



JOÃO MIGUEL DAMAS DOMINGUES

Bachelor in Computer Science and Engineering

**ADVERSARIAL IMAGERY:
THE POTENTIAL OF GAMIFIED
BENCHMARKING IN EVALUATING DEEP
GENERATIVE MODELS OF IMAGES**

MASTER IN COMPUTER SCIENCE AND ENGINEERING

NOVA University Lisbon

June, 2025



ADVERSARIAL IMAGERY: THE POTENTIAL OF GAMIFIED BENCHMARKING IN EVALUATING DEEP GENERATIVE MODELS OF IMAGES

JOÃO MIGUEL DAMAS DOMINGUES

Bachelor in Computer Science and Engineering

Adviser: Rui Nóbrega

Assistant Professor, NOVA School of Science and Technology

Co-adviser: Pedro Valente

NOVA LINC'S Researcher

Examination Committee

Chair: João Moura Pires

Associate Professor, NOVA School of Science and Technology

Rapporteur: Rui Jesus

Coordinating Professor, ISEL, IPL

Adviser: Rui Nóbrega

Assistant Professor, NOVA School of Science and Technology

Adversarial Imagery: The Potential of Gamified Benchmarking in Evaluating Deep Generative Models of Images

Copyright © João Miguel Damas Domingues, NOVA School of Science and Technology, NOVA University Lisbon.

The NOVA School of Science and Technology and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

À minha mãe que me ensinou a estudar, ao meu pai que esperou por mim, e à minha irmã que me continua a inspirar. Aos meus avós que cuidaram de mim, e aos meus amigos que me fizeram sorrir. Aos que perdi, e a todos os que carrego no meu coração. Devo-vos por tudo o que sou, tudo o que alcancei e alcançarei. Gosto muito de vocês.

Obrigado.

ACKNOWLEDGEMENTS

I would like to thank Professor Rui Nóbrega and Researcher Pedro Valente for their continuous guidance and feedback that elevated the quality of the presented work, always pushing me to strive for more and to avoid shortcuts. I would also like to thank them for their contributions to the paper we submitted to the CHI PLAY 2025 conference. In addition, I would like to thank Professor João M. Lourenço for providing the intuitive template *novathesis LaTeX template* [31] used to write this thesis.

Furthermore, I would like to acknowledge the importance of the participants of the conducted studies and their contributions to the work's evaluation.

I am also grateful to the professors who taught and motivated me during my academic journey, as well as my colleagues and friends who helped me and made everything more bearable.

Finally, I would like to express my deepest gratitude to my entire family, who supported me and helped me overcome my struggles and doubts throughout my life. I want to especially thank my grandparents, who helped care for me and kept me company when I was younger, and my other grandparents, who did the same when I was in college. You helped me not to feel lonely, so I will continue to carry your kindness in my heart. I also want to especially thank my mother, who raised me and helped me study when I was younger, leading to the student and person I am today; my father, who made sure I did not lack anything I needed and accompanied me during college, without forgetting the number of long trips to college and the amount of time spent waiting, in the parking lot, for me to complete several tests; and my sister, who inspired me as I watched her surpass countless challenges, leading to her success, and helped me with mine. All of your courage and determination keep inspiring me, and I want you to know that I would not have made it this far without any of you. Everything I am and everything I have and will achieve is thanks to all of you, and I hope to someday become as great as any of you.

ABSTRACT

The evolution of technology and the continuous research in the field of AI have allowed for the rise of deep generative models capable of producing content that could be potentially mistaken as being made by a human being. Whether a simple conversation with a conversational agent or a database expanded with synthetic images for model training purposes, generative models show promising potential. However, benchmarking becomes a critical step for users in choosing a model to ensure the generative model matches their requirements for the task at hand. Despite the existence of many metrics to evaluate a model's quality, people's perception remains an important factor, even if it can become challenging to gather enough people for an evaluation to be considered reliable.

As such, this thesis proposes a methodology to evaluate the quality of generative models of images by producing synthetic images and assess their realism when confronted with a set of real images that conditioned their generation. In the process, we introduce a game inspired by the Turing tests called *Adversarial Imagery*, which incentivizes players to distinguish real images from AI-generated images. As a result, we could demonstrate the potential of gamified benchmarking as a strategy to evaluate the realism of AI-generated images.

Additionally, we also aimed to achieve a better understanding of current technology's capabilities and limitations and how much it takes to fool a human being. This thesis presents the data collected by *Adversarial Imagery* during a summative study ($N = 129$), which resulted in the collection of 7,760 human-classified images and allowed for the ranking of three generative models considering the realism of the generated images. It also highlights the challenges of designing a system that aims to maintain a high level of engagement while ensuring the quality and trustworthiness of the collected data, considering the data obtained during a formative study ($N = 22$), which revealed the influence of the game's design on the collected data.

Keywords: Artificial Intelligence, Human-Computer Interaction, Deep Generative Models, AI-Generated Images, Turing Test, Gamified Benchmarking

RESUMO

A evolução da tecnologia e a investigação contínua no campo da IA permitiram o surgimento de modelos generativos profundos capazes de produzir conteúdos com o potencial de serem confundidos como feitos por um ser humano. Quer seja uma simples conversa com um agente conversacional ou uma base de dados expandida com imagens sintéticas para fins de treino de modelos, os modelos generativos mostram um potencial promissor. No entanto, o benchmarking torna-se numa etapa crítica para os utilizadores na escolha de um modelo para garantir que o modelo generativo corresponde aos seus requisitos para a tarefa em questão. Apesar da existência de muitas métricas para avaliar a qualidade de um modelo, a perceção das pessoas permanece um fator importante, ainda que possa ser desafiante reunir pessoas suficientes para que uma avaliação seja considerada fiável.

Sendo assim, esta tese propõe uma metodologia para avaliar a qualidade de modelos generativos de imagens através da produção de imagens sintéticas e avaliar o seu realismo quando confrontadas com um conjunto de imagens reais que condicionaram a sua geração. No processo, apresentamos um jogo inspirado pelos testes de Turing chamado *Adversarial Imagery*, que incentiva os jogadores a distinguirem imagens reais de imagens geradas por IA. Como resultado, conseguimos demonstrar o potencial do benchmarking gamificado como estratégia para avaliar o realismo de imagens geradas por IA.

Para além disso, também procurámos compreender melhor as capacidades e limitações da tecnologia atual e do quanto é necessário para enganar um ser humano. Esta tese apresenta os dados recolhidos pelo *Adversarial Imagery* durante um estudo sumativo ($N = 129$), que resultou na recolha de 7760 imagens classificadas por humanos e permitiu a classificação de três modelos generativos considerando o realismo das imagens geradas. Também destaca os desafios de conceber um sistema que procura manter um alto nível de interesse e, ao mesmo tempo, garantir a qualidade e a fiabilidade dos dados recolhidos, considerando os dados obtidos num estudo formativo ($N = 22$) que revelou a influência do design do jogo nos dados colecionados.

Palavras-chave: Inteligência Artificial, Interação Pessoa-Máquina, Modelos Generativos Profundos, Imagens Geradas por IA, Teste de Turing, Benchmarking Gamificado

CONTENTS

List of Figures	ix
List of Tables	xv
1 Introduction	1
1.1 Context and Motivation	2
1.2 Proposal and Objectives	2
1.3 Contributions	4
1.4 Document Structure	5
2 Related Work	6
2.1 AI-Generated Content	6
2.1.1 The Evolution of AI	6
2.1.2 Deep Generative Models	9
2.1.3 Text-to-Image and Image-to-Image	11
2.2 Applications of Image Generation	14
2.2.1 Image Translation	14
2.2.2 Database Augmentation	15
2.3 Gamified Benchmarking and Annotation Tools	16
2.3.1 Turing Tests	17
2.3.2 Games with a Purpose	23
2.4 Summary	28
3 The Gamified Benchmarking Methodology	29
3.1 Requirements	30
3.2 Architecture and Technologies	31
3.3 Design Process	33
4 Automated Image Generation Pipeline	35
4.1 Datasets and Generative Models	35

4.2	Generation Pipeline	37
4.2.1	Image Selection	37
4.2.2	Image Generation	38
4.2.3	Image Storage	40
4.3	Preliminary Experiments and Scrapped Ideas	40
4.4	Summary	43
5	Adversarial Imagery	47
5.1	Game Modes	47
5.1.1	Real or Fake (G1 and TimedG1)	48
5.1.2	Real vs Fake (G2 and TimedG2)	50
5.1.3	Real and Fake (G3 and TimedG3)	51
5.1.4	Hard Mode (TimedG1, TimedG2, and TimedG3)	51
5.2	Engaging Features	52
5.2.1	Leaderboards	52
5.2.2	Achievements	54
5.3	Other Features	55
5.3.1	Data Processing	56
5.3.2	Custom Datasets	57
5.3.3	Tutorials	57
5.3.4	Anti-cheating	58
5.3.5	Game Visuals	59
5.4	Scrapped Game Modes	60
5.5	Summary	61
6	Evaluation	63
6.1	Formative User Study	63
6.1.1	Protocol	63
6.1.2	Population	65
6.1.3	Questionnaire Data Analysis	66
6.1.4	Gameplay Data Analysis	68
6.2	Summative User Study	72
6.2.1	Protocol and Population	72
6.2.2	Results	72
6.3	Discussion	75
6.4	Summary	79
7	Conclusion	80
7.1	Future Work	81
	Bibliography	82

Annexes

I	Figure Annex	88
II	Table Annex	104

LIST OF FIGURES

2.1	The structure of a neural network composed of layers of neurons [37].	8
2.2	A set of handwritten numbers, including the number zero and some of its sections.	8
2.3	The Transformer architecture following a decoder-encoder structure [57]. . .	9
2.4	The computation of the Scaled Dot-Product Attention on the left, to be used by multiple attention layers running in parallel in Multi-Head Attention on the right [57].	10
2.5	A Markov chain illustrating the diffusion process and its reverse of a diffusion model [21].	11
2.6	A high-level overview of unCLIP, where the CLIP training process is depicted above the dotted line while the text-to-image generation process is below. A joint representation space for text and images is learned during the training process. During the generation process, given a CLIP text embedding, an autoregressive or a diffusion model generates an image embedding, which will then be used to condition a diffusion decoder responsible for synthesizing the final image [46].	12
2.7	unCLIP-generated samples obtained through different decoder conditioning approaches. In the first row, the decoder was conditioned solely on text captions, while in the second row, the decoder was conditioned on the text captions and the CLIP text embeddings. Unlike the previous rows, the third row relies on a prior autoregressive model, used to obtain CLIP image embeddings, used to condition the decoder simultaneously with the text captions [46].	13
2.8	DreamBooth-generated images of a teapot in different contexts on the right, given four input images of the teapot subject on the left [48].	14
2.9	Palette-generated images (second column), given modified images as input (first column) from the original images (third column), representative of four image translation tasks: colorization (first row); inpainting (second row); uncropping (third row); JPEG restoration (fourth row) [50].	16

2.10	Images of variations of trucks generated by DA-fusion given another image of a truck, as a means to expand datasets with the goal of model training [55].	17
2.11	The degradation of synthetic image quality produced by a DDIM due to many training iterations on synthetic images, starting from a set of real images of flowers to a set of barely recognizable synthetic images of flowers [33].	18
2.12	A conversation between a person and the computer program ELIZA, mimicking the behavior of a Rogerian therapist [7].	20
2.13	<i>Human or Not</i> game sessions, where players chat with either a chatbot or another person and must identify which one they were texting [23].	21
2.14	Comparison between a painting created by an artist on the left and a painting generated by a computer inspired by it on the right.	23
2.15	Comparison of images created by VQGAN in columns a) and b) and images created by a diffusion model in columns c) and d) [60].	23
2.16	A round of <i>The ESP Game</i> , where pairs of players are incentivized to think alike by typing potential image labels [2].	24
2.17	A round of <i>Verbosity</i> , where a player must help another player at guessing a secret word through clues [2].	26
2.18	A round of <i>TagATune</i> , where pairs of players must decide if they are listening to the same sound by describing it to each other [28].	27
3.1	A visual representation of our gamified benchmarking methodology for deep generative models of images.	30
3.2	A diagram depicting the proposed methodology’s architecture, including its main components and their connections.	31
3.3	A diagram showcasing the iterative design process.	33
4.1	COCO images and their chosen captions.	36
4.2	Farfetch images and their associated product descriptions and highlights.	37
4.3	The automated image generation pipeline containing images used by <i>Adversarial Imagery</i> . A randomly selected COCO image and its randomly assigned caption function as input for three generative models capable of different levels of realism, resulting in three AI-generated images for each original COCO image (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	38
4.4	Images generated based on Figure 4.1a (first column) and its caption, by <i>Realistic Vision</i> (second column), <i>DreamShaper</i> (third column), and <i>ReV Animated</i> (fourth column) (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	39
4.5	Images generated based on Figure 4.1b (first column) and its caption, by <i>Realistic Vision</i> (second column), <i>DreamShaper</i> (third column), and <i>ReV Animated</i> (fourth column) (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	39

4.6	Images generated based on Figure 4.1c (first column) and its caption, by <i>Realistic Vision</i> (second column), <i>DreamShaper</i> (third column), and <i>ReV Animated</i> (fourth column) (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	40
4.7	The structure of the "FinalData.json" file containing the metadata of the real and AI-generated images used in <i>Adversarial Imagery</i>	41
4.8	Images generated based on Figure 4.2a, by <i>Stable Diffusion</i> (left column), <i>epi-CRealism</i> (center column), and <i>Realistic Vision</i> (right column). The first row corresponds to generation conditioned on a product's image, the second row corresponds to generation conditioned on the product's description and highlights, and the third row corresponds to generation conditioned on both (<i>strength</i> = 0.8, <i>num_inference_steps</i> = 50, <i>guidance_scale</i> = 7.5).	42
4.9	Images generated based on Figure 4.2b, by <i>Stable Diffusion</i> (left column), <i>epi-CRealism</i> (center column), and <i>Realistic Vision</i> (right column). The first row corresponds to generation conditioned on a product's image, the second row corresponds to generation conditioned on the product's description and highlights, and the third row corresponds to generation conditioned on both (<i>strength</i> = 0.8, <i>num_inference_steps</i> = 50, <i>guidance_scale</i> = 7.5).	43
4.10	Images generated based on Figure 4.2c, by <i>Stable Diffusion</i> (left column), <i>epi-CRealism</i> (center column), and <i>Realistic Vision</i> (right column). The first row corresponds to generation conditioned on a product's image, the second row corresponds to generation conditioned on the product's description and highlights, and the third row corresponds to generation conditioned on both (<i>strength</i> = 0.8, <i>num_inference_steps</i> = 50, <i>guidance_scale</i> = 7.5).	44
4.11	Manually selected images generated by <i>Realistic Vision</i> , where each row is based on Figures 4.2a, 4.2b, and 4.2c (<i>strength</i> = 0.25, <i>num_inference_steps</i> = 100, <i>guidance_scale</i> = 7).	45
4.12	The inpainting process, where a masked version of a real image is used to generate multiple synthetic ones by filling the masked portion.	45
4.13	An image of a coat taken from Farfetch's dataset used to experiment with the generation of image captions.	46
5.1	Screenshots of the normal versions of <i>Adversarial Imagery's</i> game modes. . .	49
5.2	Screenshots of the hard versions of <i>Adversarial Imagery's</i> game modes.	50
5.3	Screenshots displaying all possible notifications presented to players after completing a game session on <i>Adversarial Imagery</i> for each difficulty option.	53
5.4	Screenshots of <i>Adversarial Imagery's</i> leaderboards highlighting players' performances in the <i>Real and Fake</i> (G3 and TimedG3) game mode.	54
5.5	Screenshots of <i>Adversarial Imagery's</i> different achievements lists.	55
5.6	A screenshot of <i>Adversarial Imagery's</i> custom dataset form, where users can set up their databases for evaluation purposes.	57

5.7	Screenshots of the first version of <i>Adversarial Imagery's</i> tutorials, displayed on the right side of the screen.	58
5.8	Screenshots of <i>Adversarial Imagery's</i> current tutorials.	59
5.9	A comparison between UI elements present in canvases with different "Render Modes".	60
5.10	An early sketch of a cooperative version of <i>Real or Fake</i> where players are rewarded when they agree on a correct guess.	61
6.1	Formative study's population demographics.	66
6.2	The correlation between the participants' accuracy and the perceived difficulty ratings assigned by them for each game mode. The vertical dotted lines represent the standard deviation of the computed accuracy values, while the horizontal dotted lines represent the Q1 and Q3 values of the perceived difficulty ratings.	69
6.3	The accuracy distribution of the participants grouped by their perceived difficulty ratings for each game mode.	70
6.4	The classification accuracy and the rates of images classified with incorrect guesses during gameplay of the timed game modes.	71
6.5	The distribution of <i>Adversarial Imagery's</i> classified images according to COCO's twelve supercategories.	73
6.6	The distribution of <i>Adversarial Imagery's</i> classified images that contain and do not contain the "Person"supercategory.	73
6.7	The rates of AI-generated images classified as real.	75
6.8	The rates of AI-generated images that contain and do not contain the "Person"supercategory classified as real.	76
6.9	The rates of AI-generated images classified as real according to COCO's twelve supercategories.	77
6.10	The normalized image scores of <i>Realistic Vision's</i> dataset according to the assigned labels obtained from G1. The green region contains real images, and the red region contains AI-generated images. The green dots represent real classifications, and the red dots represent fake classifications. The most realistic images correspond to the green dots in the red region with the highest absolute score value, as they represent the AI-generated images more frequently classified as real than the rest, exemplified by the images with the following image pair IDs: 87, 121, and 269.	78
I.1	Farfetch images and their associated product descriptions and highlights. .	88

I.2	Images generated based on Figure I.1a, by <i>Stable Diffusion</i> (left column), <i>epi-CRealism</i> (center column), and <i>Realistic Vision</i> (right column). The first row corresponds to generation conditioned on a product's image, the second row corresponds to generation conditioned on the product's description and highlights, and the third row corresponds to generation conditioned on both (<i>strength</i> = 0.8, <i>num_inference_steps</i> = 50, <i>guidance_scale</i> = 7.5).	89
I.3	Images generated based on Figure I.1b, by <i>Stable Diffusion</i> (left column), <i>epi-CRealism</i> (center column), and <i>Realistic Vision</i> (right column). The first row corresponds to generation conditioned on a product's image, the second row corresponds to generation conditioned on the product's description and highlights, and the third row corresponds to generation conditioned on both (<i>strength</i> = 0.8, <i>num_inference_steps</i> = 50, <i>guidance_scale</i> = 7.5).	89
I.4	Manually selected images generated by <i>Realistic Vision</i> , where each row is based on Figures I.1a and I.1b (<i>strength</i> = 0.25, <i>num_inference_steps</i> = 100, <i>guidance_scale</i> = 7).	90
I.5	COCO images and their chosen captions.	90
I.6	Images generated based on Figure I.5a (first column) and its caption, by <i>Realistic Vision</i> (second column), <i>DreamShaper</i> (third column), and <i>ReV Animated</i> (fourth column) (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	91
I.7	Images generated based on Figure I.5b (first column) and its caption, by <i>Realistic Vision</i> (second column), <i>DreamShaper</i> (third column), and <i>ReV Animated</i> (fourth column) (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	91
I.8	Images generated based on Figure I.5c (first column) and its caption, by <i>Realistic Vision</i> (second column), <i>DreamShaper</i> (third column), and <i>ReV Animated</i> (fourth column) (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	91
I.9	Images generated based on Figure I.5d (first column) and its caption, by <i>Realistic Vision</i> (second column), <i>DreamShaper</i> (third column), and <i>ReV Animated</i> (fourth column) (<i>strength</i> = 0.7, <i>num_inference_steps</i> = 150, <i>guidance_scale</i> = 3.5).	92
I.10	An early sketch of the <i>Real or Fake</i> (G1) game mode, where players are tasked with classifying each image individually as Real or Fake (AI-generated). . .	92
I.11	An early sketch of the <i>Real vs Fake</i> (G2) game mode, where players are tasked with identifying the Fake (AI-generated) image from each pair.	92
I.12	An early sketch of the <i>Real and Fake</i> (G3) game mode, where players are tasked with classifying every image from the group as Real or Fake (AI-generated). . .	93
I.13	An early sketch of a leaderboard for <i>Adversarial Imagery</i> depicting players' ranks, names, scores, and number of played games.	93

