetic images using the cosine distance.

ated with Stable Diffusion can not be compared directly with the image of the delivery place at the pixel level such as it is typical for place recognition methods based on visual descriptors or neural networks [7, 8]. For instance, if the customer inputs "a red car", then the Diffusion Model is likely to generate a red car visually dissimilar to a red car observed in the drone's camera image.

A visual comparison may not be possible between an artificially generated image and a real camera image. However, at the semantic level, if a red car exists, we want a similarity measure to reward it. Thus, the more semantic concepts are found between these images, the higher the similarity score. Therefore, we propose to explore the use of the Contrastive Language-Image Pretraining model (CLIP) [9], a neural network model that has learned to associate images with textual descriptions through the generation of numerical representation known as *embedding*, where images and textual descriptions are represented in a shared space at the semantic level.

The Diffusion Model used in this work is still slow to be used in real time. Nevertheless, we have designed a set of experiments to get insights into its potential use and discuss some avenues of research on how it could be exploited in tandem with language models such as CLIP and ChatGPT.

2 Related Work

Recently, there is a strong trend in developing multimodal approaches which aim to take advantage of more than one modality of information (visual, textual, speech, etc.). Among these, there is one combining the capabilities of Computer Vision and Natural Language Processing commonly denoted as *Vision-and-Language models*[1] which covers many very challenging tasks such as i) *image captioning* aiming to generate a natural language description from an input image, ii) *visual question-answering* which objective is to find answers by means of a question in natural language and a related image, iii) *image retrieval* which aims to retrieve the data in a given modality by the cues provided in another modality, iv) *Phrase grounding* which involves object detection from an input image and a phrase in natural language, and v) *image generation* which aims to generate an image from the information provided by a textual description in natural language.

Generating an image from a textual input is a challenging task that has been addressed from different perspectives. Generative Adversarial Networks (GANs) have been widely exploited for image generation demonstrating an impressive performance [10]. Transformers-based architectures, which have proven to be very powerful for addressing several natural language processing tasks, have also been used as a way to generate images from textual descriptions [11]. State-of-the-art has been outperformed by Diffusion models[2], which are inspired by non-equilibrium thermodynamics [12]. In particular, some of these models have attracted attention beyond the research community as they have been made publicly available online through simple interfaces causing their extensive

[1] https://huggingface.co/blog/vision_language_pretraining
[2] https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

use in many respects. Examples of these models are *Imagen*[3] and *DALLE-2*[4]. Another model of this type is *Stable Diffusion*[5] [6] which besides allows the users to access the source code and weights of the model[6].

There are other visual-and-language models capable of performing more general tasks. CLIP allows us to determine if an image and a textual description can be paired since it was pre-trained in "image-text" pairs. For doing so, both inputs are converted to an embedding representation allowing us to compare them by means of cosine similarity. On the other hand, there is VisualBERT a model for learning joint contextualised representations of vision and language [13] enabling them to capture semantics between images and texts.

A very important aspect to be considered in *text-to-image* tasks is how to evaluate the model's performance. Some options are to quantitatively measure the alignment between the images generated and the textual descriptions or to measure if the objects mentioned in the textual description are detected in the generated image [14]. However, the inherently subjective nature of textual descriptions made in natural language has made it important to assess these systems by considering human judgments on the generated images [15].

Vision-language models can have a wide range of applications [16] in different AI areas like robotics, where language and vision can provide very useful information for improving the understanding of the environment where a robot must perform. However, related work combining vision, language, and robotics is scarce. Vision-and-Language Navigation attempts to provide communication skills between humans and agents for navigating in 3D environments [17]. The use of language has also been used as a tool for monocular depth estimation by means of CLIP [18] and also by processing spoken language together with object recognition [19]. Taking advantage of DALL-E, in [20] DALL-E-Bot is presented, it is an autonomous robot able to rearrange objects in a scene by inferring a textual description, then generating an image representing a human-like arrangement of these objects, and finally performing physically the movements.

