Commission of **Computer Engineering, Cinema and Mechatronics**

Master's Degree Course in **Data Science and Engineering**
Text-to-Image Generative Models

DreamShot: Teaching Cinema Shots to Latent Diffusion Models

Supervisors
Prof. Tania Cerquitelli
Dr. Bartolomeo Vacchetti

Candidate
Tommaso Massaglia

SUMMARY

Thesis Objectives

My goal is to use the available state-of-the-art techniques, specifically *Dreambooth* [1] and *Low Rank Adaptation* [2], to finetune an existing latent diffusion model (*stable-diffusion-1-5* [3]) with the goal of generating cinema-like shots from text such that they can be used in the storyboarding task, with a specific emphasis posed on the shot scale (how far the subject is from the camera) as it was proved to be one of the most relevant in conveying the shot feeling [4].

Research Area

Since 2015, a number of models capable of creating increasingly more realistic images were introduced. Starting from Generative Adversarial Network [5], recent years' developments landed on Diffusion Probabilistic Models (DPM) [6] as the best-performing image synthesis approach. DPM models work by learning to reverse the diffusion process (which consists in adding Gaussian noise to an image) in a determined amount of time steps as a reverse Markovian process. By giving as input an image made of random noise, and by adding conditioning of some sort, usually text, it's possible to generate images that match the desired output. Of the many fields that benefit from the ability to generate high-resolution, conditioned samples, we chose to focus on cinema as it is one that widely uses and creates reference pictures to improve workflow through storyboards. By having the ability to generate realistic pictures, generating reference pictures that show expressively

an idea of the desired shot becomes suddenly a task open to anyone, without requiring the need of an extensive library of reference shots or the ability to draw the desired picture.

Contribution

Following an analysis of the state of the art, a consideration of the available resources, and an analysis of user's preference, I decided to use Dreambooth as the finetuning approach of choice. The idea behind DreamBooth is to, given a few input images ($\approx 3-5$), bind the subject to a *unique identifier* such that when it is used in the prompt, the prior knowledge of the model is used along the new information to reconstruct the subject. As Dreambooth was shown to be able to learn a style (as in artistic style) other than a specific subject, the intuition that I followed to teach specific shot types (close shot, medium shot and long shot) is that they can be considered akin to a style. Furthermore, Low Rank Adaptation, an approach to more efficiently finetune existing Large Language Models, was used to further increase the efficiency of the finetuning. As the finetuning process is very sensitive to the input data, a specific focus was posed on the creation and choice of the input dataset.

My contributions are the following:

- A methodical outlining of the process necessary to finetune a style in an existing Latent Diffusion Model using state-of-the-art techniques, and a specific application towards shot types.
- A methodical approach in building a dataset for the finetuning task which makes use of state-of-the-art tools, and its application towards building a 127.000 large cinema shots dataset from which to pool the training images from.
- The application of the two aforementioned approaches to generate three shot-types specific checkpoints (close shot, medium shot, and long shot) and their application in the storyboarding task.

Results

The testing setup closely follows the one proposed in the original Dreambooth paper, looking at the CLIP-T [7] and DINO [8] metrics, as well as using a human evaluation survey. Out of the same subset of 41750 filtered and resized shots that were used to create the training set, 1800 shots were sampled with an even distribution between the three shot types. For each sample, a caption was generated

using an external model with no supervision. The image-caption pairs were then randomly sampled to generate two pictures with the same starting seed, one using the baseline model and one using mine, 1500 times for a total of 500 pairs of generated shots per shot type. The two quantitative metrics used are CLIP-T, the average pairwise cosine similarity between the CLIP textual embeddings of the generated image and the ones of prompt that generated it. The second metric, DINO, measures the average pairwise cosine similarity between the ViTS/16 DINO image embeddings of generated and real images, essentially measuring how similar the generated image is to its real counterpart.