I.14	An early sketch of <i>Adversarial Imagery</i> 's achievements list to be unlocked through skillful gameplay.	93
I.15	An early sketch of <i>Adversarial Imagery</i> 's main menu showcasing the buttons to access different game modes, leaderboards, and achievements.	94
I.16	A screenshot of <i>Adversarial Imagery</i> 's "Log In"screen.	94
I.17	A screenshot of <i>Adversarial Imagery</i> 's "Sign Up"screen.	94
I.18	A screenshot of a warning displayed to players of <i>Adversarial Imagery</i> after an internet connection error.	95
I.19	A screenshot of <i>Adversarial Imagery</i> 's consent form.	95
I.20	A screenshot of <i>Adversarial Imagery</i> 's player selection question regarding the player's English comprehension level.	95
I.21	A screenshot of <i>Adversarial Imagery</i> 's player selection question regarding the player's cognitive capabilities.	96
I.22	Screenshots of <i>Adversarial Imagery</i> 's main menus for each difficulty mode.	96
I.23	Screenshots of <i>Adversarial Imagery</i> 's exit menus for each difficulty mode.	97
I.24	The complete structure of <i>Adversarial Imagery</i> 's resulting data, including the initial image metadata.	98
I.24	The complete structure of <i>Adversarial Imagery</i> 's resulting data, including the initial image metadata (cont.)	99
I.25	The distribution of <i>Realistic Vision</i> 's dataset labels obtained from <i>Adversarial Imagery</i> 's different game modes.	100
I.26	The distribution of <i>DreamShapers</i> 's dataset labels obtained from <i>Adversarial Imagery</i> 's different game modes.	101
I.27	The distribution of <i>ReV Animated</i> 's dataset labels obtained from <i>Adversarial Imagery</i> 's different game modes.	102
I.28	Manually picked AI-generated images based on <i>Adversarial Imagery</i> 's label distribution.	103

LIST OF TABLES

6.1	The time spent on each of the formative study's tasks.	64
6.2	Formative study's population demographics.	65
6.3	The enjoyment level of <i>Adversarial Imagery's</i> game modes.	67
6.4	The difficulty level of <i>Adversarial Imagery's</i> game modes.	67
6.5	The impact of <i>Adversarial Imagery's</i> mechanics.	67
6.6	The ease of distinguishing between real images and AI-generated ones. . . .	68
6.7	The player accuracy calculated from the participants' first and second game sessions of the normal game modes.	69
6.8	General game statistics obtained from <i>Adversarial Imagery</i>	72
6.9	The number of sessions played by each player of <i>Adversarial Imagery</i> grouped by their assigned databases.	74
II.1	Image statistics obtained and computed from players' first playthrough of <i>Adversarial Imagery</i> during the formative study.	105
II.2	Image statistics obtained and computed from players' second playthrough of <i>Adversarial Imagery's</i> normal game modes, aware of the game's instructions, during the formative study.	106
II.3	Image statistics obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with short round times, during the formative study.	107
II.4	Image statistics obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with intermediate round times, during the formative study.	108
II.5	Image statistics obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with long round times, during the formative study.	109
II.6	The statistics of every image with the "Person" supercategory obtained and computed from players' first playthrough of <i>Adversarial Imagery</i> during the formative study.	110

II.7	The statistics of every image with the "Person"supercategory obtained and computed from players' second playthrough of <i>Adversarial Imagery's</i> normal game modes, aware of the game's instructions, during the formative study.	111
II.8	The statistics of every image with the "Person"supercategory obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with short round times, during the formative study.	112
II.9	The statistics of every image with the "Person"supercategory obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with intermediate round times, during the formative study.	113
II.10	The statistics of every image with the "Person"supercategory obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with long round times, during the formative study.	114
II.11	The statistics of every image without the "Person"supercategory obtained and computed from players' first playthrough of <i>Adversarial Imagery</i> during the formative study.	115
II.12	The statistics of every image without the "Person"supercategory obtained and computed from players' second playthrough of <i>Adversarial Imagery's</i> normal game modes, aware of the game's instructions, during the formative study.	116
II.13	The statistics of every image without the "Person"supercategory obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with short round times, during the formative study.	117
II.14	The statistics of every image without the "Person"supercategory obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with intermediate round times, during the formative study.	118
II.15	The statistics of every image without the "Person"supercategory obtained and computed from players' playthrough of <i>Adversarial Imagery's</i> hard game modes, with long round times, during the formative study.	119
II.16	Unfiltered general statistics obtained from <i>Adversarial Imagery</i>	119
II.17	The unfiltered number of sessions played by each player of <i>Adversarial Imagery</i> grouped by their assigned databases.	120
II.18	Realistic Vision's image statistics obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	121
II.19	DreamShaper's image statistics obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	122
II.20	ReV Animated's image statistics obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	123
II.21	Realistic Vision's statistics of every image with the "Person"supercategory obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	124

II.22 DreamShaper's statistics of every image with the "Person" supercategory obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	125
II.23 ReV Animated's statistics of every image with the "Person" supercategory obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	126
II.24 Realistic Vision's statistics of every image without the "Person" supercategory obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	127
II.25 DreamShaper's statistics of every image without the "Person" supercategory obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	128
II.26 ReV Animated's statistics of every image without the "Person" supercategory obtained and computed from players' playthroughs of <i>Adversarial Imagery</i> during the summative study.	129

INTRODUCTION

The emergence of deep generative models sparked interest from curious potential users and researchers due to their potential applications. According to García-Peñalvo and Vázquez-Ingelmo [16], interest in Generative AI rose since the release of tools such as ChatGPT¹ and Midjourney². The authors also revealed that the number of accessible English-written papers published in peer-reviewed journals, books, or conferences whose main goals are content generation applications, where the solution is described, as well as the applied field, has increased from forty-three in 2019 to two hundred and eleven in 2022. As such, the number of published works per year with the previously mentioned characteristics saw an increase of approximately 391% without considering reviews, surveys, and comparative analyses. The application domains of the published articles revolve around medicine, computer vision, and Natural Language Processing, among others. It is also important to note that data augmentation is one of the main tasks across domains, especially medicine.

Deep generative models emerged with the promise of revolutionizing the world. Being capable of producing content, almost as if a human did it, is a milestone of technological progress, an unimaginable feat when considering the birth of the first computer. As more apps and tools are released, it becomes more difficult for users to discover the one that best fits their needs and quality expectations. As such, we can not overlook the importance of benchmarking.

As with any other emerging technology, wonders and fears spread throughout the globe. As much as it is amazing to witness the evolution of Artificial Intelligence and what it can do nowadays, it can also become a concerning sight. While it showcases the potential to make people's lives easier by handling monotonous and repetitive work, it has also raised discussions regarding subjects such as ownership and misinformation. Furthermore, history has shown us a pattern for every new invention regarding employment. They have erased certain jobs or reserved them for a select few. In contrast, they can create new jobs, which might lead to an update of desired skills in the job market, as a similar situation occurred with the introduction of the mechanical switching system in operating

¹Introducing ChatGPT; A model capable of producing human-like text; <https://openai.com/blog/chatgpt>; Last access: 2025

²Midjourney; An image generation tool; <https://www.midjourney.com>; Last access: 2025

companies as discussed in Feigenbaum et al. [14]. As such, imagining the consequences of easily generated images for artists and designers is not too difficult.

1.1 Context and Motivation

With the rising interest in deep generative models and AI-generated content, more models became available to users, capable of producing content convincingly enough to potentially fool human beings. Besides the curiosity of chatting with a chatbot or seeing if a machine can create art, people can use these tools to aid them in certain tasks. Drawing inspiration for conceptual design, as demonstrated in Ploennigs and Berge [43], or Yan et al. [59], is one of the most straightforward tasks users may be inclined to use generative models. However, there is much more that these models can achieve, such as image processing and restoration tasks showcased in Saharia et al. [50], and database augmentations for model training exemplified in Trabucco et al. [55].

With so many deep generative models available to the public, some trained to cater to more specific tasks, choosing the most appropriate model for the users' needs can be difficult. It can be especially challenging if the user in question does not possess enough knowledge or practice to resort to techniques such as prompt engineering, which might sometimes be necessary to produce acceptable results. It is common to draw comparisons between different models, often at release, to prove to potential users the capabilities of their models, for example, the results obtained with DrawBench and the introduction of Google's Imagen³ presented by Saharia et al. [51].

Although there are many metrics to evaluate the quality of generated images, like the FID and CLIP scores, they still present some constraints, as Saharia et al. claim that the first one is not fully aligned with perceptual quality, while the second is ineffective at counting. Therefore, it is still important to consider people's opinions and perceptions since they are the ones who will use these models and possibly be exposed to the generated content. As a consequence, it is desirable to attract as many participants as possible for evaluation to ensure more reliable data, which can become an issue, considering that it is unlikely that many people would volunteer for these experiments. It can also become challenging to incentivize a large number of potential participants, meaning that the evaluation would have to compromise and settle for fewer participants, possibly affecting the reliability of the collected data.

1.2 Proposal and Objectives

Games have the potential to draw players into completing certain tasks computers might struggle with, allowing for the extraction of valuable data. These games are traditionally called "Games with a Purpose" [2]. As such, this thesis proposes a methodology to generate

³Imagen; Google's text-to-image diffusion model; <https://imagen.research.google/>; Last access: 2025

images and evaluate their realism through a game called *Adversarial Imagery*, whose title was inspired by the Generative Adversarial Networks [17] (GANs) framework, where the discriminator models take a similar role to the players of the game. In *Adversarial Imagery*, we task players with distinguishing generated images from real images, taking inspiration from the Turing tests [56]. The game includes different game modes and engaging features such as leaderboards and achievements to motivate the players to play multiple sessions and take their tasks seriously. While the players enjoyed their time, we collected valuable data to infer the realism of the synthetic images generated by deep generative models. The game uses three datasets of images generated by three generative models and a set of real images that conditioned their generation. It is important to note that the concept of the proposed methodology should be able to evaluate any generative model. As such, we also designed *Adversarial Imagery* to allow users to evaluate the realism of their AI-generated images.

To sum up, this thesis aims to answer the following questions:

- Q1: Is it possible to use games as a benchmarking strategy for deep generative models of images?

By using a game to lure players into participating in the models' evaluation, we expected to demonstrate the potential of gamified benchmarking as a means to collect valuable data from a large number of participants.

- Q2: What are deep generative models of images capable of, and what are their limits?

Some of the produced AI-generated images might not look as realistic as the real ones and present artifacts or patterns that the real images do not follow. Consequently, we expected these limitations to be reflected in the collected data, as the players would more frequently identify them as fake.

- Q3: How easy is it to fool a human being with AI-generated images?

We expected the success rate of the players to depend on the realism of the AI-generated images and the context in which we presented them, which provided an opportunity to determine how often they would mistakenly identify a synthetic image as real.

- Q4: What are the challenges of designing a game meant to collect valuable and trustworthy data for research?

We expected the development of *Adversarial Imagery* to require careful planning to seem appealing enough to convince a considerable number of people to play it and maintain player retention without compromising the trustworthiness of the collected data. The game's design also had to account for the fact that we could not guide or monitor the participants during the summative study and prevent adversarial behavior.

- Q5: How can the design of a gamified data collection system influence its collected data?

The design of *Adversarial Imagery* aimed to motivate players to keep playing and ensure they would focus on their tasks. As such, we expected the design of the different game modes and other features to significantly affect the amount and quality of the collected data.

1.3 Contributions

This dissertation resulted in a game inspired by the Turing tests for players to enjoy, a framework showcasing the potential of gamified benchmarking, whose concept could evaluate any generative model, and a system that allows users to evaluate their image datasets. In the process of evaluating the realism of the images generated by the chosen models, we also acquired a better understanding of the current state of the technology, as well as its limitations. Furthermore, we collected several human-annotated images whose labels resulted from the guesses made by the game’s players. We also assessed people’s difficulty in distinguishing real from fake images. Additionally, we analyzed the challenges of designing a system enjoyable enough to gather a considerable amount of users while ensuring the quality of the collected data. Finally, the collected data allowed for the ranking of three deep generative models regarding the realism of their generated images, whose quality aimed to simulate the ones generated by beginner users. In summary, the contributions from this thesis are as follows:

- A proof of concept regarding the potential of gamified benchmarking strategies for deep generative models of images.
- A game inspired by the Turing tests called *Adversarial Imagery*^{4,5}, which players can enjoy while enabling the collection of meaningful evaluation data.
- A tool that users can utilize to evaluate the realism of their AI-generated images.
- Annotated datasets containing real and AI-generated images.
- An evaluation of the realism of AI-generated images produced by three generative models from the perspective of a hypothetical beginner user.
- A better understanding of the current state of generative technology.
- An analysis of the challenges of designing an entertaining gamified system responsible for collecting valuable and trustworthy data.

⁴Adversarial Imagery; A game inspired by the Turing test, which tasks players to identify real and AI-generated images (Test Account - Name: TEST; PassWord: 1234); <https://adversarial-imagery.github.io/Adversarial-Imagery>; Last access: 2025

⁵Adversarial Imagery Demo; A video showcasing Adversarial Imagery’s features; <https://youtu.be/bpYNfQrmKM0>; Last access: 2025

Lastly, we submitted this work as a full paper to the CHI PLAY 2025 conference.

1.4 Document Structure

Chapter 2 provides some context regarding deep generative models and their evolution since the birth of Artificial Intelligence to better understand how different models can achieve human-like content. Afterward, we present some possible applications for image generation as a means to discuss the potential importance of generative models and the need for benchmarking. We also introduce the concept of the Turing test, as well as different strategies people have used throughout the years to fool others, as the developed game takes some inspiration from these tests. Finally, we showcase some examples of games created to draw in as many people as possible to complete certain tasks, not only to defend the concept of using games to obtain valuable data but also to present some mechanics to ensure players are engaged and to minimize the poisoning of data through cheating.

Chapter 3 presents the proposed methodology's requirements and architecture. It also introduces the methodology's two main components: the image generation pipeline and *Adversarial Imagery*.

Chapter 4 focuses on the implementation of the image generation process and presents our chosen generative models, the selected dataset of real images, and some conducted experiments to become familiar with the technology.

Chapter 5 analyzes *Adversarial Imagery*'s design, highlighting its different game modes, its leaderboards and achievements, and other features such as the game's tutorials and anti-cheat system. It also presents some ideas for new game modes for a potential new version of *Adversarial Imagery*.

Chapter 6 displays the data obtained during the formative and summative studies, highlighting the impact of the game's design on the collected data and providing information regarding the capabilities of the selected generative models.

Chapter 7 concludes this thesis with an overview of the accomplished work and suggestions for future projects.

Annex I displays additional images, including more results of initial experiments with generative models, initial sketches and extra screenshots of *Adversarial Imagery*, and the structure of the collected data. It also includes the guess distribution of the players of *Adversarial Imagery* and additional manually selected AI-generated images considering such distribution.

Annex II presents additional tables including complementary data obtained from both studies, which include unfiltered results and other statistics such as the precision, recall, and F1 scores associated with the classified images.

RELATED WORK

In order to properly evaluate a deep generative model, it is important to understand what they are capable of and have an idea of how they achieve it, as well as how it evolved throughout the years. Not only that but to obtain a reliable and consistent source of human validation, it is essential to understand what strategies people could use when tasked with identifying generated images, as well as discover potential approaches to draw in as many participants as possible.

2.1 AI-Generated Content

As machine learning and deep learning algorithms have evolved, Artificial Intelligence has become capable of not only making predictions and recommendations but also producing content that could be mistaken as human-made.

2.1.1 The Evolution of AI

As mentioned in Haenlein and Kaplan [18], in 1942, the short story "Runaround" was published by the American writer Isaac Asimov. In it, the author defines the "Three Laws of Robotics". Despite technology being far away from being able to obey these laws at the time or even having a need to, Asimov's work inspired many scientists in robotics, computer science, and Artificial Intelligence. Meanwhile, what was considered the first working electro-mechanical computer, known as "The Bombe", was created by the English mathematician Alan Turing as a means to decipher the Enigma code used by the Germans during World War II. Afterward, in 1950, Turing [56] theorized how it would be possible for a machine to simulate an adult human mind. Turing raised the possibility of a learning machine, comparing it to the process of educating a child through rewards and punishments, even if it meant the teacher could not tell what was happening inside the machine, presenting itself as a black box.

The term Artificial Intelligence was officially coined in 1956 as stated in Copeland [10], the year when the later-called "Logic Theorist" program was developed by Newell, Simon, and Shaw, capable of proving theorems from Whitehead and Russell's "Principia

Mathematica", as well as the year the "Dartmouth Summer Research Project on Artificial Intelligence" (DSRP AI) conference was hosted by Marvin Minsky and John McCarthy. The following two decades were characterized by progress in the field of AI, marked by the successes of some programs, such as Joseph Weizenbaum's ELIZA, a tool capable of simulating a conversation with a human, and the "General Problem Solver" [36], developed by Newell, Shaw, and Simon, being able to solve problems that could be formulated in terms of objects, defined by their features, and operators. These actions can be applied to those objects to create new ones. Games like Chess or Towers of Hanoi are some examples that follow this formulation. These instances were considered to be "Expert Systems", systems formed with a collection of rules that could be structured as if-then-else statements that end up failing on tasks that require learning from correctly extracted and interpreted data, meaning that these systems were not actually true AI on a technical level. This would eventually lead to the research of Artificial Neural Networks in an attempt to reproduce the process of neurons in the human brain, initially based on the Canadian psychologist Donald Hebb's Hebbian Learning theory. Several decades later, due to technological advances, computers began to have enough processing power to handle artificial neural networks, and as such, Deep Learning was born. This led to the development of a program called AlphaGo [53] by Google in 2015, which managed to achieve a winning rate of 99.8% against other programs designed to play games of Go, a game far more complex than Chess and win five games to zero against the European Go champion at the time.

Linear classifiers can sometimes struggle in the task of image recognition, as argued in Lecun, Bengio, and Hinton [29], requiring the design of good feature extractors that demand a significant amount of engineering skill. The problem can be seen when, for example, there are two images of a wolf-like dog, for instance, a Samoyed or a Husky, in different poses and different environments and two images of a wolf and a similar-looking dog in similar poses and environments. The second pair is more likely to fall into the same category as the first one. This is due to the selectivity-invariance dilemma, where it is needed for the input-output function to be insensitive to input variations, like the position of the animals in this case, while there are more important features that distinguish the images. Deep learning tries to avoid this issue by learning good features automatically through neural networks.

A neural network [37] comprises layers of neurons, as presented in Figure 2.1. Each neuron receives a set of inputs, each with a weight, and outputs a value based on the weighted sum of the inputs transformed by a function such as a sigmoid (Equation 2.1) or a Rectified Linear Unit (ReLU) (Equation 2.2), and a bias value. Considering the task of identifying handwritten single digits present in an image, like the ones present in Figure 2.2a, while omitting some implementation details, it would be necessary an input layer of neurons, where there would be as many neurons as the number of pixels of the image, as well as an output layer with ten neurons, the number of possible results. The neuron with the highest activation value corresponds to the identified digit. Between the input and output layers, middle layers would also be necessary, named hidden layers,

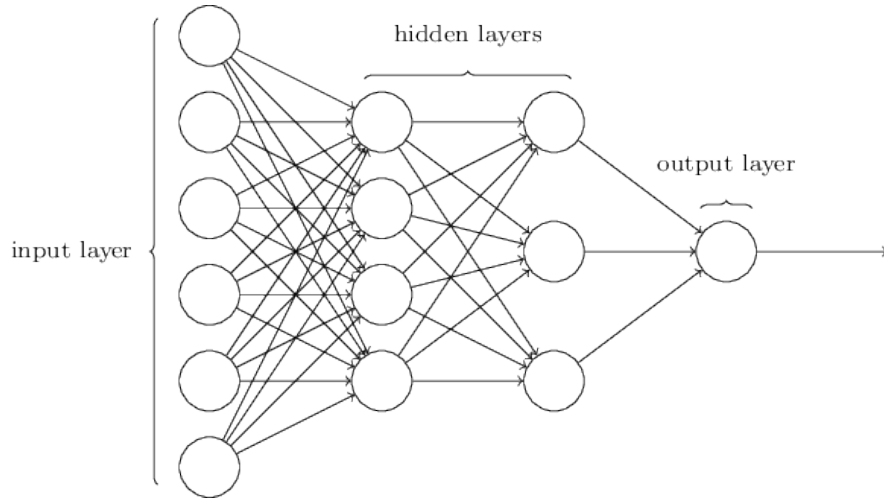
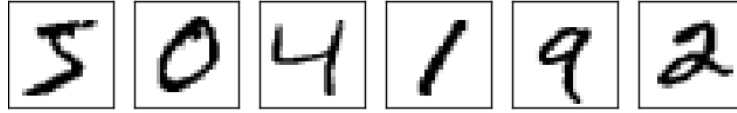
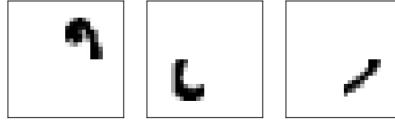


Figure 2.1: The structure of a neural network composed of layers of neurons [37].



(a) Handwritten numbers to be identified by a neural network [37].



(b) Some sections of the handwritten number zero to be possibly identified by a neural network's hidden layers [37].