To the best of our knowledge, diffusion models have not been used for helping on autonomous drone navigation. Our proposal is aimed to contribute in this direction by evaluating the performance of such models to generate images from a textual description provided by users and enable a direct comparison with target images that could be captured with the drone's onboard camera.

3 METHODOLOGY

The CLIP model represents an image or a textual description via a numerical embedding vector that can be compared using a similarity score based on the cosine distance (with values between -100 and 100). In a straightforward manner, we could compute the embedding of the textual description provided by the customer and compare it against the computed embedding of a target image. We explored this approach in the first part of the experiments, and it will be noted that the similarity score between a textual description and a target image reaches an average of 30. In contrast, two embeddings of the same image achieve a score of 100 and if the image begins to change in appearance, the less similar, the smaller the score with a tendency towards zero.

We also observed that different textual descriptions with significant changes for the same target image produced scores with no significant difference. This would make it difficult to assess whether a textual description is better than another or the opposite, this is, whether one image corresponds better to a textual description than another. For this reason, we propose to map the textual description to the visual space through the use of the Stable Diffusion model. Thus, an image embedding can be computed with CLIP and compared directly with the corresponding image embedding of the target image. In this work, textual and image embeddings are numerical vectors of 512 float numbers.

For the comparison, we employ the cosine distance between two vectors as proposed in the CLIP methodology. Suppose we have an embedding for a target image $\mathbf{e_t}$ and an one for the image generated with Stable Diffusion $\mathbf{e_s}$, then the similarity score is (Note that for the sake of interpretation, the cosine distance is multiplied by 100):

$$score = 100.0 \times \cos(\mathbf{e_t}, \mathbf{e_s}) = 100.0 \times \frac{\mathbf{e_t} \cdot \mathbf{e_s}}{\|\mathbf{e_t}\|\|\mathbf{e_s}\|} \quad (1)$$

In sum, given a textual description, we use Stable Diffusion to generate a synthetic image, which is expected to contain semantic information described in the text and this is why we can compare it against a target image at the semantic level using CLIP.

4 EXPERIMENTAL FRAMEWORK

To evaluate our proposal, we assume that we have target images depicting different outdoor areas representing delivery locations, and then we asked subjects to provide textual information about the target images by only looking at them. The subjects had not seen the images before, and only at the moment of the request can they pay attention to the image and write down their description. We seek to emulate the scenario where a customer can visually observe a delivery place and describe it in text. Then, the AI tools can use this textual description to generate an image that will resemble what the subject is visualising so that the delivery drone can use such an image to search out the delivery place even if it has never been there before.

For our experiments, we pulled out 5 images from the internet with no particular appearance in mind: 1) a house with

[3]https://imagen.research.google/
[4]https://openai.com/dall-e-2
[5]https://stablediffusionweb.com/
[6]https://github.com/CompVis/stable-diffusion

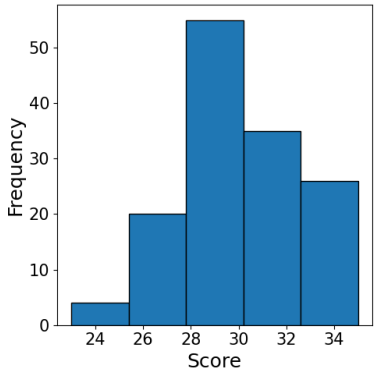

Figure 2: Distribution of scores obtained with text embeddings from subject prompts compared to target image embeddings.