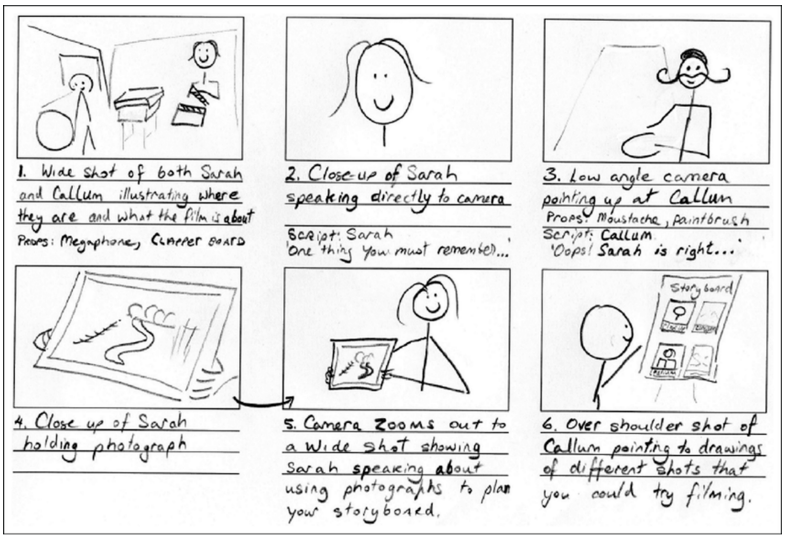
	CLIP-T	**DINO**
baseline	0.3221	0.4163
ours	**0.3269**	**0.4989**

The results show a slight (although significant for the considered metrics) increase for both the CLIP-T and DINO scores over the baseline model. For the human evaluation, each subject was shown a total of 36 pairs of pictures A and B generated with the same setting and the same prompt as the quantitative test, one from the baseline model and one from the finetuned one. Whether a picture was labelled A or B was randomized. For each pair of pictures, three questions were asked, and possible answers for each question were A, B, or *neither/same*. A total of 55 subjects with no required domain knowledge and a high degree of reliability answered the survey.

Question	**baseline**	**ours**	**neither/same**
Which picture do you like most?	26.41	**57.53**	16.06
Which picture is closer to the associated shot type?	20.35	**56.82**	22.83
Which picture is closer to the associated prompt?	19.95	**49.6**	30.45

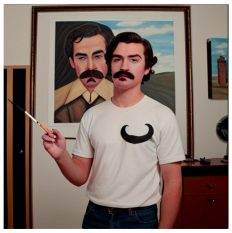
Table 1: Scores expressed as a percentage.

Except for picture likeability, we can see that the baseline model obtained the lowest score of the three, suggesting that even in the worst cases, the generation is of equal quality to the non-finetuned one. Comparing the survey to the CLIP-T and DINO metrics, the results are in line with each other. The higher likeability and shot type closeness are directly related to DINO and they are noticeably higher than prompt closeness and CLIP-T when compared to the baseline.

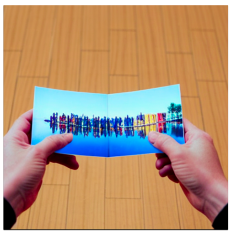


Long Shot, Sarah and Callum talking to each other in the middle of a room

Close Shot, Sarah speaking directly to the camera



Medium Shot, A front view of Callum with a mustache holding a paintbrush



Close Shot, two hands holding a photograph of something over a floor background

Long Shot, Sarah holding a small photograph in a room

Medium Shot, Callum seen from behind pointing at drawings

Figure 1: A practical application of my method, on the top a reference storyboard (provided to entrants of the BBC's 'my place my space' competition) and on the bottom the *equivalent* generated using my finetunings along with the prompt that was used to generate it.

Bibliography

[1] 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. arXiv: 2208.12242 [cs.CV] (cit. on p. i).

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL] (cit. on p. i).

[3] Runaway ML Stability AI. *Stable Diffusion release blog post*. https://stability.ai/blog/stable-diffusion-public-release. (accessed 23-May-2023). 2022 (cit. on p. i).

[4] Brendan Rooney and Katalin E. Balint. «Watching More Closely: Shot Scale Affects Film Viewers Theory of Mind Tendency But Not Ability». In: *Frontiers in Psychology* 8 (2018). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2017.02349 (cit. on p. i).

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML] (cit. on p. i).

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG] (cit. on p. i).

[7] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV] (cit. on p. ii).

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. *Emerging Properties in Self-Supervised Vision Transformers*. 2021. arXiv: 2104.14294 [cs.CV] (cit. on p. ii).