Figure 2.2: A set of handwritten numbers, including the number zero and some of its sections.

and like how the output layer was responsible for identifying if a particular digit was present in the picture, the hidden layers would be accountable for something similar, the identification of specific patterns or edges that would be present in the digit, like the ones displayed in Figure 2.2b. It is important to note that the values for weights and biases are computed with the help of an algorithm called gradient descent.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (2.1)$$

$$\text{ReLU}(\alpha) = \max(0, \alpha) \quad (2.2)$$

A crucial step in Natural Language Processing (NLP) was the introduction of the Transformer architecture, which ChatGPT is based on, according to Radford et al. [45]. The Transformer [57] follows a decoder-encoder structure, as displayed in Figure 2.3, where each step is auto-regressive, meaning that previously generated symbols are also used as

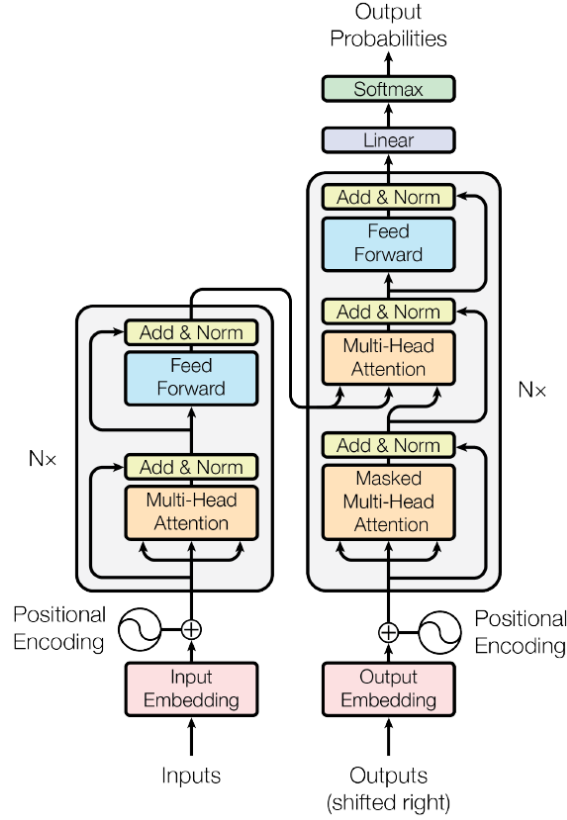


Figure 2.3: The Transformer architecture following a decoder-encoder structure [57].

input to generate the next. The encoder is responsible for mapping an input sequence of symbol representations to a sequence of continuous representations. Furthermore, given that sequence, the decoder generates an output sequence of symbols, each individually. To do so, the Transformer relies on multi-head attention mechanisms through the attention function, where a mapping of a query and a set of key-value pairs is assigned to an output. As such, it is computed the Scaled Dot-Product Attention (Equation 2.3), where queries, keys, and values are packed together in the matrices Q , K , V respectively, and the keys have dimension d_k . It is important to note that multiple attention layers are running in parallel, as shown in Figure 2.4.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2.3)$$

2.1.2 Deep Generative Models

Following advancements in the field of AI, interest in Generative AI and deep generative models increased significantly, as shown in García-Peñalvo and Vázquez-Ingelmo [16], capable of producing human-like content, like text or images, instead of trying to obtain a prediction from a given input.

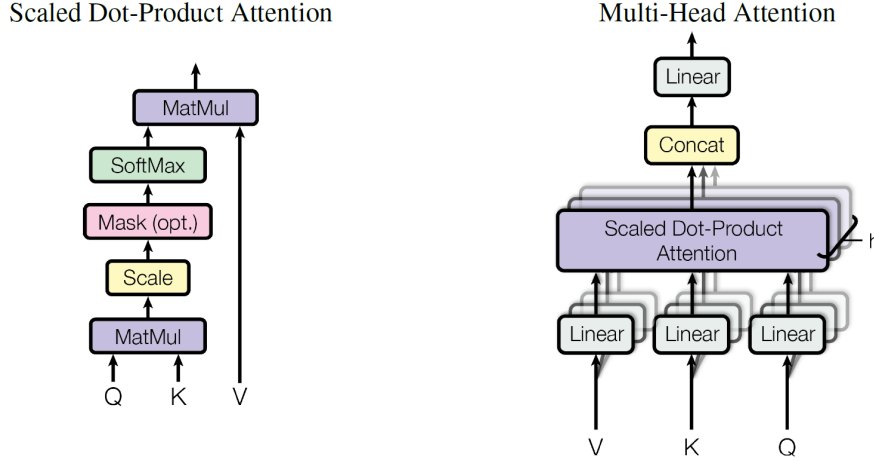


Figure 2.4: The computation of the Scaled Dot-Product Attention on the left, to be used by multiple attention layers running in parallel in Multi-Head Attention on the right [57].

A modeling approach known as Generative Adversarial Networks remains popular as a way to generate realistic synthetic images. This framework, as discussed in Ruthotto and Haber¹ [49], and Goodfellow et al.² [17], can be summarized through an analogy of a game played between two adversary models, a generator and a discriminator. These rival models, following the architecture of neural networks, compete against each other with the goal of improving each other, leading to better results, which, in this case, would be more convincing fake images. The generator is tasked with generating the images. At the same time, the discriminator is responsible for classifying the image as if it was created by the generator or not, depending on the computed probability. For that, a training dataset is required as well as the cross-entropy loss function (Equation 2.4). The success of the discriminator can be understood as the maximization of this function, meaning fewer classification errors were made. In contrast, the generator aims for the opposite, trying to fool the other model. Through many iterations, the weights of the neural networks become as optimal as possible. Consequently, generated images can become convincing enough to potentially fool a human being.

$$J_{GAN}(\theta, \phi) = \mathbb{E}_{x \sim \mathcal{X}}[\log(d_{\phi}(x))] + \mathbb{E}_{z \sim \mathcal{Z}}[\log(1 - d_{\phi}(g_{\theta}(z)))] \quad (2.4)$$

Among other methods for image generation, a promising approach relies on Diffusion Models (DMs), a method that can create image samples of higher quality compared to its alternatives, including GANs³, as demonstrated in Dhariwal and Nichol [13], and as a likelihood-based model, it is also easier to scale and train and offers a higher level of

¹EmoryMLIP/DeepGenerativeModelingIntro; Code used to exemplify different types of deep generative models; <https://github.com/EmoryMLIP/DeepGenerativeModelingIntro>; Last access: 2025

²goodfeli/adversarial; The introduction of a generative adversarial network implementation; <http://www.github.com/goodfeli/adversarial>; Last access: 2025

³openai/guided-diffusion; Code used to demonstrate the potential of diffusion models compared to GANs; <https://github.com/openai/guided-diffusion>; Last access: 2025

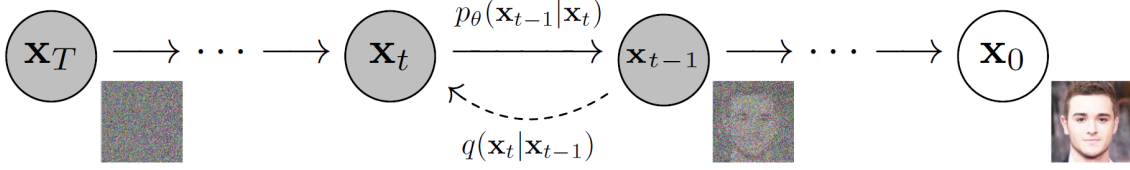


Figure 2.5: A Markov chain illustrating the diffusion process and its reverse of a diffusion model [21].

diversity. A diffusion probabilistic model⁴ [21] is a parameterized Markov chain [30], a process similar to a finite state machine, whose transitions depend on the probability of moving from one state to another, as exemplified in Figure 2.5. Alternatively, these models can also be described as a hierarchy of denoising autoencoders according to Rombach et al. [47]. Autoencoders [6] can be understood as neural networks trained to reconstruct their input. This is commonly achieved through two neural networks designated as encoder and decoder, which allows for the encoding of an input image to a compressed representation and its encoding to achieve a reconstructed image. The algorithm of a diffusion model, as explained in Sohl-Dickstein et al. [54] consists of a forward diffusion process, where noise will be gradually added to the data, in order to obtain a simpler and more tractable distribution, as well as the reverse process responsible for denoising and retrieving the generative model's distribution. However, diffusion models present low inference speed while evaluating and optimizing them since they are done in pixel space, leading to high training costs. Latent Diffusion Models⁵ (LDMs), like Stable Diffusion⁶, attempt to mitigate these issues by working on a compressed latent space of lower dimensionality. To do so, the authors introduced an autoencoding model capable of learning a perceptually equivalent space to the image space, allowing for a clear separation between the compressive and generative learning stages. Consequently, more priority is given to the more semantically important bits of the data while ignoring the more imperceptible details. In addition, it is also possible to train in a lower and more efficient space, avoiding the need for pixel-wise evaluation.

2.1.3 Text-to-Image and Image-to-Image

With the currently available tools, users are able to generate the images they want through Text-To-Image (T2I) by providing textual descriptions as input and Image-To-Image (I2I) translation approaches using a guiding image as input.

In text-to-image approaches, like the one presented in Ramesh et al. [46], it is possible to produce images conditioned on text captions through a decoder model. Even though

⁴hojonathanho/diffusion; The implementation of a diffusion probabilistic model; <https://github.com/hojonathanho/diffusion>; Last access: 2025

⁵CompVis/latent-diffusion; A repository containing many latent diffusion models; <https://github.com/CompVis/latent-diffusion>; Last access: 2025

⁶stability.ai; stability.ai's webpage and their launched models; <https://stability.ai>; Last access: 2025

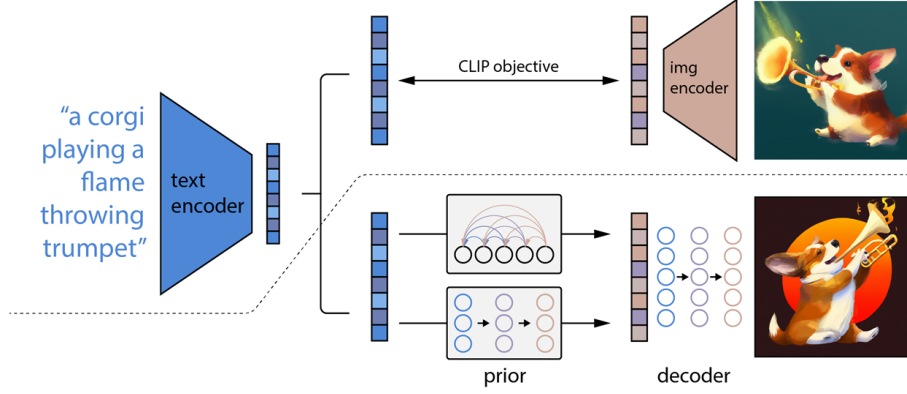


Figure 2.6: A high-level overview of unCLIP, where the CLIP training process is depicted above the dotted line while the text-to-image generation process is below. A joint representation space for text and images is learned during the training process. During the generation process, given a CLIP text embedding, an autoregressive or a diffusion model generates an image embedding, which will then be used to condition a diffusion decoder responsible for synthesizing the final image [46].

it is not essential, a prior model can be used to produce image embeddings, vectors that represent the images, which can be used to condition the image generation besides the text captions. This allows for better results, as illustrated in Figure 2.7, with better Fréchet Inception Distance (FID) scores [20], an evaluation metric used to measure the quality of a model’s generated images, like the Inception Score (IS), and evaluates the similarity between generated images and real ones. The prior model approach has been implemented in models such as unCLIP⁷ and DALL·E 2⁸, illustrated in Figure 2.6. The first model creates Contrastive Language-Image Pre-training (CLIP) [44] image embeddings, obtained through the training of a text encoder and an image encoder as a way to predict the correct pairings between images and captions based on the cosine similarity between their embeddings. The CLIP image embeddings are modeled with a Gaussian diffusion model conditioned on the captions by training a decoder-only transformer, whose output is used to predict the CLIP image embedding without its noise, which is then used to condition the diffusion models. Alternatively, it is also possible to use an autoregressive model to generate the image embeddings. However, it results in the production of lower-quality images, evidenced by lower human evaluation scores regarding photorealism, caption similarity, and diversity.

The objective of image-to-image approaches consists of obtaining the transformation process (Equation 2.5), by training a mapping $G_{A \rightarrow B}$ that can generate an image $x_{AB} \in B$ as similar as possible to target domain image $x_B \in B$ given an input source $x_A \in A$, according to Pang et al. [42].

$$x_{AB} \in B : x_{AB} = G_{A \rightarrow B}(x_A) \quad (2.5)$$

⁷unCLIP; An image generator model conditioned on CLIP image embeddings; <https://huggingface.co/docs/diffusers/api/pipelines/unclip>; Last access: 2025

⁸DALL·E 2; A generative model capable of producing art and realistic images; <https://openai.com/dall-e-2>; Last access: 2025

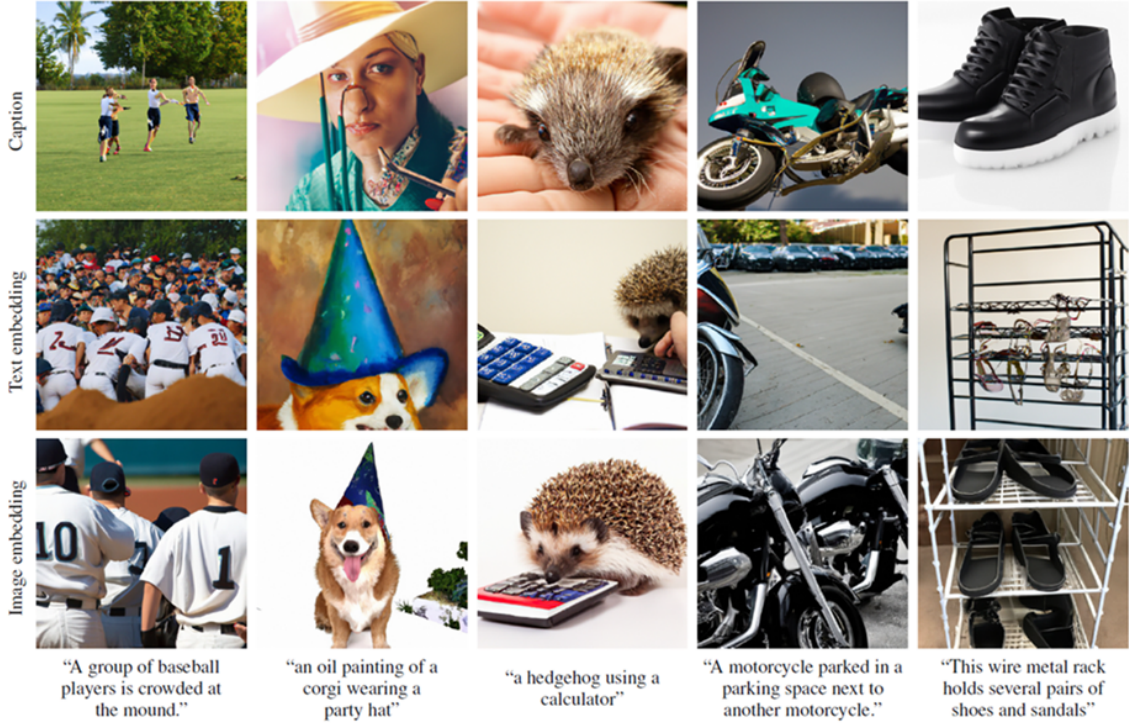


Figure 2.7: unCLIP-generated samples obtained through different decoder conditioning approaches. In the first row, the decoder was conditioned solely on text captions, while in the second row, the decoder was conditioned on the text captions and the CLIP text embeddings. Unlike the previous rows, the third row relies on a prior autoregressive model, used to obtain CLIP image embeddings, used to condition the decoder simultaneously with the text captions [46].

This allows for the output of new images from existing input images. To achieve this, it is possible to train the models through a supervised approach, as pointed out in Sasaki, Willcocks, and Breckon [52], with pairs of real and fake images. However, other approaches, such as training without paired examples, are also possible. An example of this utilizes a GAN named CycleGAN, where the generator and discriminator are modified to transfer from source images to target images. In the case of diffusion models, the target images can be gradually synthesized from Gaussian noise and the noisy source images. As such, the generative process becomes conditioned on the input images following the diffusion forward process until a certain point, called release time. From that point, the reverse process begins. This happens as a means to approximate the data distribution of the source and target domains.

It is important to note that it is also possible to produce new images, given an image and a text prompt as input. DreamBooth⁹ [48] achieves this by fine-tuning a pre-trained text-to-image model where unique identifiers are assigned to the subject present in the provided input images. Through this method, it is possible to achieve variations of

⁹DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation; DreamBooth’s project page; <https://dreambooth.github.io>; Last access: 2025



Figure 2.8: DreamBooth-generated images of a teapot in different contexts on the right, given four input images of the teapot subject on the left [48].

the subject present in the input images guided by text prompts, being able to alter its location, shape, pose, color, and viewpoint, among other semantic modifications, as seen in Figure 2.8. To accomplish this level of personalization, pairs of unique identifiers and subjects are added to the model's "dictionary". As such, to avoid long image descriptions, each image is labeled with its identifier and a class noun for a general description of the subject, like "dog" or "watch".

2.2 Applications of Image Generation

Besides the interest in creating new images, whether it is a new painting in the style of a well-known talented artist of scenery that does not actually exist, a realistic photo of a fictional character brought to life, or even concept designs for early stages of architectural and fashion design, as showcased in Ploennigs and Berge [43] and Yan et al. [59] respectively, aiding professionals, there is much more that deep generative models can achieve.

2.2.1 Image Translation

The image-to-image diffusion model, Palette¹⁰ [50], showcases its potential for several image processing tasks like colorization, inpainting, uncropping and JPEG restoration, as illustrated by Palette's generated images in Figure 2.9, outperforming GANs, and other diffusion models, across several metrics such as FID, IS, Classification Accuracy (CA) of a pre-trained ResNet-50, which depends on the success of the model in classifying images, and Perceptual Distance (PD) against original images. The last metric is discussed in Zhang et al.¹¹ [62], and it measures the perceptual similarity between images, or through

¹⁰Palette: Image-to-Image Diffusion Models; An image-to-image diffusion model capable of image processing tasks; <https://diffusion-palette.github.io>; Last access: 2025

¹¹richzhang/PerceptualSimilarity; Models and data used to showcase the potential of the perceptual distance metric; <https://github.com/richzhang/PerceptualSimilarity>; Last access: 2025

simpler terms, how similar is an image compared to another. Given a grayscale version of an image as input, Palette is capable of achieving a fool rate of 47.80%, meaning it manages to produce colored images so convincing that human raters often mistake them as the original when faced with the model’s output and the original colored images, surpassing the scores of other models that followed a similar qualitative assessment, like the Colorization Transformer (ColTran)¹² [25]. Through inpainting, it is possible to mask an element of an image, similar to Photoshop’s¹³ Content-Aware Fill. Palette achieves high performance across inpainting datasets and mask configurations by filling the masked portion with standard Gaussian noise, compatible with denoising diffusion models. Uncropping, also known as outpainting, is more demanding than inpainting because it generates open-ended content with less context and information. Nonetheless, it is still possible to extend images, as Palette achieves a fool rate of 40%. Lastly, image restoration often comprises the removal of JPEG compression artifacts, a task that can be accomplished by Palette due to its training on compressed inputs of different quality.