cars parked in front of it; 2) a food truck selling hot dogs; 3) a kiosk; 4) a basketball court; 5) a house with a swimming pool. Next, we asked three subjects to look at these images in order to provide textual descriptions of the places, which we will refer as "prompts". It is important to mention that the subjects never used *text-to-image* generative models before and that they were not provided with any guidelines for doing this task. Each prompt is passed to the Stable Diffusion model tuned to generate 10 images per prompt to have more than one synthetic image per prompt. Under these settings, we propose two types of prompt generation: i) subject 1 will provide 10 prompts for each one of the target images before using Stable Diffussion; ii) subject 2 and 3 will provide a prompt waiting to see the 10 synthetic images generated by Stable Diffusion and use it as feedback to improve the following prompt; they will continue this iteration until 10 prompts for each target image have been written. This variation in the prompt generation is intended to observe whether Stable Diffusion can improve its output through interaction with subjects 2 and 3; or whether just good prompt writing is enough. Note that for each subject, 10 prompts per target image will be written down, thus generating a total of 100 images per target image. Finally, in a step forward towards what is known as "prompt engineering", we also propose to use ChatGPT-4, a Large Language Model (LLM) [21], to generate the best textual description possible out of the 10 prompts generated by each subject for each target image and pass it to Stable Diffusion. The goal is to assess whether a powerful LLM can contribute to the generation of better synthetic images.

4.1 Comparison using Text Embedding

Before describing the results obtained with images generated with prompts via Stable Diffusion, and as mentioned in Section 3, we evaluated a direct comparison of the prompts given by the subject against the target images. Our goal was to assess the capacity of the text embedding to represent a textual description of a place rather than just short sentences describing a few objects, as it has been a common practice in previous works using CLIP. It is worth noticing that the average number of words in subject prompts is 32 words; CLIP fails to compute an embedding for prompts larger than 80 words. Thus, embeddings for the 150 prompts generated by the subjects were computed and compared against their corresponding target image using the similarity score (Eq. 1). The distribution of the scores is shown in Figure 2. Note that the average value is around 30, with 23 for the lowest and 35 for the highest, which makes it difficult to assess whether one textual description could be better than another if these are directly compared against the image embedding. Hence, we propose comparing image embeddings where visual information may encode better semantic information.

4.2 Results from Subject Prompts

Figure 1 presents a mosaic of the best images generated by Stable Diffusion, as determined by the similarity scores in comparison to the target image (displayed in the first column). These images were produced using prompts from Subject 1. The subsequent column showcases the best image from the ten synthetic images generated by the first prompt. In a similar fashion, the next column represents the best image from the second prompt, and this pattern continues. Each row across all prompts represents the best synthetic images corresponding to each target image. Upon observing these images, one can note their striking similarity to the target image. This indicates that Stable Diffusion when paired with CLIP, is capable of generating and identifying viable candidate images, potentially useful in delivery location recognition, as demonstrated by their visual quality.

Due to the lack of space, we cannot include the respective mosaics for subjects 2 and 3, but we can state that the quality in similarity holds. We can summarise the score distribution of the 100 synthetic images generated for each target image per subject as shown in Figure 3. Establishing which subject produced the best prompts is complicated since the distributions resemble each other. Note that subject 1 created very good prompts for the fifth target image, the house with a pool, whereas subject 3 did it for the food truck, the second target image. We must recall that subject 1 wrote their prompts first and then used Stable Diffusion, and subjects 2 and 3 produced a prompt, had a look at the resulting 10 synthetic images, and used their judgment to propose a following prompt, seeking to improve the next batch of synthetic images. However, it seems this did not have a significant effect and is worth investigating with more subjects and images in a future study. Meanwhile, we should highlight that the best values reach around 80 and, for some images, a score of 90. By looking at the distributions, one may think that the synthetic images exhibit a score of 70 on average. Nevertheless, the important fact is that with our proposed methodology, we are able to generate and identify synthetic images with scores

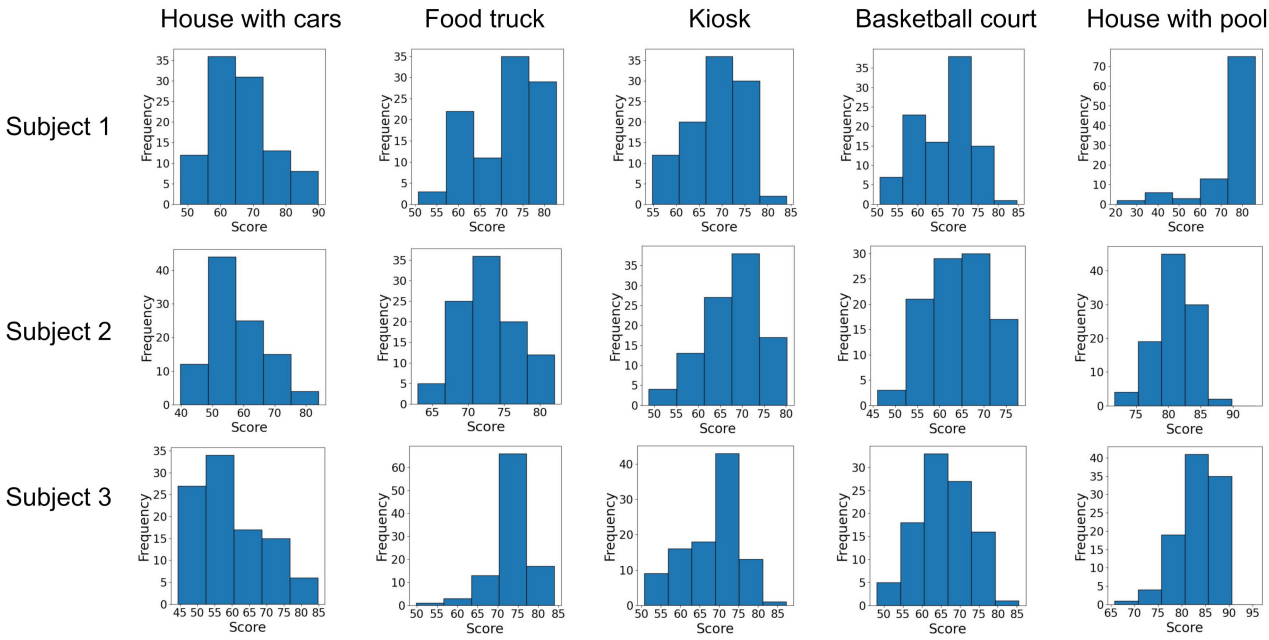

Figure 3: Score distribution for all the images generated with Stable Diffusion using the subject prompts for each target image.

bigger than 80 with a notable visual similarity.

Inspired by the literature assessing the performance of generative models by means of human judgements, we also asked all subjects to choose the image that, in their opinion, was most similar to the target image. To do this, given that for each prompt, Stable Diffusion generates 10 images; then the subject selects 1 out of these 10 images marking it up as the most similar to the target image before applying the evaluation with the CLIP embeddings. We found that the user coincided in 30% with the most similar image found via CLIP embeddings and the similarity score. However, if we take the selected image by the user with the highest score out of the 100 images per target image, then this percentage increases to 47%.

To summarise and illustrate the above, for each target image and using the data of subject 1, Figure 4 shows the image selected by the user with the highest score shown under the "Subject 1 Selected" column and the prompt passed to Stable Diffusion to generate it. The image with the highest score is shown under the "CLIP" column and the corresponding prompt. It stands out that in 3 out of 5 images (60%), subject 1 coincides in their choice of the most similar image with the highest score. Moreover, the similarity between images is also notable for the other two target images. For completeness but also to save space, Figure 5 only shows the images in the same fashion, this is, those selected by the user versus that of the highest score for each target image. Note that in some cases, the resemblance between the synthetic and target images is uncanny.

4.3 Subjects' Prompt Analysis

The rapid widespread use of generative models (like the ones mentioned in Section 2 or even ChatGPT-4) has led to the emergence of a new research area denoted as "prompt engineering". It involves the practice and skills for writing effective prompts for generative models. A very important aspect of obtaining the desired images from a textual description is the ability to select the correct words. However, this is not the only key to warrant the performance of generative models, other aspects like a certain format and style, adding keywords, and key phrases have been recognised as a way to improve prompts. People having more experience with writing prompts usually execute the prompt, assess the result, and try to adapt the prompt to improve the obtained output [22]. These modifications are often called "prompt modifiers" a taxonomy of six different types of them is proposed in [23].