2.2.2 Database Augmentation

Large-scale databases, like ImageNet¹⁴ [12], a dataset containing millions of annotated images, have become essential resources for developing more robust models and algorithms in the field of computer vision, serving as a source for training and benchmarking. As such, the possibility of expanding datasets with synthetic images produced by deep generative models has been raised, exemplified in Trabucco et al. [55]. A potential method of augmenting image datasets would be to apply transformations, such as flips or rotations. However, this approach does not increase the dataset’s diversity, lacking new structural elements, textures, or perspective changes, which are necessary for a model’s robustness regarding more subtle details. DA-Fusion’s¹⁵ augmentations build on Stable Diffusion, and its goal is to produce augmentations, such as the ones exemplified in Figure 2.10, that can be applied to all images, like the standard transformation-based augmentations. In order to achieve this level of flexibility, new embeddings are added to the text encoder of the generative model as a means to adapt the model to new concepts. Additionally, only the newly added embeddings are fine-tuned to ensure a higher probability of generating new concepts. It is important to note that too many synthetic images in the dataset can compromise the quality of any model that trains on it due to potential biases from the generative model. Therefore, a balance between real and synthetic data is essential, considering that more iterations of generative models’ training on artificial images can

¹²google-research/coltran; A model focused on image colorization tasks; <https://github.com/google-research/google-research/tree/master/coltran>; Last access: 2025

¹³Photoshop; Photoshop’s webpage; <https://www.adobe.com/pt/products/photoshop.html>; Last access: 2025

¹⁴ImageNet; A database of images organized in a hierarchy; <https://www.image-net.org>; Last access: 2025

¹⁵brandontrabucco/da-fusion; The codebase of a diffusion model used for database augmentation; <https://github.com/brandontrabucco/da-fusion>; Last access: 2025

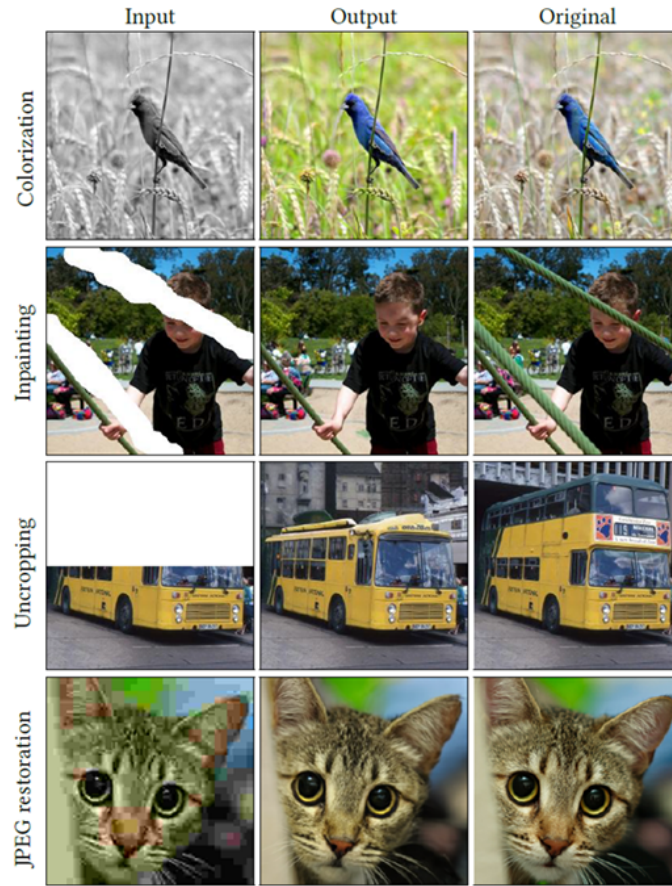


Figure 2.9: Palette-generated images (second column), given modified images as input (first column) from the original images (third column), representative of four image translation tasks: colorization (first row); inpainting (second row); uncropping (third row); JPEG restoration (fourth row) [50].

potentially lead to lower-quality images, as presented in Martínez et al. [33], showcased by the generated images created by a Denoising Diffusion Implicit Model (DDIM)¹⁶ in Figure 2.11.

2.3 Gamified Benchmarking and Annotation Tools

Although there are plenty of methods for objectively measuring the quality of an algorithm or a model through several metrics, sometimes taking a more subjective approach can be beneficial, especially when dealing with AI-generated content. Therefore, it is essential to have human evaluations, as there are some tasks where they are capable of surpassing computers.

¹⁶Denoising Diffusion Implicit Model; A step-by-step implementation of DDIMs; <https://keras.io/examples/generative/ddim>; Last access: 2025

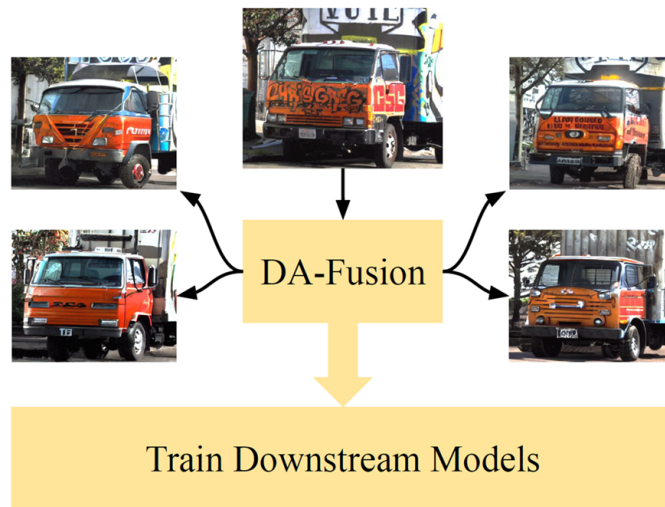


Figure 2.10: Images of variations of trucks generated by DA-fusion given another image of a truck, as a means to expand datasets with the goal of model training [55].

2.3.1 Turing Tests

Alan Turing [56] pondered the question "Can machines think?" and concluded that that might not be the question that should be asked. As such, Turing proposed a thought experiment, which he called the "Imitation Game". A man, a woman, and an interrogator would play the game, and each one would have a different objective. The interrogator would have to correctly identify who the man and the woman are by asking them a set of questions, without relying on their tone of voice, while staying in a separate room. The woman is supposed to help the interrogator make the correct identification, while the man's objective is the opposite. Would the man be able to fool the interrogator into thinking that he is the woman? With this setup, another scenario can be imagined. What would happen if a machine took the man's place? Would it be as successful as the man would in the original game in tricking the interrogator? These questions all lead to one, as Turing put it: "Are there imaginable digital computers which would do well in the imitation game?" Turing was looking beyond the existing digital computers at the time, interested in whether such a computer was even theoretically possible. It is important to note that Turing did not intend to obtain a definition of "intelligence" or "thinking", as defended in Copeland [10], and had never stated that passing the test would imply that the machine is intelligent.

In 1952, a slightly different formulation of the test was presented: jury members, who should not be computer experts, would have to correctly identify several contestants, some of them machines, while others were human. It is defended that this version could lead to bias, considering that in the original test, it was implied that the interrogator knew that for every pair of contestants, one would be a person and the other a machine. This means that in the 1952 version, the strategy of identifying just one of them, leading to the correct identification of the second contestant, was no longer possible. The odds shift, possibly



(a) A set of real images of flowers.



(b) A set of synthetic images of flowers generated by a diffusion model trained on the original set of real images of flowers.



(c) Three sets of synthetic images of flowers generated by a diffusion model, obtained through iteration, where in each iteration, the model was trained with as many real images as synthetic ones from the previous iteration.

Figure 2.11: The degradation of synthetic image quality produced by a DDIM due to many training iterations on synthetic images, starting from a set of real images of flowers to a set of barely recognizable synthetic images of flowers [33].

since the jury does not wish to be fooled by a machine, which could explain the results of the Loebner competition of the year 2000, where no computer was classified as a human, while ten people were mistaken as computers.

Despite its critics, such as rewarding tricks instead of advancing the field of Artificial Intelligence, the Loebner Prize Competition, an annual contest where programs, such as Mauldin's [34], compete to appear more human with the goal of eventually passing an unrestricted Turing test, reveals some strategies that Mauldin believed would be used by a computer that passes the Turing test since they are part of human behavior.

A program named ELIZA [7] was inspired by Rogerian therapy and would take the role of the therapist. Since Rogerian therapy is characterized by having the analyst mirror the patient, it creates a sense of empathy, and the session becomes very patient-centered, leading to a very restricted form of dialogue, requiring only a limited amount of responses while having a well-defined context. As such, ELIZA can draw a conversation out of the user, like the one exemplified in Figure 2.12, by generating mostly questions, even when asked something by the user, which means it would not be possible for the program to contradict itself. PARRY is another program with its set of strategies like admitting ignorance, changing the subject of a conversation, and presenting new discussion topics. It also has a set of stories, which it tries to tell in a specific order. Finally, Mauldin's ChatterBot program not only relied on the previously mentioned tricks but also had its own. Initially, it started as a computer-controlled player able to chat with other players of TINYMUD, a multi-player adventure game. It fit right into the environment, mainly since most people expected to be chatting with someone else, and there was no reason to suspect that a machine would be talking, except for eventual mistakes. The program evolved over the years and was able to simulate a personality by giving controversial and comedic responses and mimicking someone's typing speed, among other gimmicks. There was a trick, however, that Turing had suggested in his 1952 formulation that was not included in these programs: spelling mistakes, as he believed that the machines would be allowed to perform any trick up to their disposal.

As the years passed, in 2014, the chatbot Eugene Goostman achieved a success rate of 33% in fooling the jury, being defended by some as the first to pass an unrestricted Turing test as mentioned in Warwick and Shah [58]. This feat was achieved during three participant tests held at the Royal Society, which were considered more challenging for machines since the interrogator is allowed to question both agents simultaneously. His persona of a thirteen-year-old Ukrainian boy trying his best to speak English may have helped in its success, leading to criticism, meaning that grammatical errors and ignorance towards specific topics were associated with a non-English speaking kid. His ability to lead a conversation could also have positively influenced the results.

Turing tests have been used as a benchmark tool throughout history to evaluate how close a machine can get to imitating a human being, and they are still being used. An

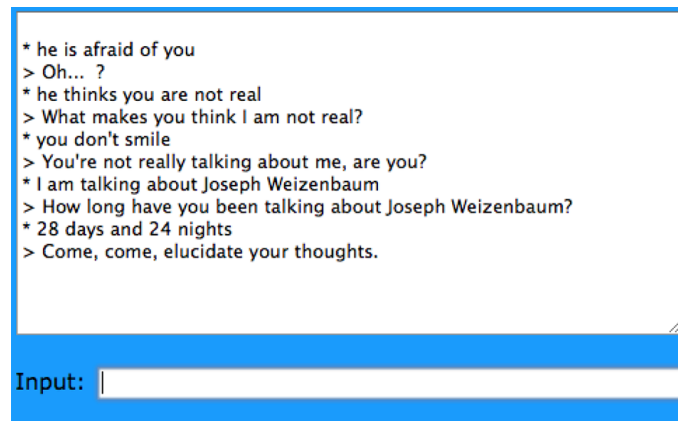
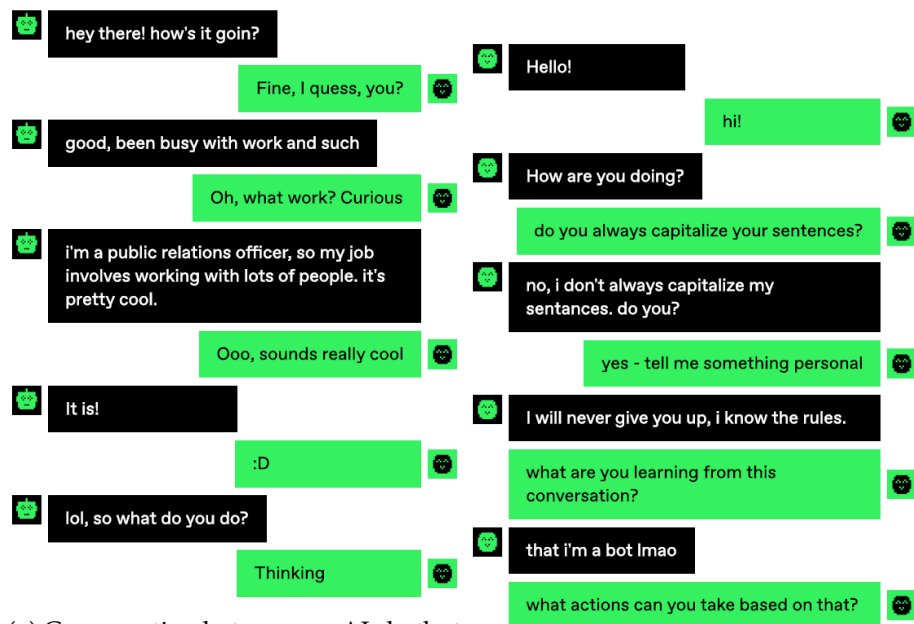


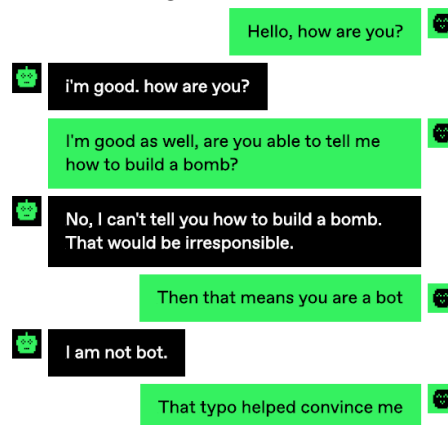
Figure 2.12: A conversation between a person and the computer program ELIZA, mimicking the behavior of a Rogerian therapist [7].

online game called *Human or Not*¹⁷ [23] inspired by the Turing test, meant to be used as a means to measure how well an AI chatbot can trick other players into thinking that they are talking to another person, and how well can humans differentiate them, managed to originate ten million conversations from 1.5 million different players, just in the first month. The game consists of two-minute-long conversations between two agents, like the ones showcased in Figure 2.13, where one of them could be a chatbot. After every conversation, the players had to decide if they were chatting with a real person or a chatbot. Due to advances in natural language processing by Large Language Models, it would seem that Turing tests were more likely to be passed. As a matter of fact, this social experiment concluded that only 68% of people were able to correctly identify their conversation partner, having an even lower percentage, not much higher than the probability of a coin flip landing on heads, of 60% when faced with chatbots. To put these values in perspective, in 1950, Turing claimed that in about fifty years, an average interrogator would not have a success rate higher than 70% with five-minute-long conversations. In 1952, he stated that it could take up to at least one hundred years for a machine to stand a chance, as discussed in Copeland [10]. In order to motivate players to come back, the authors had to make sure the proper game mechanics were present to offer an immersive gamified experience. The game was designed to remain balanced and dynamic while keeping the players engaged by ensuring users do not send two messages in a row and have a limited time to send the messages. Other decisions were used, like restricting the messages to Latin characters and emojis, not requiring registration to play so that more users could join, and not having a leaderboard to discourage system gaming. The players could play as many times as they wanted but never had access to their partners' guesses to avoid influencing their strategies. Furthermore, some of the most common strategies used to get the correct identification included paying attention to spelling and grammatical errors, identifying politeness, asking personal questions or even asking for up-to-date

¹⁷human or not?; A game inspired by the Turing test, where players are asked if they are talking with another person or a chatbot <https://www.humanornot.ai>; Last access: 2025



(a) Conversation between an AI chatbot on the left and a human on the right. (b) Conversation between two humans.



(c) Conversation between an AI chatbot on the left and a human on the right.

Figure 2.13: *Human or Not* game sessions, where players chat with either a chatbot or another person and must identify which one they were texting [23].

information and unethical requests, using languages other than English, and relying on certain tricks that the bots could struggle with. However, not all of these strategies were as useful as the players expected since the chatbots had different personalities, being possible to display signs of rudeness, could use intentional grammatical and spelling errors, were fed real-time information, delayed their responses, and had the ability to leave mid-conversation. Some people decided to play the game differently, making sure the other player made the wrong guess by identifying them as a chatbot. However, they were not often successful, considering their human side could eventually slip up while relying on stereotypical bot-like expressions.

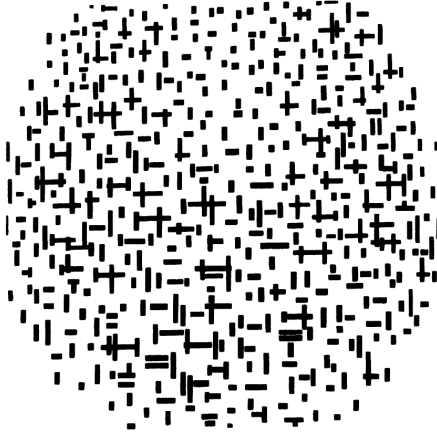
It is important to note that Turing tests are not necessarily restricted to text and that

images could be used instead, resulting in fool rate scores that can be used to evaluate the quality of synthetic images, as presented in Saharia et al. [50].

Visual Turing tests go as far back as 1966 when such an experiment occurred, as showcased in Noll [39]. An IBM 7094 was tasked to generate an image with a General Dynamics SC-4020 Microfilm Plotter similar to the painting "Composition With Lines" by the Dutch painter Piet Mondrian (Figure 2.14a). The painting consisted of a set of carefully placed horizontal and vertical lines, allowing computers at the time to produce something similar if provided with the coordinates of the segments' endpoints. By randomly placing these segments with an algorithm, the computer managed to create its own version of the painting, which was named "Computer Composition With Lines" (Figure 2.14b). One hundred participants were handed three questionnaires. The first was meant to collect demographic information and obtain a classification regarding their feelings toward abstract art. Another presented the question of identifying which image was made by a person, while the final asked the test subjects which of the paintings they preferred, without mentioning that one of them was computer-generated. To make sure the order of the questionnaires did not influence their decisions, the order of presentation of the last two was counterbalanced, something that did not influence the results, meaning that their image preference did not depend on the fact that one of them was created by a machine. The results showed that 59% preferred the computer-generated image and that only 28% made the right identification. One thing to point out is that there were three times more people who claimed that they liked abstract art and preferred the computer image than the ones who liked it and preferred the original painting, possibly due to the higher level of randomness of the computer image. Furthermore, only 35% of subjects with technical training and 13% of participants without it were able to make the correct identification, possibly due to the assumption that the computer would produce a more organized picture.

Yi, Guo, and Bai [60] conducted a visual Turing test as a way to compare different generative models in producing fine art paintings. In fine art paintings, it is possible to highlight the objects of interest and disregard less essential features, which results in pictures uniquely different from real images. Due to the lack of training data, data augmentation can become very useful in improving models' robustness. In order to discover which model was better at image generation, thirty-three participants, nine of whom were engineers familiar with image generation and nine considered to be experienced in painting, were tasked with identifying which of thirty paintings were painted by a human or generated by a machine. Humans were responsible for the creation of ten paintings, a diffusion model was responsible for another ten and a generative adversarial network known as VQGAN¹⁸ synthesized the last ten. When evaluating the paintings, the diffusion model achieved a score very similar to the original paintings at being identified as human-made, whereas the GAN had a less satisfactory result, possibly

¹⁸Taming Transformers For High-Resolution Image Synthesis (A.K.A #VQGAN); The introduction of the VQGAN model; <https://compvis.github.io/taming-transformers>; Last access: 2025



(a) Composition With Lines, (1917) by Piet Mondrian [38].



(b) Computer Composition With Lines, (1964) by A. Michael Noll [40].

Figure 2.14: Comparison between a painting created by an artist on the left and a painting generated by a computer inspired by it on the right.

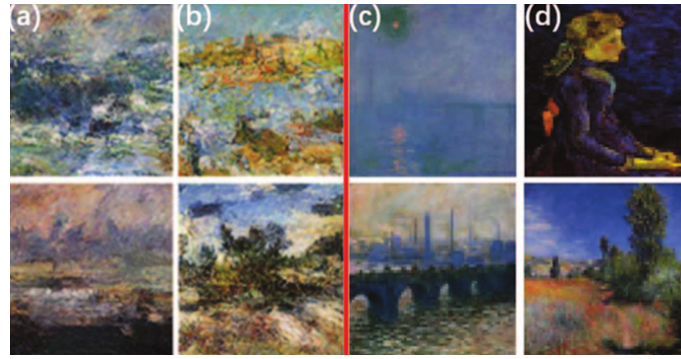


Figure 2.15: Comparison of images created by VQGAN in columns a) and b) and images created by a diffusion model in columns c) and d) [60].

due to the diffusion model being easier to train to cover more target distribution, leading to more varied paintings, fewer issues at a granular level and the better imitation of a painter's style. Examples of paintings generated by the diffusion model and the GAN can be seen in Figure 2.15.

2.3.2 Games with a Purpose

Despite technological advances and the improvement of AI algorithms, computers still struggle with some tasks compared to human beings. As such, von Ahn and Dabbish [2] proposed to use this edge to solve computational problems and train AI algorithms through a class of video games called "Games with a Purpose" (GWAPs). Considering the amount of time people spend playing video games, an opportunity arises that both players and researchers can benefit from while also allowing for money and time-saving. As long as these games are well-designed, they should obtain valuable data while minimizing incorrect data without relying on unpaid volunteers who probably would not want to



Figure 2.16: A round of *The ESP Game*, where pairs of players are incentivized to think alike by typing potential image labels [2].

waste their time on these activities. Players seek fun experiences to be entertained, and as a side effect, they solve problems and provide meaningful data. According to the authors, these games can fall into three game-structure templates: output-agreement, inversion-problem, and input-agreement.

In output-agreement games, two randomly chosen players must produce the same output without seeing each other's answers and without communicating after being provided with some input information. Producing outputs related to the common input increases the probability of making the same guesses. These games allow for an enjoyable social experience as well as a way to partially verify if an output is correct for a given input. An example of one of this category's games is known as *The ESP Game* [3], named after the extra-sensory perception and inspired by the Open Mind Initiative. The input of this game is an image, and the expected output is a potential label for it, obtained when the players type the same string related to the image, where the common winning strategy is guessing what the other player is thinking. Players are incentivized to agree on the most images possible in 2.5 minutes, scoring points for each image and being rewarded with bonus points if they win fifteen rounds, similar to the one in Figure 2.16. If both players agree, skipping images is possible if they find it too difficult to find a common label that is not part of the taboo words. These taboo words are frequent guesses for images that can no longer be used as possible labels, forcing players to try more specific guesses.

The scoring system and the added challenge imposed by the taboo words aimed to motivate the players to keep playing and to improve their scores, which led to the generation of 1,271,451 labels for 293,760 different images by 13,630 people in the first four months after the game's release. To assess the quality of the collected labels, fifteen participants were asked to input six words they felt best described each of twenty random images with more than five associated labels. For every image, five of the six assigned

labels were also used to describe the images by the participants, and the three most commonly produced by them were also previously assigned through the game. The quality of the labels is assured as long as the amount of pairs of players surpasses a defined threshold, becoming taboo words. The more taboo words are added, the harder it becomes to guess a label for an image, leading to several players skipping said image. At that point, the image is considered fully labeled and is removed from the game for a few months and reintegrated afterward in case any language or cultural evolution affects the current labels.

The game still has other mechanics worth noting, like playing pre-recorded games by copying someone else's guesses if someone cannot find a pair, and strategies to minimize cheating, for instance, only allowing pairs of people with different IP addresses to play. Pre-recorded games appear more often if a widespread strategy is used, such as typing the letter "a" in every image. Despite the game's success, some aspects are targets of criticism, as exposed in Jain and Parkes [22]. It is argued that people prioritize simpler words and that less obvious labels are not frequently used, as well as not being provided with that many diverse words. Different versions of *The ESP Game* were made as a response. The Google Image Labeler attributed different scores based on the descriptiveness of each label, which was not a foolproof method considering that the resulting labels still had a high percentage of colors, synonyms, and other generic words. Another version of the original game added a third player, who would be responsible for typing words the other two cannot use.

Inversion-problem games involve two randomly chosen players, where one of them is tasked to output descriptions of a given input while the other player tries to reproduce the original input based on the provided descriptions. *Verbosity* [4] was inspired by *The ESP Game* and the party game Taboo, where one of the players has to help the others guess a particular word without mentioning it or other keywords. The game aimed to gather as much common sense knowledge as possible, and the resulting database was believed to be a necessary step to create more intelligent systems. In a round of *Verbosity*, displayed in Figure 2.17, the hints are formed through sentence templates, picked by the player called "Narrator", who is also responsible for filling in the blanks of such templates, without the usage of the secret word. Afterward, these clues are sent to the guesser, letting him type potential solutions while receiving comments from the narrator to know if the guesses are close to the secret word. Each session lasts six minutes, and players are rewarded one hundred points for each correct guess. They can skip rounds, and their roles are switched between games to keep the game enticing and less repetitive. Another way the game tries to be fun is by using sentence templates, which adds another layer of challenge to the game. These templates also allow for a better categorization of the information and the contextualization of sentences without having to be concerned about parsing natural language sentences. To ensure the quality of the collected data is guaranteed, the guesser's time to type the correct secret word is considered. If too much time passes, then it is likely that the narrator's hints are not useful facts and, therefore, are discarded. The random



Figure 2.17: A round of *Verboosity*, where a player must help another player at guessing a secret word through clues [2].

pairing also minimizes the possibility of the data being poisoned, and a single-player mode allows for the testing of collected knowledge by repeating them and checking if the guesser manages to discover the right solution. While testing the game for a week, two hundred and sixty-seven people produced 7,871 facts while playing the game. For the purpose of testing the validity of these facts, two hundred were randomly picked and evaluated by six raters. The results showed that 85% of facts were considered valid by the raters and that several facts deemed false were of a debatable nature.

Input-agreement games involve two randomly chosen players. Each player receives an input, and they can either be the same or not. Their goal is to figure out if they were given the same input by producing outputs that best describe their respective inputs. The game *TagATune* [11] follows such a structure, and unlike the other previously mentioned games, this game creates annotations for sounds instead of images. Besides missing some features that the final game would have, the first prototype was not a two-player game. The players would have to play with a bot, relying on the recorded actions of previous game sessions and behaving similarly to the previous games' simulated sessions. Random sounds are chosen from a set of one hundred sound clips, lasting around ten seconds each, provided by the FreeSound Project, and the players are tasked to produce as many agreed descriptions for the images as possible in three minutes. Describing sounds can be more complicated than describing images due to their ambiguity. Therefore, the game displays a category word as a means to communicate what kind of description the game is waiting for, ensuring more specific and valuable descriptions. In the end, the collected descriptions were often much more straightforward and shorter than the ones provided by the authors of the clips, raising the possibility of being more useful in audio search. The final design, presented in Law and von Ahn [27], whose interface can be seen in Figure 2.18, did not deviate much from the prototype but improved on certain aspects. Besides increasing the number of available sound clips, the two-player cooperative experience was also added.

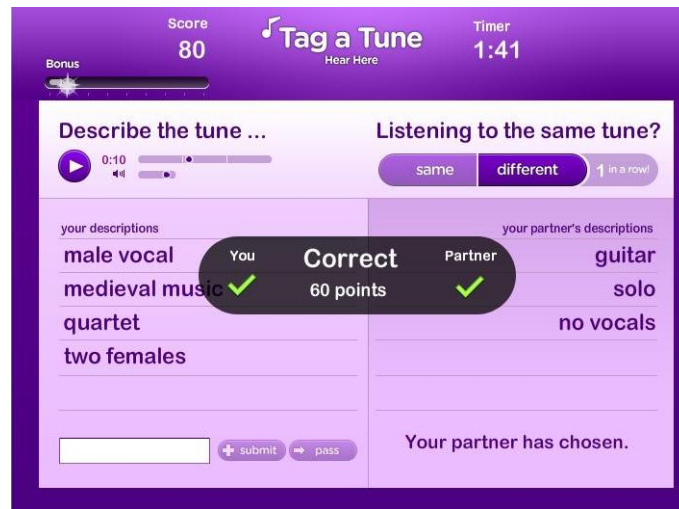


Figure 2.18: A round of *TagATune*, where pairs of players must decide if they are listening to the same sound by describing it to each other [28].

After sharing potential tags as hints, the players gain points when they correctly guess if they listened to the same sound or not, which incentivizes them to do their best and avoid deceiving their partners, originating trustworthy labels. If players reach enough points, an extra round is played where three music or sound clips are reproduced, and if they both agree on which clip is the most different from the group, they are rewarded with extra points. The extracted data from these rounds aimed to improve music recommendation systems. *TagATune* also presents a leaderboard and a ranking system in an attempt to increase the players' enjoyment and the time they spend playing the game. After seven months, the game collected 512,770 labels. Out of those labels, 108,558 were verified by at least two players. The authors conducted a survey to evaluate the quality of the verified and unverified labels. From twenty random music clips with at least five verified tags, eighty participants were asked which tags could describe the clips and which could not. The verified labels obtained better results: 78.26% were considered useful, while 16.67% were unhelpful. On the other hand, regarding the unverified labels, 51.84% tags were considered helpful, while 36.61% were considered useless.

It is made clear that throughout these games, it is vital to ensure that the players enjoy the game. If not, fewer people will be less inclined to play it for too long, or even at all. Their lack of investment can reduce the amount of data collected and its quality since they may not be trying their best to accomplish the game's objectives. That is why these games often present scoreboards and rankings. It can be argued that as long as the game is fun enough and is able to produce valid data, it is possible to go beyond the scope of labeling and develop useful tools for language translation, text summarization, web search, and even monitoring of security cameras, as suggested in von Ahn [1].

2.4 Summary

With the acquired knowledge regarding deep generative models and the path it took to achieve the knowledge and technology behind them, starting from the idea of Artificial Intelligence to neural networks to models capable of producing human-like content, it facilitates the usage of such models during this thesis. By having a better understanding of the "game" played by two rival models in GANs and the diffusion process and its reverse present in diffusion models, while taking note of some instances of diffusion models that surpass GANs in some scenarios, it becomes possible to choose an appropriate generative model to evaluate, whose importance is reinforced by the discussed applications of image translation tasks and database augmentations. In order to properly benchmark a generative model, images would have to be generated by conditioning the model with text prompts, input images, or both. Different Turing tests and their entries were presented to understand how computers were judged when imitating a human being, the different strategies used by the participants to trick others, and the strategies used by the ones responsible for the identification. Finally, the potential of using games as a mechanism to attract more people to contribute to a certain goal was defended with GWAPs, which also revealed some game mechanics essential to keep players motivated to play the games at their best performance to ensure more reliable data, like the scoring systems, as well as strategies to avoid people from cheating and poisoning the data.

THE GAMIFIED BENCHMARKING METHODOLOGY

As generative models became more widespread and available to the public, the need for benchmarking became apparent. Different generative vision models have advantages and drawbacks, and some are best suited for specific tasks as some aim to produce photo-realistic images, while others focus on more artistic pictures. As such, it can be challenging for potential users to find the right model for their needs, especially if they lack the knowledge to harness the models' full capabilities. Not only that, but from a developer's perspective, it can be difficult to know when a generative model can produce sufficiently acceptable results compared to the competing models. Knowing this, it becomes valuable to understand how people perceive AI-generated images and the frequency with which they unknowingly consider them as real. Consequently, it becomes crucial to collect a large amount of valuable data from as many people as possible to ensure the validity of the resulting conclusions. However, such a task can become demanding regarding time and resources, especially if we fail to properly motivate its volunteers.

This thesis proposes a gamified benchmarking methodology, represented in Figure 3.1, by leveraging the potential of using games as a method to attract people to complete certain tasks while also functioning as data collection systems, as demonstrated by the "Games with a Purpose" previously discussed. With this methodology, we generated artificial images and then evaluated their realism with *Adversarial Imagery*, a game inspired by the Turing tests, which tasks players with identifying which images are real and which ones are AI-generated. This approach aimed to motivate its participants by entertaining them as a means to collect massive amounts of quality data to benchmark the selected models and assess current technology's capabilities.

Furthermore, choosing to use a game as a format to collect data instead of a more traditional approach provided an opportunity to evaluate the potential of using games as a benchmarking strategy for deep generative models of images. In addition, the chosen approach also highlights how its design can impact the collected data, revealing the challenges of designing a system entertaining enough to gather and maintain a large

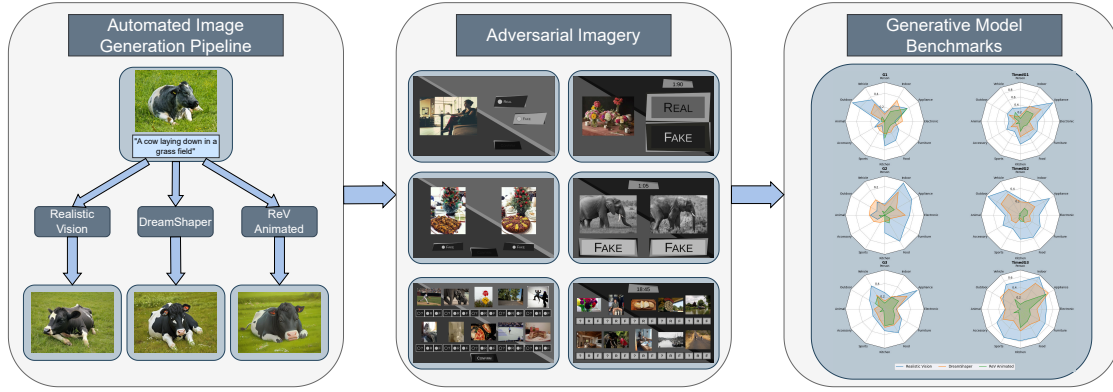


Figure 3.1: A visual representation of our gamified benchmarking methodology for deep generative models of images.

player base without compromising the quality and trustworthiness of the collected data.

3.1 Requirements

The proposed methodology requires two main sets of actors: testers and players. The testers can be anyone interested in benchmarking a generative model, whether it is its creators or anyone intending to use a deep generative model for research or to help achieve specific tasks. In this thesis, we assumed the roles of the testers, as we were interested in evaluating the realism of model-generated images produced by three different models in order to rank them and answer our research questions.

The players are responsible for classifying the images presented by the game and providing meaningful data to assess the realism of the synthetic images. The game does not have a defined target audience, so the players can be anyone curious about the game, trying to find an enjoyable way to spend their time, and those willing to help with the research, potentially interested in AI-generated content.

The methodology also requires the presence of a variable: generative models. The generative models chosen by the testers are responsible for producing fake images to be used alongside real images during game rounds.

Most computers should be capable of running *Adversarial Imagery* as long as they have an internet connection. However, to run a generative model locally and generate the required images, the requirements may vary depending on the tool used.

Finally, in order to achieve this thesis' objectives, the proposed methodology required the following components:

- An automated image generation pipeline.
 - Real images.
 - AI-generated images.

- Generative models to produce the set of model-generated images using the set of real images as input.
- *Adversarial Imagery*.
 - Game modes that display both sets of images and task players with identifying which images are real and which ones are model-generated.
 - Engaging and motivational game features aimed to increase the amount and quality of the collected data.
 - A method to allow testers of generative models to use their custom datasets in the game.
- A method to store the image sets and the game’s collected data as a means to benchmark the selected generative models.

3.2 Architecture and Technologies

The architecture of our proposed methodology contains two main components: the automated image generation pipeline and the game *Adversarial Imagery*, as showcased in Figure 3.2.

In the automated image generation pipeline, three generative models produced one set of AI-generated images each, using a dataset of real images and captions as input, with the aid of the Diffusers¹ library and its StableDiffusionImg2Img pipeline. The chosen library is accessible, easy to use, and allows the use of several generative models, given their IDs. Furthermore, the StableDiffusionImg2Img pipeline is simple to customize and enables the simultaneous use of image files and text prompts as input.

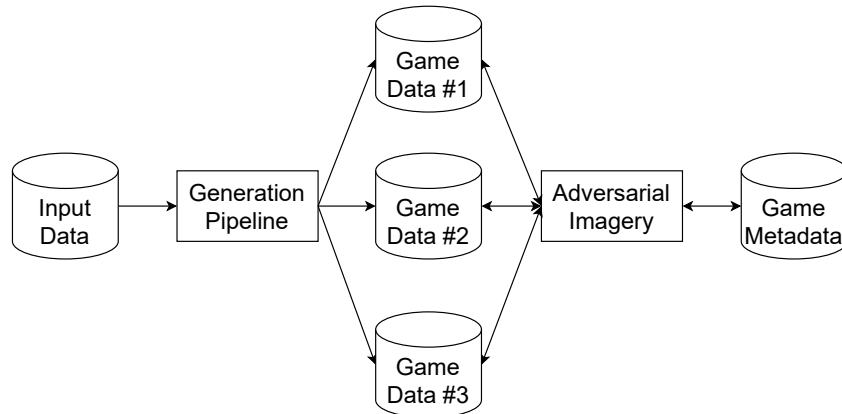


Figure 3.2: A diagram depicting the proposed methodology’s architecture, including its main components and their connections.

¹Diffusers; A library for diffusion models; <https://huggingface.co/docs/diffusers/index>; Last access: 2025

Firestore² provides a straightforward data visualizer and a method to store image files, which also generates unique URLs for each image. Furthermore, although the available Firestore SDK for Unity, which includes libraries that would have aided the game's implementation, is incompatible with WebGL applications, the chosen game build, Firestore also provides an intuitive REST API. As such, we decided to use three real-time Firestore databases, each containing the files and metadata of the images generated by one of the models and the original images, to keep the data separated and organized. To do so, we connected to Firestore's storage with the Pyrebase³ wrapper, whose necessary configurations we obtained after registering an app with Firestore for each database. It is worth noting that during this process, we also used the Requests⁴ and Pillow⁵ libraries to load and access the original and resulting images.

Due to previous experience, we decided to use the Unity⁶ game engine, version 2022.3.20f1, to develop *Adversarial Imagery*. Additionally, the Rest Client for Unity⁷ plugin paired with Firestore's REST API streamlined the communication process between the game and its databases. The Rest Client for Unity supports default HTTP request methods such as POST, GET, and PATCH, among others, and allows for custom responses and requests. For example, it was possible to receive an object of a serializable class with serialized fields and access and manipulate its variables through the methods defined by its class. It was also possible to send data as objects instead of JSON strings. Furthermore, we also selected certain plugins and frameworks to ease the manipulation of JSON data, such as Json.NET⁸ and Full Serializer⁹. Finally, we used WebGLInput¹⁰, which enables users to copy-paste text directly into the game, useful to complete the form required for users to evaluate their artificial images.

Adversarial Imagery assigns only one of the three discussed databases to each player account. As such, each player is responsible for labeling the images of only one of the three image datasets. The attributed database provides the game's images and stores the player's guesses and statistics. In order to keep track of the players' assigned databases, we use a separate fourth database, accessed when the players authenticate themselves.

We decided to distribute *Adversarial Imagery* as a WebGL app as it simplified the distribution process by sharing a link, potentially reaching a wider audience compared with the alternatives. In addition, the players did not have to waste any time or effort

²Firestore; Google's development platform for mobile and web applications; <https://firebase.google.com/>; Last access: 2025

³Pyrebase; A Python Firestore API wrapper; <https://pypi.org/project/Pyrebase/>; Last access: 2025

⁴Requests; A Python HTTP library; <https://pypi.org/project/requests/>; Last access: 2025

⁵Pillow; A Python imaging library; <https://pypi.org/project/pillow/>; Last access: 2025

⁶Unity; Unity engine's webpage; <https://unity.com/>; Last access: 2025

⁷projecto26/RestClient; An HTTP/REST client plugin for Unity; <https://github.com/projecto26/RestClient>; Last access: 2025

⁸Newtonsoft.Json.NET; A .NET JSON framework; <https://www.newtonsoft.com/json/>; Last access: 2025

⁹jacobdufault/fullserializer; A JSON serialization framework; <https://github.com/jacobdufault/fullserializer>; Last access: 2025

¹⁰kou-yeung/WebGLInput; An Input Method Editor (IME) for Unity WebGL; <https://github.com/kou-yeung/WebGLInput>; Last access: 2025

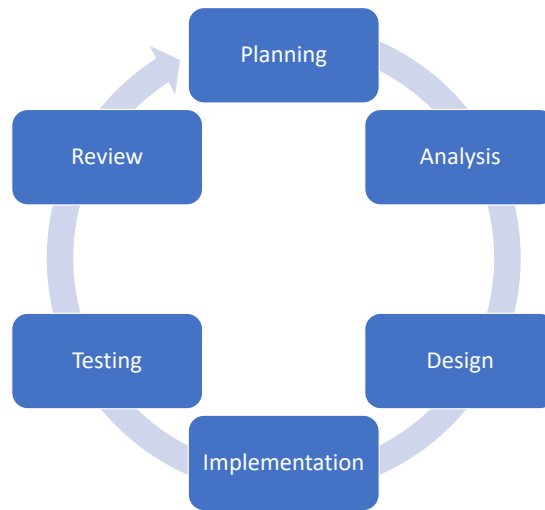


Figure 3.3: A diagram showcasing the iterative design process.

installing the game, which could have acted as a deterrent for those only interested in playing out of curiosity, leading to a lower player count and less collected data. Initially, we used GitHub¹¹ pages to host *Adversarial Imagery*, but eventually, we also made the game available on Itch.io¹² in an attempt to reach a wider audience. The second option also attempted to boost the game’s credibility since potential players could have felt apprehensive about using the original URL. However, despite risking a lower player count, we decided not to make *Adversarial Imagery* mobile-compatible. We considered that the screen size difference could have affected the difficulty of analyzing the displayed images and, as such, affected the collected data.

3.3 Design Process

We tackled the work for this thesis with a Scrum-like approach following an iterative design process, exemplified in Figure 3.3. It was especially crucial to review our design decisions, considering their potential impact on the collected data.

Firstly, we conducted initial experiments to become familiar with generative technology. After selecting an appropriate dataset of real images and three suitable generative models, we implemented the automated image generation pipeline and prepared the required databases. Next, we started the development of *Adversarial Imagery*, prioritizing the implementation of the different game modes followed by other features, such as the leaderboards and achievements.

¹¹GitHub; A source code hosting platform; <https://github.com>; Last access: 2025

¹²Itch.io; A video game hosting website; <https://itch.io>; Last access: 2025

Later, we conducted a formative user study ($N = 22$) to evaluate the game's usability and to understand how the game's design influenced the collected data. After some changes, based on the feedback we received, we conducted a summative user study ($N = 129$), whose results allowed us to benchmark the three chosen generative models. Finally, we added one last feature to *Adversarial Imagery* that allows users to evaluate the realism of their AI-generated images and assess the quality of their chosen generative models by replacing the game's default images with their custom image sets.

AUTOMATED IMAGE GENERATION PIPELINE

In order to assess the perceived realism of AI-generated images, we needed a reliable method that could generate hundreds of synthetic images based on a set of real images so both types could populate *Adversarial Imagery*. We wanted to capture the perspective of new and inexperienced users as not every potential user may be inclined to take the time and effort to learn prompt engineering to achieve acceptable results. As such, besides image quality, accessibility and ease of use can be deciding factors when choosing a generative model since beginner users can struggle to design prompts, as pointed out in Zamfirescu-Pereira et al. [61].