Despite its implications for improving the use of generative models, there are only a few available resources helping and guiding users on these tasks for example on the use of OpenAI's documentation[7] or the online templates created by the AI-generated art community [24]. Unfortunately, for a task like the one addressed in this paper, there are no available resources for guiding in a prompt generation.

As already mentioned, in the experiments carried out, we rely on the subjects' backgrounds for creating the prompts describing the target images. A qualitative analysis on the basis of two of the types of "prompt modifiers" in the aforementioned taxonomy[8] was done. All the prompts contain at least

[7] https://platform.openai.com/examples

[8] The remaining aspects were not considered since they are strongly re-

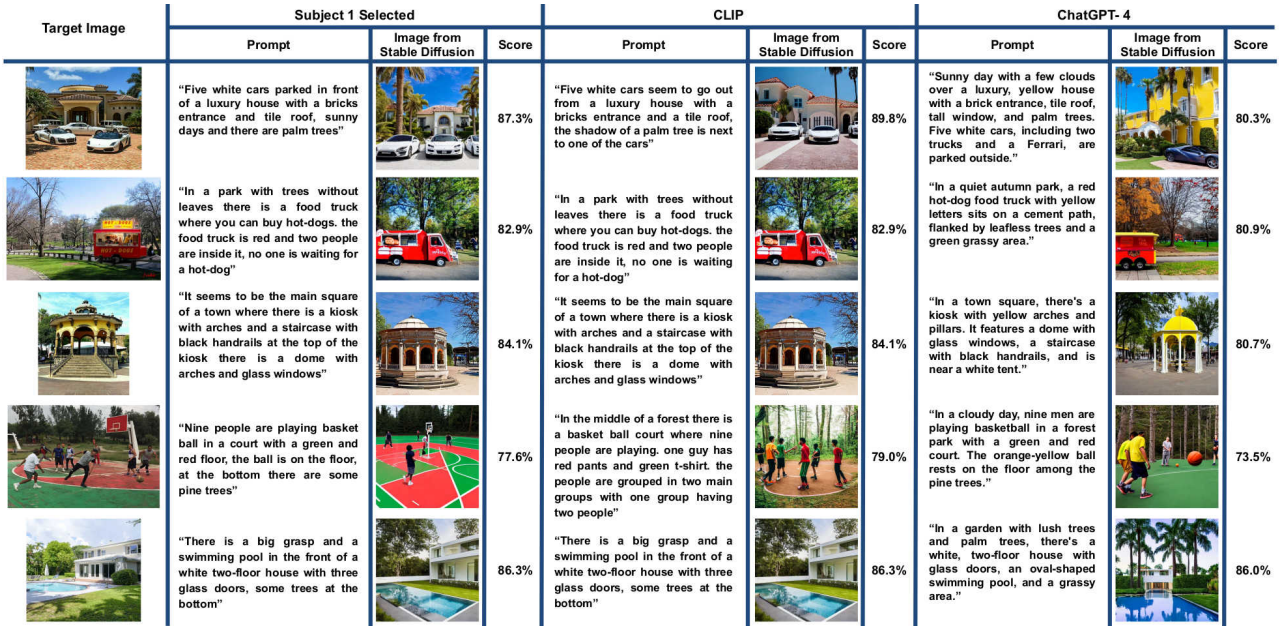

Figure 4: Images generated with Stable Diffusion using the prompts from subject 1 in this study. We compare the image with most similarity to the target image in three modes: 1) selected by the subject (Selected); 2) with the highest score (CLIP); 3) generated with the prompt produced with ChatGPT-4 whose image score is the highest.

a mention of the most salient *"subject terms"* in the target image (i.e., "pool", "basketball", etc.) and in some cases, these terms were written twice in a given prompt *"repetition"*. It is also interesting to note that, one of the subjects who performed an iterative process during prompt generation wrote some consecutive prompts by adding more terms at the end of the phrase, while the prompts generated by the subject who never received feedback from the outcome images are very different between them, in this case, it seems that this subject intentionally tries to describe the target image in a very different way each time.