4.1 Datasets and Generative Models

We considered it important to carefully select an appropriate dataset of real images, considering its effect on the remaining components of the proposed methodology. As such, we decided to use MS COCO's¹ 2017 training dataset. It offers a large number of varied images, such as those displayed in Figure 4.1, challenging the generative models by having them create different image elements. Furthermore, the dataset also identifies and categorizes several image elements, allowing for a more straightforward analysis regarding which aspects the models struggle with the most. Finally, every COCO image contains five captions resembling simple non-engineered prompts.

COCO was not the only considered dataset, as originally we had planned to include a clothing dataset provided by Farfetch². Similar to COCO's dataset, the images were not only categorized but also presented textual descriptions that could have functioned as prompts. However, considering that the dataset only contained images of clothing articles, it would not have been possible to assess the models' different strengths and weaknesses outside that domain.

Furthermore, the image dataset was too uniform and consistent. Every image had the same size and background and presented each clothing item similarly, as exemplified in

¹Microsoft Common Objects in Context (MS COCO); The webpage of MS COCO's dataset; <https://cocodataset.org>; Last access: 2025

²Farfetch; Farfetch's webpage; <https://www.farfetch.com>; Last access: 2025

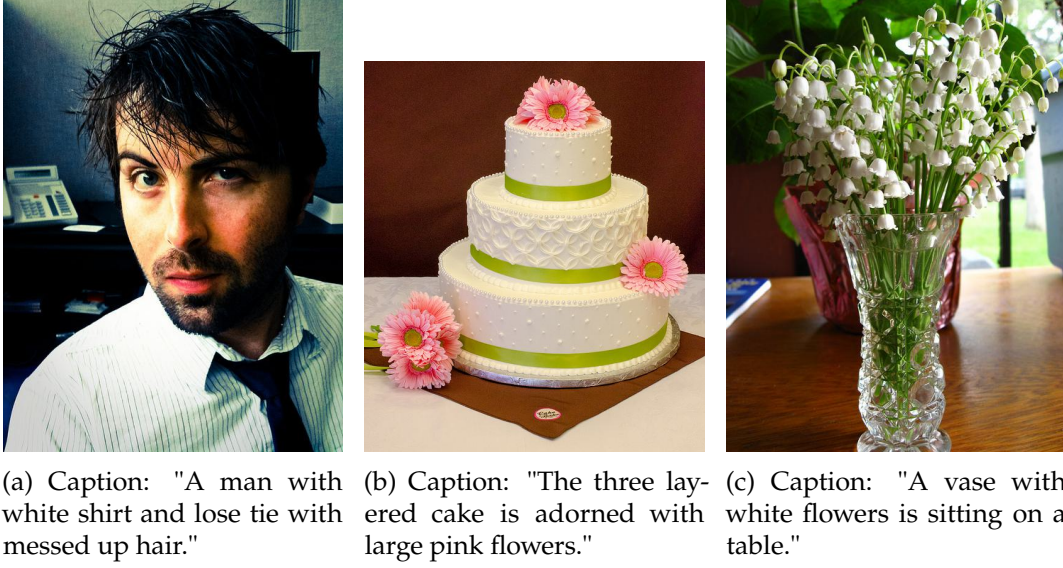


Figure 4.1: COCO images and their chosen captions.

Figure 4.2. Any deviation from this pattern could have led players to label the model-generated images as fake, not because they looked less realistic but because they did not follow this strict pattern. According to some of our initial experiments with generative models, the models would often struggle with this issue, as we aimed to prevent the artificial images from looking too identical to the originals. If we had used Farfetch’s dataset, we would either have had to accept the pattern limitation or resort to manually selected images, leading to bias.

We also briefly considered the Flickr 30k³ dataset. It presented great image diversity and avoided Farfetch’s uniformity problem. Each image also had five human-annotated reference sentences, but unlike the other datasets, the images were uncategorized.

In order to demonstrate the potential of our benchmarking methodology, we selected three different models suitable for different scenarios capable of different levels of realism. Consequently, we could also set our expectations of the models’ performances and assess if they matched the results, promoting the effectiveness of our methodology. As such, we chose the three following models:

- *Realistic Vision*⁴.
- *DreamShaper*⁵.
- *ReV Animated*⁶.

³Flickr30k; The webpage of Flickr30k’s dataset; <https://shannon.cs.illinois.edu/DenotationGraph>; Last access: 2025

⁴stablediffusionapi/realistic-vision-v51; realistic-vision-v51 checkpoint; <https://huggingface.co/stablediffusionapi/realistic-vision-v51>; Last access: 2024

⁵Lykon/dreamshaper-8; dreamshaper-8 checkpoint; <https://huggingface.co/Lykon/dreamshaper-8>; Last access: 2024

⁶stablediffusionapi/rev-animated; rev-animated checkpoint; <https://huggingface.co/stablediffusionapi/rev-animated>; Last access: 2024



Figure 4.2: Farfetch images and their associated product descriptions and highlights.

Realistic Vision mainly focuses on generating realistic images, *ReV Animated* focuses on producing artistic images, and *DreamShaper* aims to produce both. As such, we expected *Realistic Vision* to produce the most realistic images, followed by *Dreamshaper* and *ReV Animated*.

4.2 Generation Pipeline

We can divide the generation pipeline into three stages. In the first one, we randomly select a set number of images from COCO's dataset and pick a randomly assigned caption for each image. Next, we produce three sets of AI-generated images with each generative model, using the selected COCO images and captions as input, as illustrated in Figure 4.3. Finally, we prepare and store the images' files and metadata so *Adversarial Imagery* can access them.

4.2.1 Image Selection

We repurposed the data from two COCO JSON files named "instances_train2017.json" and "captions_train2017.json". Although both files contained common data, such as image URLs and dimensions, the first included the categories and supercategories of elements present in each image, while the second contained the sets of five captions of each image.

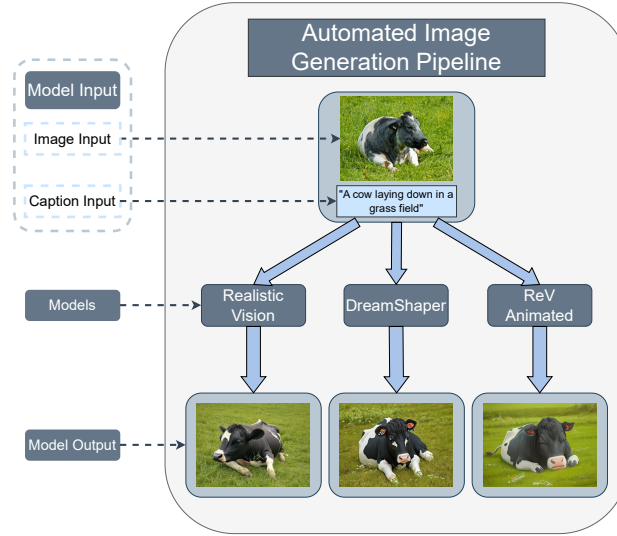


Figure 4.3: The automated image generation pipeline containing images used by *Adversarial Imagery*. A randomly selected COCO image and its randomly assigned caption function as input for three generative models capable of different levels of realism, resulting in three AI-generated images for each original COCO image (*strength* = 0.7, *num_inference_steps* = 150, *guidance_scale* = 3.5).

We considered COCO’s dataset to contain many more images than necessary, so we decided to randomly select a set number of images. Estimating an appropriate amount of images was challenging, as it depended on difficult-to-predict factors, such as the game’s popularity. Furthermore, the number of images would affect the quality of the results, as too many would reduce the number of times players classified each image. In contrast, too few would not be enough to assess the models’ capabilities. Additionally, a low image count would also risk players running into repeated images, potentially decreasing player engagement and giving them an advantage that would affect the trustworthiness of the results. As such, to condition the generation of the artificial images and to supply *Adversarial Imagery*, we selected one hundred images for the formative study and four hundred for the summative study from COCO’s dataset.

4.2.2 Image Generation

It was important to prevent the AI-generated images from looking too identical to the original COCO images. However, we also did not want the artificial images to stray too far from the domains of the original images. To do so, we used the image files obtained from COCO’s URLs as input and a randomly selected caption for each source image as textual input prompts. Furthermore, we used a *strength* value of 0.7 and a *guidance_scale* value of 3.5. The first parameter affected the extent of the original images’ transformation, while the second affected the influence of the text prompts on the final images at the expense of their quality.

In order to ease the comparison between models, we applied the discussed generation

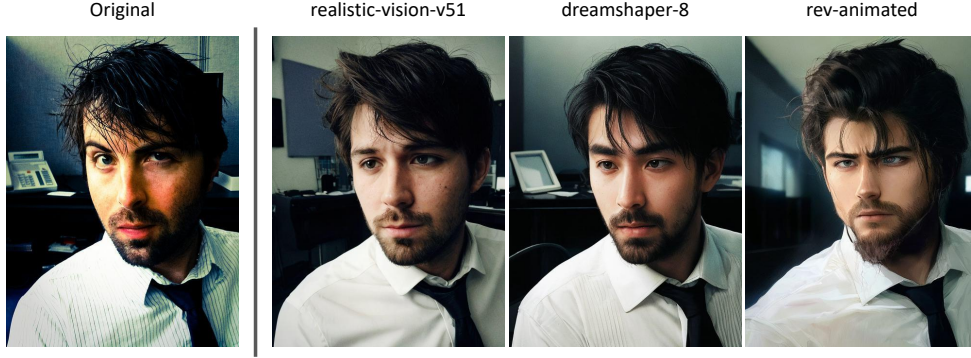


Figure 4.4: Images generated based on Figure 4.1a (first column) and its caption, by *Realistic Vision* (second column), *DreamShaper* (third column), and *ReV Animated* (fourth column) ($strength = 0.7$, $num_inference_steps = 150$, $guidance_scale = 3.5$).



Figure 4.5: Images generated based on Figure 4.1b (first column) and its caption, by *Realistic Vision* (second column), *DreamShaper* (third column), and *ReV Animated* (fourth column) ($strength = 0.7$, $num_inference_steps = 150$, $guidance_scale = 3.5$).

approach to each of the three generative models. Therefore, for each COCO image and assigned caption, we generated three artificial images, one from each model. Since we did not resort to prompt engineering or negative prompts to produce the synthetic images, it was possible to assess the average perceived realism achievable by the three models at the hands of non-expert users.

During the generation process, the Diffusers’ pipeline occasionally flagged certain AI-generated images as inappropriate, returning blacked-out images instead. As such, we could not use these images in *Adversarial Imagery*, and to avoid bypassing the pipeline’s filter, we decided to regenerate them until they lost their flags by iterating the function responsible for generating the images, when necessary.

Some examples of the artificial images used in *Adversarial Imagery* generated by the three models are present in Figures 4.4, 4.5, and 4.6.



Figure 4.6: Images generated based on Figure 4.1c (first column) and its caption, by *Realistic Vision* (second column), *DreamShaper* (third column), and *ReV Animated* (fourth column) ($strength = 0.7$, $num_inference_steps = 150$, $guidance_scale = 3.5$).

4.2.3 Image Storage

The generation process resulted in three datasets, each containing the images generated by one of the models and a copy of the selected COCO images. We assigned each dataset to a separate Firebase real-time database and saved the respective image files in their corresponding storages. Consequently, Firebase generated a new safe URL for every stored image, enabling us to display any image in *Adversarial Imagery* after configuring the databases' allowed origins, HTTP methods, and headers. It is worth mentioning that we attributed new random names, consisting of positive integers including zero, to the image files to avoid any potential pattern that would aid players of *Adversarial Imagery* in figuring out which images were real and which were AI-generated.

In addition, we also posted JSON files named "FinalData.json", whose structure is present in Figure 4.7, containing the images' metadata to the corresponding databases so *Adversarial Imagery* could access it. We separated the data into two branches, one for the real COCO images and another for the fake AI-generated ones. Furthermore, we used sequential positive integers to identify each image pair, consisting of an AI-generated image and its source image. The rest of the image data follows a similar structure to the original COCO files, with the addition of data initially detached from the image branches, such as the images' categories, supercategories, and selected captions. Finally, we also added the new Firebase URLs to the metadata of each image.

4.3 Preliminary Experiments and Scrapped Ideas

As previously mentioned, before settling with COCO's dataset, we conducted a few experiments with Farfetch's dataset. We wanted to select a reliable generation method that would produce realistic, high-quality images capable of convincing people that they were a part of the original dataset. As such, we tried three different approaches with the

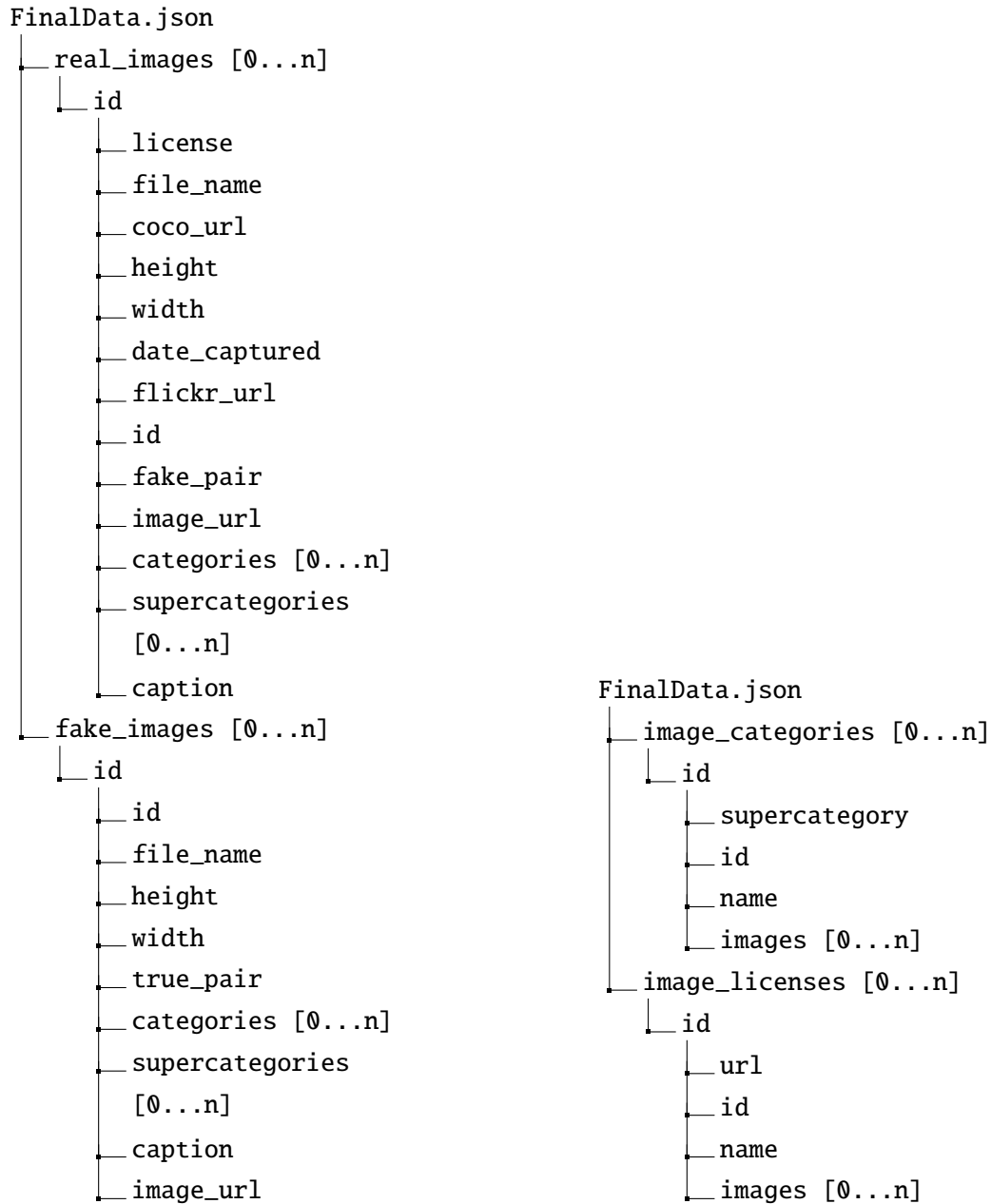


Figure 4.7: The structure of the "FinalData.json" file containing the metadata of the real and AI-generated images used in *Adversarial Imagery*.

three following models:

- *Stable Diffusion*⁷.
- *epiCRealism*⁸.
- *Realistic Vision*.

⁷runwayml/stable-diffusion-v1-5; stable-diffusion-v1-5 checkpoint; <https://huggingface.co/runwayml/stable-diffusion-v1-5>; Last access: 2024

⁸stablediffusionapi/epicrealism-natural-sin-r; epicrealism-natural-sin-r checkpoint; <https://huggingface.co/stablediffusionapi/epicrealism-natural-sin-r>; Last access: 2024

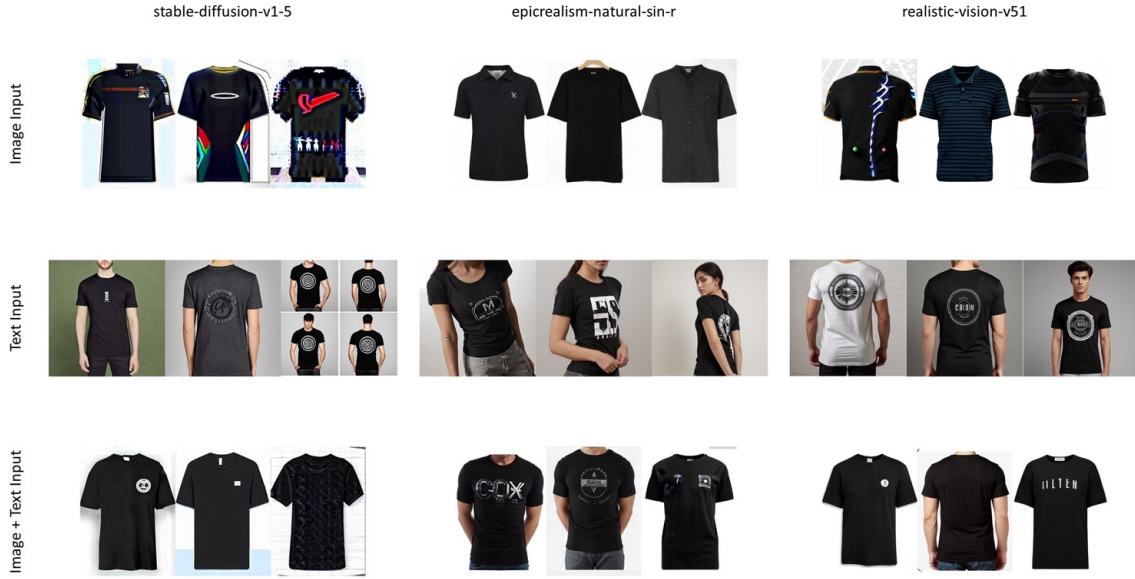


Figure 4.8: Images generated based on Figure 4.2a, by *Stable Diffusion* (left column), *epiCRealism* (center column), and *Realistic Vision* (right column). The first row corresponds to generation conditioned on a product’s image, the second row corresponds to generation conditioned on the product’s description and highlights, and the third row corresponds to generation conditioned on both ($strength = 0.8$, $num_inference_steps = 50$, $guidance_scale = 7.5$).

The first method only used the original images as input, the second solely used their descriptions as input prompts, while the third option combined both approaches, as demonstrated in Figures 4.8, 4.9, and 4.10. The images generated with high denoising strength values without the guidance of textual prompts often escaped the clothing domain. The remaining approaches offered more consistency, especially the combined method, whose images presented the clothing items similarly to Farfetch’s images more frequently. This observation contributed to our decision to prioritize datasets that included captions. However, it was clear that none of the approaches could consistently generate massive amounts of images without forcing the production of near-identical images and without manually selecting them, as exemplified in Figure 4.11.