As mentioned before, we are interested in better taking advantage of prompt engineering for improving the prompts and hence the generated images, we use human-created prompts for feeding the following textual prompts to ChatGPT-4: a) *I have 10 descriptions of a place:* [here the list of human-created prompts were added] *for each one of the 10 description, could you tell me the average number of words?*; and b) *ok, then given these 10 descriptions, could you mix them to generate the best description whose number of words is around* [NUM] *words?* (where NUM was replaced by the average length of the prompts generated by each subject). As output, we obtained a new prompt generated according to the profile of each subject. Afterward, we generated 10 images per prompt that were evaluated in the same way as human-created prompts. Table 1 shows the best scores obtained for each target image by both the subjects and the ChatGPT-4. As it can be observed, for all cases, subject's prompts outperform the scores of ChatGPT, however it still shows a very competitive performance since the obtained scores are very similar to the prompts generated by humans. Interestingly, ChatGPT shows the most salient drop in the score rate when compared with subject 2, which prompts are the shorter and therefore the summary of them are even shorter provoking to lose important details for generating an effective prompt. We are planning to further explore how to better leverage the capabilities of LLMs for applying prompt engineering as an auxiliary tool for generating images. Due to the lack of space, the images with the highest score for each target with the summary prompt are shown in Figure 4 for subject 1 and in Figure 5 for the remaining subjects.

Table 1: Comparison of the best scores for synthetic images generated with Stable Diffusion using subject prompts versus the best score obtained with synthetic images generated with a prompt from ChatGPT-4 that combined all the prompts of a corresponding subject.

Target Image	Best Score					
	S1 Best	GPT-4 (S1)	S2 Best	GPT-4 (S2)	S3 Best	GPT-4 (S3)
1	**89.84**	80.03	84.08	64.99	84.77	83.84
2	82.86	80.91	83.35	79.00	**83.99**	79.30
3	**84.08**	80.71	80.27	75.93	81.20	77.39
4	79.00	73.49	77.69	64.21	**79.30**	66.85
5	86.28	86.04	89.84	84.52	**90.62**	90.33

lated to AI-generated art community.