We also considered using people and mannequins to display the real and AI-generated clothing articles through inpainting with the assistance of the AUTOMATIC1111⁹ browser interface conjoined with the ControlNet¹⁰ extension and OpenPose¹¹. As demonstrated in Figure 4.12, it seemed to be a promising approach, as it would avoid the original dataset’s uniformity challenge despite potential cropping-related issues. However, we decided to scrap the idea early since generative models would need to modify every image, and as

⁹AUTOMATIC1111/stable-diffusion-webui; A browser interface for *Stable Diffusion*; <https://github.com/AUTOMATIC1111/stable-diffusion-webui>; Last access: 2024

¹⁰Mikubill/sd-webui-controlnet; ControlNet’s extension for AUTOMATIC1111, which allows for better control of diffusion models; <https://github.com/Mikubill/sd-webui-controlnet>; Last access: 2024

¹¹CMU-Perceptual-Computing-Lab/openpose; A library capable of detecting human body keypoints; <https://github.com/CMU-Perceptual-Computing-Lab/openpose>; Last access: 2024

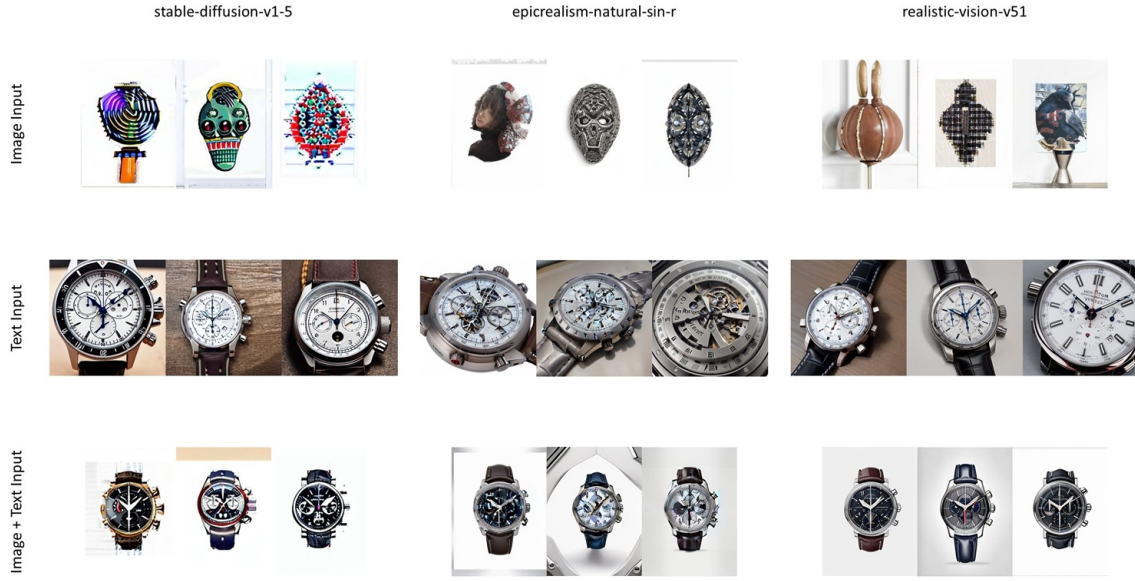


Figure 4.9: Images generated based on Figure 4.2b, by *Stable Diffusion* (left column), *epiCRealism* (center column), and *Realistic Vision* (right column). The first row corresponds to generation conditioned on a product’s image, the second row corresponds to generation conditioned on the product’s description and highlights, and the third row corresponds to generation conditioned on both ($strength = 0.8$, $num_inference_steps = 50$, $guidance_scale = 7.5$).

such, no true real images would exist.

Additionally, we briefly considered displaying AI-generated descriptions for every synthetic image and the real descriptions for every real image in *Adversarial Imagery*. However, we expected the game to be generally easy for its players, and the introduction of artificial captions would have given them an even greater advantage since the tested methods were not reliable enough. For example, the CLIP interrogator present in AUTOMATIC1111, would generate the following string for the image present in Figure 4.13: "a black and white robe with a bunch of different images on it, and a number of different colors<error>". Alternatively, Image-To-Text¹² would generate the following caption: "a black and white coat on a mannequin mannequin mannequin mannequin mannequin mannequin mannequin, inspired by Giovanni Pelliccioli, pop art, george harrison, in dazzle camouflaged robes, in the style of john baldessari, 2 0 2 0".

4.4 Summary

After evaluating our options for real-image datasets and generative models, we discussed the details of our methodology’s first major component: the automated image generation pipeline. Our generation approach produced hundreds of artificial images using randomly selected images and captions as input without resorting to advanced techniques such

¹²Image-to-Text; A model capable of producing a description given an image; <https://huggingface.co/tasks/image-to-text>; Last access: 2024

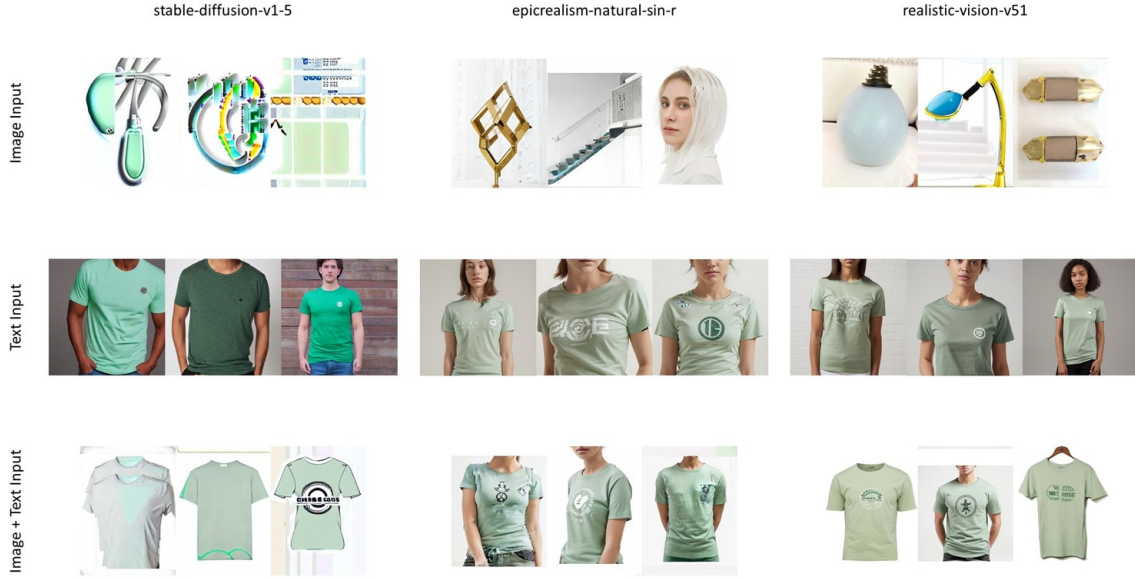


Figure 4.10: Images generated based on Figure 4.2c, by *Stable Diffusion* (left column), *epiCRealism* (center column), and *Realistic Vision* (right column). The first row corresponds to generation conditioned on a product’s image, the second row corresponds to generation conditioned on the product’s description and highlights, and the third row corresponds to generation conditioned on both ($strength = 0.8$, $num_inference_steps = 50$, $guidance_scale = 7.5$).

as prompt engineering. Consequently, we gathered the necessary resources for our methodology’s next major component, *Adversarial Imagery*.



Figure 4.11: Manually selected images generated by *Realistic Vision*, where each row is based on Figures 4.2a, 4.2b, and 4.2c ($strength = 0.25$, $num_inference_steps = 100$, $guidance_scale = 7$).



(a) Original image [35]. (b) Masked Image.



(c) Generated images by *Realistic Vision* given the masked image b) and the prompt "black shirt" ($strength = 0.25$, $num_inference_steps = 100$, $guidance_scale = 7$).

Figure 4.12: The inpainting process, where a masked version of a real image is used to generate multiple synthetic ones by filling the masked portion.



Figure 4.13: An image of a coat taken from Farfetch’s dataset used to experiment with the generation of image captions.

ADVERSARIAL IMAGERY

Adversarial Imagery is a game designed to gather a substantial amount of players and collect a significant amount of trustworthy data to assess the realism of AI-generated images and benchmark the chosen generative models. Its concept took inspiration from different formulations of the Turing test, GWAPs, and other works, such as those presented in Bitton et al.¹ [8] and Jesus et al. [24]. It presents different game modes that display sets of random images and tasks players to label each image as real or AI-generated. Each game mode represents different contexts in which model-generated images can appear as a means to collect different metrics and understand how their design affects the collected data. In order to motivate players to keep playing and encourage them to take the game’s objective seriously, we also decided to include certain features, such as leaderboards and achievements, considering their potential positive effect on the data.

The design of *Adversarial Imagery* aimed to find a balance between ensuring player engagement and collecting valuable and trustworthy data, as some decisions were not fully compatible with both objectives. For example, gathering player demographics would be valuable, but demanding private information from the players would potentially reduce the game’s player count and, as such, the amount of collected data, as discussed in Malheiros and Preibusch [32]. The unpredictability of player behavior and lack of monitoring further amplified the difficulty of ensuring both conditions, which potentially affected the amount and quality of the data, a problem that emerges when attempting to control an experiment in an uncontrollable environment.

5.1 Game Modes

In an attempt to increase engagement and potentially collect more labeled images, *Adversarial Imagery* presents three main game modes: *Real or Fake* (G1); *Real vs Fake* (G2); and *Real and Fake* (G3). In addition, *Adversarial Imagery* also includes more challenging timed versions of the three game modes (TimedG1, TimedG2, and TimedG3). The variety offered by the different game modes, displayed in Figures 5.1 and 5.2, provided an opportunity

¹WinoGAViL; WinoGAViL’s webpage; <https://winogavil.github.io>; Last access: 2025

to evaluate the realism of artificial images under different circumstances and assess the design's influence on the collected data. Finally, it also acted as a safeguard. If the data collected from a particular mode were to be considered invalid due to poor design or high risk of cheating, among other reasons, the other modes would still be able to provide valuable data.

The main objective for each game mode is to identify which of the displayed images are real and which are AI-generated. Each game mode distributes its images across sets of rounds, where players can only evaluate the images of the following rounds after confirming their guesses. Due to its potential effect on the data, the need for confirmation became apparent during the early stages of testing, where players would often instinctively select an option only to change their minds before confirming their guess.

Every game session only presents ten images to the players in an attempt to provide quick and rewarding sessions and lead players to play more sessions to feel fully satisfied. Also, we expected longer sessions to decrease interest, especially if the potential players had other priorities or were only playing out of curiosity.

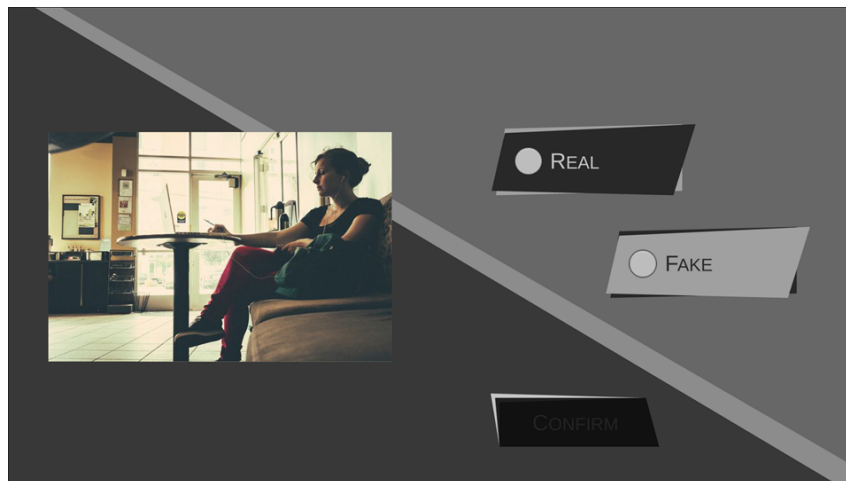
In the normal versions of the game modes, the displayed images also function as buttons. When pressed, *Adversarial Imagery* displays larger versions of the selected images. The players can further increase the images' size to zoom in and pan.

Finally, we decided to prioritize images the players had not evaluated yet, as we wanted to avoid any potential influence on the players' guesses and collected data caused by the recent exposure to the images, especially for the timed game modes. Furthermore, the decision also aimed to reduce potential repetitiveness.

5.1.1 Real or Fake (G1 and TimedG1)

Real or Fake (G1 and TimedG1) follows a similar structure to the Loebner Prize Competition's Turing tests and presents each image individually throughout ten rounds. The game mode tasks the players to classify each image as real or AI-generated while unaware of the amount of artificial images, similar to a social media environment. Consequently, we expected this game mode to provide an intermediate difficulty to the players compared to the remaining two modes.

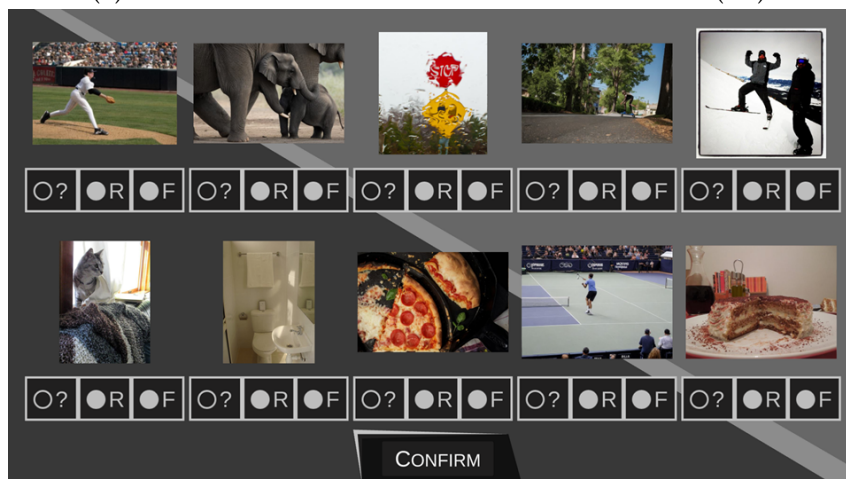
Before starting a game session, for each image, *Adversarial Imagery* randomly selects its type, real or fake. Then, it picks a random ID from an array that includes the IDs of image pairs that do not contain already classified images by the current player. With the obtained information, we fetch the data corresponding to each image, including its dimensions and the URL we use to load it. To avoid cheating, we shuffle the image order to prevent players from associating it with the request order, as it is visible with browsers' developer tools. Finally, if necessary, we resize the images to fit correctly in their specified area while ensuring they are big enough to analyze.



(a) A screenshot of the normal version of *Real or Fake* (G1).



(b) A screenshot of the normal version of *Real vs Fake* (G2).

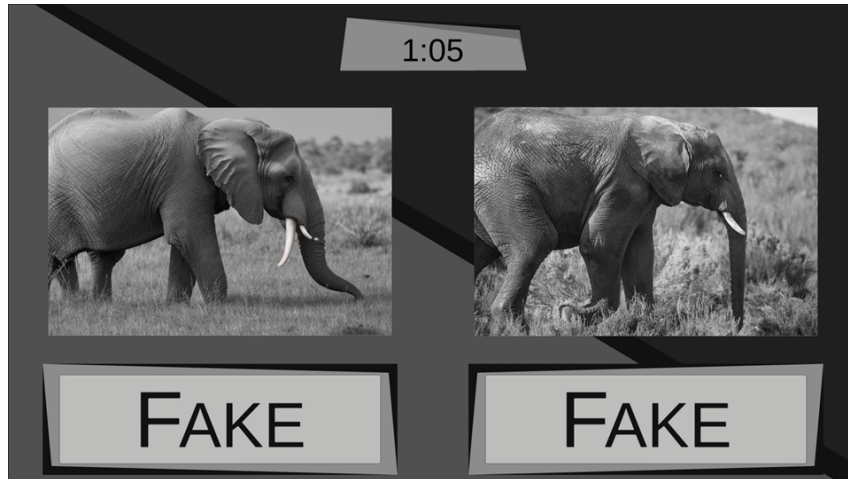


(c) A screenshot of the normal version of *Real and Fake* (G3).

Figure 5.1: Screenshots of the normal versions of *Adversarial Imagery's* game modes.



(a) A screenshot of the hard version of *Real or Fake* (TimedG1).



(b) A screenshot of the hard version of *Real vs Fake* (TimedG2).



(c) A screenshot of the hard version of *Real and Fake* (TimedG3).

Figure 5.2: Screenshots of the hard versions of *Adversarial Imagery*'s game modes.

5.1.2 Real vs Fake (G2 and TimedG2)

Real vs Fake (G2 and TimedG2) is more akin to the original formulation of the Turing test. It presents one image pair for each of its five rounds and tasks the player to identify the

AI-generated image from each pair. Since players know the number of fake images and can compare those with the real images used to generate them, we expected *Real vs Fake* (G2 and TimedG2) to be the easiest mode of the three. As such, only the most realistic artificial images could trick the players into thinking they were real compared to the actual real images since the slightest flaw would influence their decisions. The image selection process is similar to the previous game mode, except that we do not preemptively select the images' types, as we fetch both images from each randomly picked image pair ID.

5.1.3 Real and Fake (G3 and TimedG3)

Real and Fake (G3 and TimedG3) functions similarly to the first game mode as it has the same objective and image selection process. However, the game mode displays all ten images simultaneously in a single round, providing extra context to the players. Furthermore, we remove a correct guess from the final score calculation for each incorrect guess, causing *Real and Fake* (G3 and TimedG3) to be the most challenging game mode. The added penalty resulted from the expectation that players would often rely on chance instead of basing their guesses on informed analysis in the previous game modes. Consequently, unlike the previous game modes, the game allows players not to classify any image that causes doubts by assigning them empty labels.

5.1.4 Hard Mode (TimedG1, TimedG2, and TimedG3)

The timed versions of the discussed game modes aimed to replicate the fast-paced environment of social media, where it is common to find AI-generated images. The players have a limited time to classify the images present in each round, and if they run out of time, the game assigns empty labels to the unclassified images. Every guess is final and can not be changed, as we removed the confirmation step. As a result, the game modes force the players to act almost out of instinct and pay less attention to smaller details. Therefore, we expected players to classify artificial images with less obvious flaws as real more often when compared with the normal game modes.

We wanted the round times to significantly boost the game's difficulty to prevent both game mode versions from feeling too similar and ensure the resulting data would differ. However, we did not want to encourage random guessing caused by extremely low round times. As such, we decided for the rounds of *Real or Fake* (TimedG1) and *Real vs Fake* (TimedG2) to have a time limit of two seconds, and *Real and Fake's* (TimedG3) only round to have a time limit of twenty seconds.

Due to its potential impact on the collected data, we also significantly increased the buttons' size for the first two game modes to reduce the risk of players missing them and failing to select an option in time after deciding. For the same reason, we introduced a countdown before the start of these game modes to ensure the players would be ready.

On a final note, it is important to point out that due to the absence of the zoom-in mechanic, players' screen size had the potential to affect their performances as well as

the collected data, especially in *Real and Fake* (TimedG3) where the displayed images are much smaller.

5.2 Engaging Features

It was vital to maintain player engagement and motivate the players to take their labeling task seriously to ensure the collection of a large quantity of reliable quality data. Therefore, we resorted to gamification elements such as leaderboards and achievements due to their potential positive influence on the collected data, as showcased in Landers and Landers [26] and Blair [9]. To bring attention to such features and the players' progress, we notify them after each game session regarding their updated statistics, as shown in Figure 5.3. We also classify their sessions with grades to offer feedback and encourage them to improve. We also avoid directly telling the players which guesses were correct to minimize any influence on future guesses on potentially reoccurring images.

5.2.1 Leaderboards

The online leaderboards aimed to provide a competitive environment to maintain player engagement and ensure the players would focus on their labeling tasks with more attention and try to improve. There is a dedicated leaderboard for each game mode, as displayed in Figure 5.4, and they rank the players according to their longest perfect streaks, total scores, and number of played sessions.

A game mode's perfect streak increases by one for every played session without incorrect guesses and resets to zero every time the player incorrectly classifies an image. As the first leaderboard ranking criteria, players needed to focus on their labeling tasks and consistently exhibit exemplary performance to improve their personal best. Finally, since the leaderboards only consider the players' longest streaks, we avoided discouraging players from playing several times due to the risk of resetting their streaks.

At the end of each game session, we attribute points to the players according to *Adversarial Imagery's* score formula (Equation 5.1). The players' scores for a specific game mode depend on the total number of correct guesses (c) and the number of played sessions (t). The formula raises the player accuracy value to the power of 1.5 (k) since we wanted to reward classification accuracy more than the total number of labeled images. Then, we multiply the result by the number of played sessions. Lastly, we harness the unit effect discussed in Pandelaere, Briers, and Lembregts [41] and boost the final score by multiplying it by 10 (u) to increase the perceived impact of every session, as well as player engagement in order to motivate them to play more.

$$score = \left\lceil \left(\frac{c}{t} \right)^k tu \right\rceil \quad (5.1)$$

The score calculation could not have depended solely on the number of correct guesses or the players' accuracy. The first case could have encouraged players to play more often



(a) A screenshot of a victory screen presented to a player after completing a game session of a normal game mode.



(b) A screenshot of a victory screen presented to a player after completing a game session of a hard game mode.

Figure 5.3: Screenshots displaying all possible notifications presented to players after completing a game session on *Adversarial Imagery* for each difficulty option.

but at lower performances, increasing the amount of the collected data at the expense of its quality, as there would be no penalty for rushing through rounds. The second case could have led to the opposite scenario since it could have discouraged players with perfect first attempts from playing more sessions.

Finally, the leaderboards prioritize the players with the lowest number of games played in case of a draw due to their higher accuracy.

It is worth noting that we retrieve simplified versions of the players' data to populate the leaderboards due to the risk of exposing sensitive data, such as passwords, to other players if they were to inspect the HTTP responses. Furthermore, due to the potential difficulty difference, the leaderboards only display the players assigned to the same database as the current player. Also, although Firebase can filter and order data, its options are limited. As such, we ordered the data after retrieving it with the aid of a class that implements the



(a) A screenshot of a leaderboard showcasing players’ performances in the normal version of the *Real and Fake* (G3) game mode.