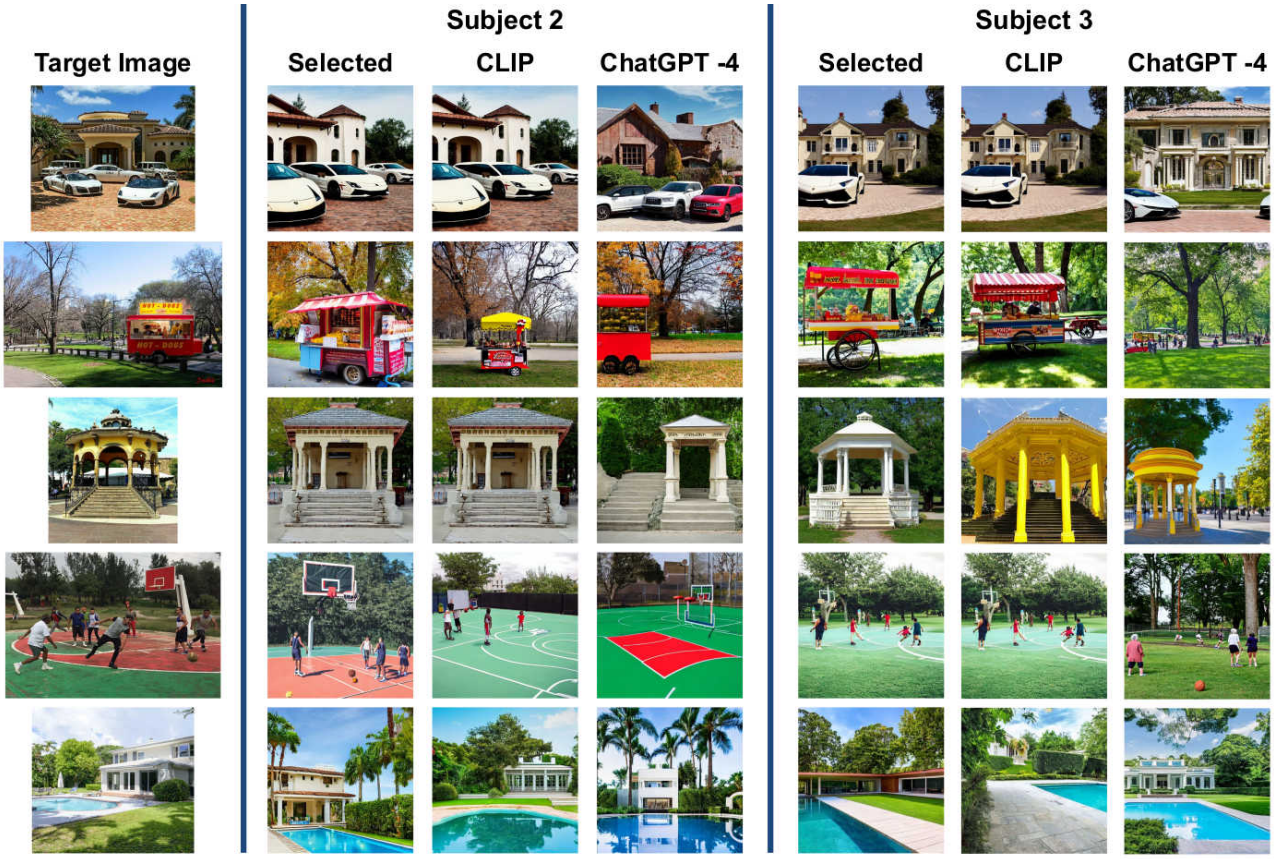

Figure 5: Same comparison as in Figure 4, but without showing the prompts and scores. Note that the resemble between the target and synthetic images is uncanny.

5 CONCLUSION

We have presented the preliminary findings of a methodology which utilises language and generative models such as CLIP, Stable Diffusion, and ChatGPT-4 to provide insights into the exploitation of a textual description in the context of parcel delivery conducted by an autonomous drone. Our research scenario envisages that the drone might reach the vicinity of a delivery point using GPS, but may still require assistance to pinpoint the exact delivery location. In the final stage of delivery, commonly referred to as the 'last-mile delivery', a delivery person often encounters difficulty in identifying the precise target location. Consequently, companies now frequently request customers to provide textual descriptions of the delivery location. Inspired by this scenario, we suggest that a delivery drone might also utilise these textual descriptions to generate synthetic images which reflect the delivery point, even if the drone has never visited the location previously.

Our initial results demonstrate that it is indeed possible to generate synthetic images that could serve as a guide to search for the target location using images captured by an onboard drone camera, given that such images can be compared in a semantic shared space. We acknowledge that a more extensive testing phase involving more subjects and additional image data is needed. However, our preliminary findings are promising and indicate that synthetic generation with generative models such as Stable Diffusion model, can be leveraged with prompt engineering. Therefore, our future work will focus on automatic prompt generation given a single textual description. Indeed, we foresee that in terms of customer interaction, it may not be long before we are able to *speak to* a delivery drone to direct it to its destination, either for delivery or collection purposes.

ACKNOWLEDGEMENTS

We used ChatGPT-4 for prompt generation in the experiments. L.O.R.P. and A.A.C.P. are thankful to CONAHCYT in Mexico for their scholarships with numbers 924254 and 802791.

REFERENCES

[1] Robin Kellermann, Tobias Biehle, and Liliann Fischer. Drones for parcel and passenger transportation: A literature review. *Transportation Research Interdisciplinary Perspectives*, 4:100088, 2020.

[2] Nils Boysen, Stefan Fedtke, and Stefan Schwerdfeger. Last-mile delivery concepts: a survey from an operational research perspective. *Or Spectrum*, 43:1–58, 2021.

[3] Xuping Wang, Linmin Zhan, Junhu Ruan, Jun Zhang, et al. How to choose "last mile" delivery modes for e-fulfillment. *Mathematical Problems in Engineering*, 2014, 2014.

[4] Elżbieta Macioszek. First and last mile delivery–problems and issues. In *Advanced Solutions of Transport Systems for Growing Mobility: 14th Scientific and Technical Conference" Transport Systems. Theory & Practice 2017" Selected Papers*, pages 147–154. Springer, 2018.

[5] Giuseppe Loianno, Chris Brunner, Gary McGrath, and Vijay Kumar. Estimation, control, and planning for aggressive flight with a small quadrotor with a single camera and IMU. *IEEE Robotics and Automation Letters*, 2(2):404–411, 2016.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, 2021.

[7] Jose Martinez-Carranza, Richard Bostock, Simon Willcox, Ian Cowling, and Walterio Mayol-Cuevas. Indoor mav auto-retrieval using fast 6d relocalisation. *Advanced Robotics*, 30(2):119–130, 2016.

[8] L Oyuki Rojas-Perez and Jose Martinez-Carranza. Deeppilot4pose: a fast pose localisation for mav indoor flight using the oak-d camera. *Journal of Real-Time Image Processing*, 20(1):8, 2023.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[10] Vinicius Luis Trevisan de Souza, Bruno Augusto Dorta Marques, Harlen Costa Batagelo, and João Paulo Gois. A review on generative adversarial networks for image generation. *Computers Graphics*, 114:13–25, 2023.

[11] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. CogView: Mastering Text-to-Image Generation via Transformers. In *Advances in Neural Information Processing Systems*, volume 34, pages 19822–19835. Curran Associates, Inc., 2021.

[12] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[13] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A Simple and Performant Baseline for Vision and Language, 2019.

[14] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Semantic Object Accuracy for Generative Text-to-Image Synthesis. *CoRR*, abs/1910.13321, 2019.

[15] Vitali Petsiuk, Alexander E. Siemenn, Saisamrit Surbehera, Zad Chin, Keith Tyser, Gregory Hunter, Arvind Raghavan, Yann Hicke, Bryan A. Plummer, Ori Kerret, Tonio Buonassisi, Kate Saenko, Armando Solar-Lezama, and Iddo Drori. Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark, 2022.

[16] Lanxiao Wang, Wenzhe Hu, Heqian Qiu, Chao Shang, Taijin Zhao, Benliu Qiu, King Ngi Ngan, and Hongliang Li. A survey of vision and language related multi-modal task. *CAAI Artificial Intelligence Research*, 1(2):111–136, 2022.

[17] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7606–7623, Dublin, Ireland, May 2022.

[18] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can Language Understand Depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6868–6874, New York, NY, USA, 2022. Association for Computing Machinery.

[19] Jose Martinez-Carranza, Delia Hernández-Farías, Leticia Oyuki Rojas Pérez, and Aldrich Cabrera Ponce. Language meets yolov8 for metric monocular slam. *Journal of Real-Time Image Processing*, 20, 05 2023.

[20] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. DALL-E-Bot: Introducing Web-Scale Diffusion Models to Robotics. *IEEE Robotics and Automation Letters*, 8(7):3956–3963, jul 2023.

[21] OpenAI. ChatGPT [Large language model], 2023.

[22] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, 2022.

[23] Jonas Oppenlaender. A Taxonomy of Prompt Modifiers for Text-To-Image Generation, 2022.

[24] Anna Notaro. State of the Art: A.I. through the (artificial) artist's eye. In *EVA*, 2020.