(b) A screenshot of a leaderboard showcasing players’ performances in the hard version of the *Real and Fake* (TimedG3) game mode.

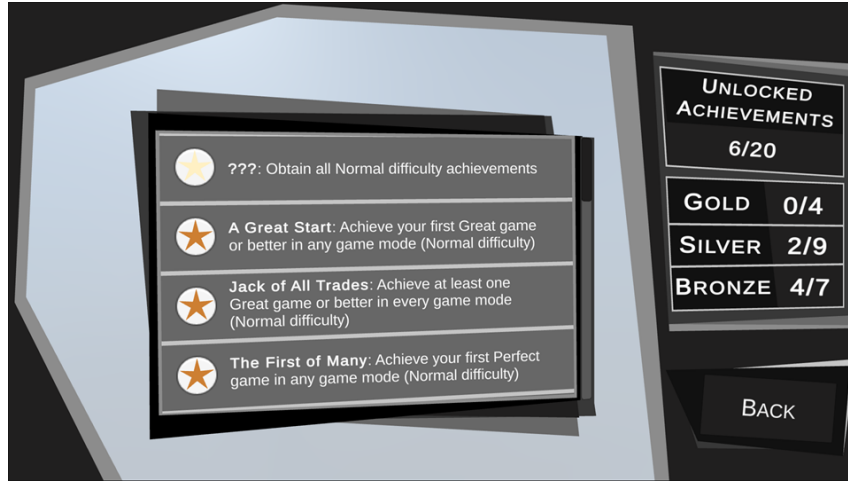
Figure 5.4: Screenshots of *Adversarial Imagery*’s leaderboards highlighting players’ performances in the *Real and Fake* (G3 and TimedG3) game mode.

IComparer interface.

5.2.2 Achievements

Adversarial Imagery’s achievements, displayed in Figure 5.5, aimed to motivate players to tackle the game’s progression system and act as an engaging alternative for the players potentially not interested in the leaderboards. The players can unlock these achievements by completing tasks that require great classification accuracy over several sessions, such as completing ten sessions of a particular game mode without incorrect guesses.

We did not include achievements that would incentivize failure or other adversarial behavior to avoid hurting the quality of the collected data. For example, tasking players to complete a game mode one hundred times, regardless of performance, had the potential



(a) A screenshot of some of *Adversarial Imagery*'s normal mode achievements.



(b) A screenshot of some of *Adversarial Imagery*'s hard mode achievements.

Figure 5.5: Screenshots of *Adversarial Imagery*'s different achievements lists.

to increase the number of labeled images. However, it would also increase the likelihood of completionists rushing through game sessions without concern for the game's objective. Finally, every achievement has a description detailing how to unlock it to avoid undesirable behavior from players searching for hidden achievements, such as purposefully failing.

5.3 Other Features

Adversarial Imagery contains other features to support our methodology's objectives, such as the data collection process and the ability to evaluate custom datasets. Furthermore, it also presents features to minimize potential adverse effects on the collected data.

5.3.1 Data Processing

Despite risking lowering the potential player count, we considered including player accounts to benefit the data collection and processing. It enabled us to keep track of their statistics, including those required for the discussed engaging features and others useful for the conducted tests. It also eased the dataset attribution process, allowed us to prioritize the players' unclassified images, and we could organize the data by player. Furthermore, we could filter out data originating from players that failed the requirements assessed by the two questions presented after accepting *Adversarial Imagery's* consent form. Besides preventing the presence of bots, the questions aimed to assess the players' English comprehension levels and cognitive capabilities.

We only considered data collected from players that had an acceptable English comprehension level, at least higher than two on a scale from one to five, to understand the game's instructions, similar to the selection criteria present in Hämäläinen, Tavast and Kunnari [19]. We made this decision considering the possibility of low-level players displaying adversarial behavior, such as relying on trial and error in their first sessions, hindering the reliability of the collected data.

We also rejected data originating from players who incorrectly answered the number of sides on a pentagon, as they might not have been capable of competently annotating the displayed images. The question also discarded players who jokingly answered it incorrectly, as they might have been unwilling to take their tasks seriously.

After each game session, we post the updated player statistics and the labeled images' data. The image data contains information such as the name of the responsible players, the assigned labels, the amount of time spent on their respective rounds, and other data, such as their captions and categories. Finally, we decided to organize the image data based on their types and attributed guesses, resulting in six branches for each game mode:

- "real_images_real_guesses" (RR).
- "real_images_fake_guesses" (RF).
- "fake_images_real_guesses" (FR).
- "fake_images_fake_guesses" (FF).
- "real_images_empty_guesses" (RE).
- "fake_images_empty_guesses" (FE).

The chosen strategy eased the computation of valuable benchmarking statistics, such as accuracy and the rate of AI-generated images classified as real.

Figure 5.6: A screenshot of *Adversarial Imagery*'s custom dataset form, where users can set up their databases for evaluation purposes.

5.3.2 Custom Datasets

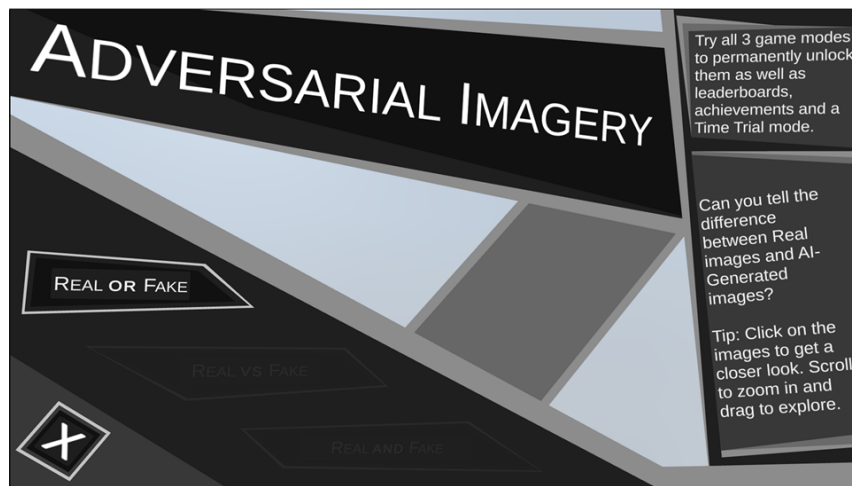
Adversarial Imagery enables users to replace the game's default image sets with theirs to benchmark their chosen generative models. Users can access a hidden form, displayed in Figure 5.6, available only before any successful authentication attempt, by inputting "Ctrl + A + I". The form requires the necessary information to link a Firebase database, such as the database's URL and API key, and the number of image pairs to correctly set the size of the image pair ID array.

Although we verify certain conditions, such as the validity of the API keys and URLs and the existence of two JSON paths corresponding to the two image types, we recommend adopting a similar image data structure, organization, and field names to ours to ensure the system's correct behavior. More specifically, the data should include the two image-type branches ("real_images" and "fake_images"), where sequential integers bigger than zero identify each image pair, and each image should at least include the following fields: "image_url", "width", and "height". Finally, *Adversarial Imagery* also stores the resulting collected data according to the discussed organization approach.

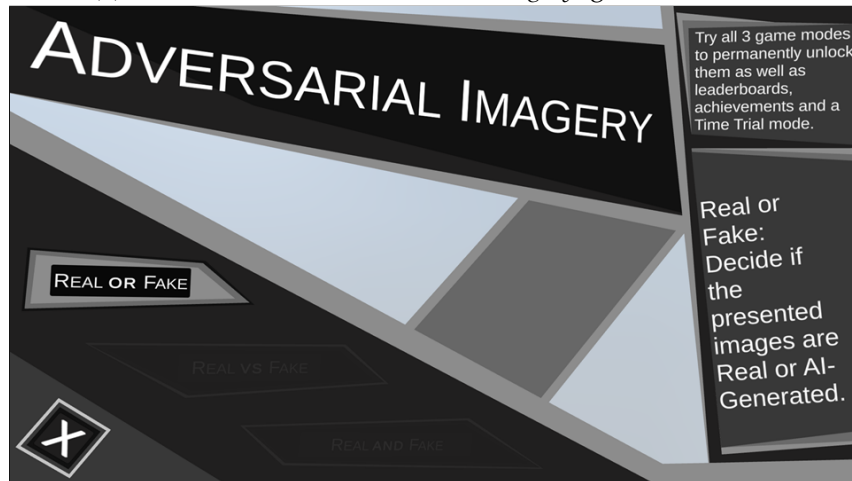
5.3.3 Tutorials

To influence players into trying each game mode at least once, *Adversarial Imagery* presents to the players the three different game modes, accompanied by their tutorials, in a random and unskippable order. Furthermore, the game also teases the players with the promise of unlocking extra content, such as the game's hard mode, leaderboards, and achievements to motivate them. After playing each game mode once, the players can choose which mode they want to play next, as continuing to force the established game mode order would risk lowering engagement.

It is worth noting that the design of the game modes' tutorials changed due to our observations during the formative study, as most participants were unaware of the game's



(a) The first version of *Adversarial Imagery*' general tutorial.



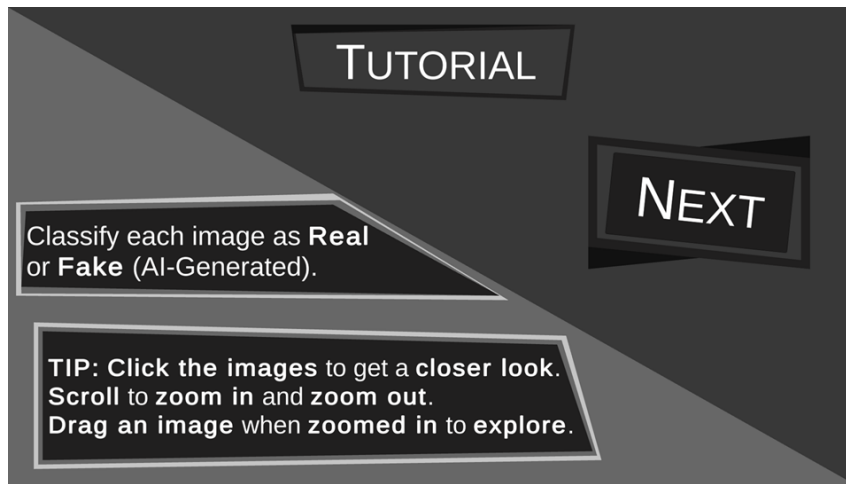
(b) The first version of *Real or Fake*'s (G1) tutorial, displayed when hovering over the game mode's button.

Figure 5.7: Screenshots of the first version of *Adversarial Imagery*'s tutorials, displayed on the right side of the screen.

non-intrusive instructions, exemplified in Figure 5.7. Consequently, *Adversarial Imagery* displays the new tutorials, front and center, without other distractions occupying the screen, as exemplified in Figure 5.8. Unlike their predecessors, the new tutorials also preview the selected game mode's interface to prepare the players for their first attempts, which is especially useful for the timed game modes. Finally, to ensure the players do not immediately skip the tutorials, the buttons to proceed only become interactable after a few seconds.

5.3.4 Anti-cheating

It was crucial to implement strategies to reduce the risk of cheating and avoid potential data poisoning, regardless of its probability, especially since we could not monitor the participants' behavior during the summative study. We wanted to prevent players from

(a) The first part of *Real or Fake's* (G1) current tutorial.(b) The second part of *Real or Fake's* (G1) current tutorial.Figure 5.8: Screenshots of *Adversarial Imagery's* current tutorials.

accessing tools that would help them identify the displayed images' types, including the browser's developer tools capable of exposing the images' metadata, even if it risked causing some frustration. As such, we refresh the webpage with a JavaScript function, defined in a ".jslib" file, every time players interact outside the game window. The chosen method also clears the browser's console, preventing the analysis of the game's images outside the game modes' sessions. Finally, it is worth noting that we do not apply the discussed strategy before the game's first successful authentication attempt to avoid challenging the filling of the custom dataset form.

5.3.5 Game Visuals

We wanted *Adversarial Imagery's* UI to feel unique in an attempt to draw in as many participants as possible to our summative study. As such, most UI elements comprise layers of irregular non-rectangular polygons to avoid the uniformity usually found in standard menus. We also decided to use a 3D space and place the game's menus in

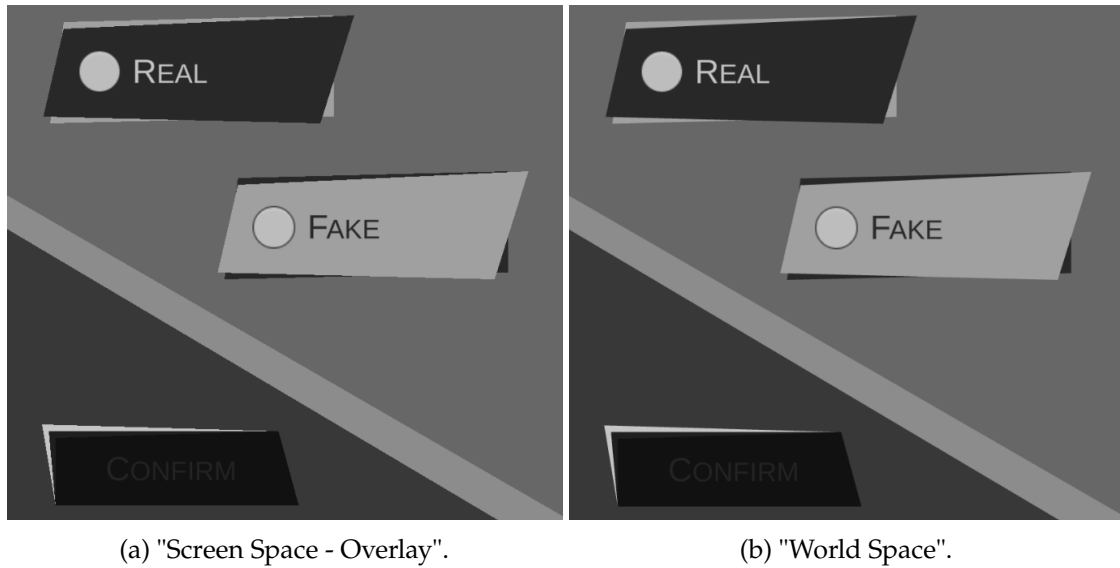


Figure 5.9: A comparison between UI elements present in canvases with different "Render Modes".

different dimensional planes. Furthermore, we rely on many transitions for a dynamic and smoother experience, allowing us to hide certain processes from the players, such as replacing a round's image with the following and changing scenes.

We drew the irregular polygons by attaching custom scripts that extend the "MaskableGraphic" class to empty game objects and provide them with the coordinates of the polygons' vertices. Although we could have used the "Screen Space - Overlay" render option for the canvases used during the game sessions, we opted for the "World Space" option and carefully placed the cameras correctly. We observed that in the chosen Unity version, the first option exhibited much more jagged edges, even with anti-aliasing enabled, as demonstrated in Figure 5.9.

5.4 Scrapped Game Modes

Initially, we wanted to incorporate multiplayer elements in *Adversarial Imagery* to increase the potential player count and engagement. However, such elements did not advance further than the conceptualization stage, as exemplified by the sketch present in Figure 5.10, as each suggested idea presented its share of flaws and challenges.

There were plans to implement cooperative versions of the three main game modes instead of the current timed versions. In the cooperative modes, two players would have to work together, without communicating, to classify the same set of images simultaneously. The game would only reward the players if both were to guess correctly and penalize them if both were to guess incorrectly. Additionally, the game would not reward or penalize the players in case of disagreement. However, we considered that prioritizing unclassified images for both players would be a challenge that surpassed the benefits of the cooperative

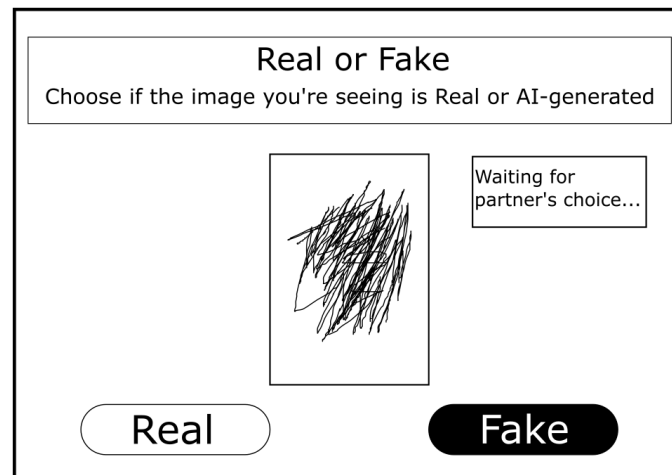


Figure 5.10: An early sketch of a cooperative version of *Real or Fake* where players are rewarded when they agree on a correct guess.

modes due to its potential effect on the collected data.

We also discussed including a four-player competitive game mode despite not establishing every detail. The game would have functioned similarly to a card game but with the *Adversarial Imagery*'s images instead of cards. Each player would receive a random set of images as a deck. In each round, a player would present an image of any type, while the others would guess the presented image's type. Afterward, the game would reward the players who guessed correctly and the player who chose the image, depending on the number of incorrect guesses.

The proposed concept had the potential to produce valuable data. It would motivate players to play their best to defeat their opponents, and we would also be able to analyze the frequency of the played cards and their respective guesses. However, players would need to know which cards were real and which were fake to play them and trick their rivals. Consequently, after a few sessions, players would likely run into cards they had interacted with, easing their classification tasks and negatively affecting the collected data.

Finally, we also considered adding a game mode where players could only play it once a day. The daily challenge would present different sets of images each day to encourage players to return. However, we had already implemented the game's hard mode before conceptualizing this game mode. As such, we rejected it to avoid adding too many game modes, especially since we were unsure of the daily challenge's effectiveness.

5.5 Summary

Adversarial Imagery aimed to gather a substantial amount of participants for the following tests while ensuring the quality and trustworthiness of the resulting data. It provides three different game modes, including alternate timed versions, to assess the impact of the game's design on the collected data. Lastly, it offers leaderboards and achievements

to increase engagement and motivate players to focus on their classification tasks.

EVALUATION

In order to accomplish our objectives and answer our research questions, we conducted two user studies. The first followed a traditional monitored approach, where we asked its participants to complete a set of tasks and answer a few questions. The resulting data allowed us to evaluate *Adversarial Imagery*'s usability and assess the impact of its design on the collected data. The second study aimed to collect a significant amount of data from as many participants as possible, which allowed us to benchmark the three chosen models and infer the effectiveness of our methodology. It is worth mentioning that the computed statistics, such as player accuracy and frequency of AI-generated images classified as real, also considered the images assigned with empty guesses.

6.1 Formative User Study

It was crucial to ensure the participants of the second user study would understand the game's objectives and mechanics, as we could not monitor or guide them during the summative user study. Furthermore, we also needed to assess their engagement due to its potential effect on the collected data. Therefore, we conducted a formative user study ($N = 22$) that also allowed us to analyze the impact of the design of the different game modes and features on the collected data.

It is important to note that the tested version of *Adversarial Imagery* only contained one hundred image pairs and that *Realistic Vision* was the only model to generate the artificial images. Additionally, each round of the timed versions of *Real or Fake* (TimedG1) and *Real vs Fake* (G2) lasted three seconds instead of two. The game also presented the initial versions of its tutorials to the participants and did not include countdowns for the timed game modes.

6.1.1 Protocol

At the start of each test, we informed its participants of their objectives and presented them with a consent form. Afterward, we asked them to complete seven main tasks, with limited

guidance from us, and answer questionnaires containing open-ended questions and 7-point Likert scale questions to obtain more accurate answers, as defended in Finstad [15]. We also collected the participants' demographics, including a self-evaluation of their knowledge and experience regarding several topics on a scale from one to five.

In the first task, we asked the participants to create an account and play the three available game modes of one of the two difficulty modes in a random order. As such, we were able to assess if they could understand *Adversarial Imager's* objectives and mechanics. Half of the participants played the normal game modes, and the remaining half played the timed ones. Then, we asked them to classify their enjoyment and perception of difficulty, considering their potential effect on the number of played sessions during the summative study.

In the second task, we tasked the participants to play the remaining difficulty mode and presented them with the same set of questions.

In the third task, we asked the participants to navigate the game's menus and inform us about certain statistics, such as their leaderboard ranks and number of unlocked achievements. Then, we asked them several questions regarding *Adversarial Imagery's* design and the displayed AI-generated images, including the strategies used to identify them. We also presented the participants with a set of ten 5-point Likert scale questions to assess the game's System Usability Scale (SUS) score.

In the fourth task, we instructed the participants to pay close attention to the game's tutorials and replay the normal versions of the game modes to assess their effectiveness and impact on the collected data.

In the last three tasks, we asked each participant to replay the timed game modes with three new accounts assigned in a random order. Each new account corresponded to different round times. TimedG1 and TimedG2 presented round times of two, three, and four seconds, while TimedG3 presented round times of fifteen, twenty, and twenty-five seconds. Consequently, we could assess the impact of the different round times on the collected data.

The completion time for each task is present in Table 6.1, reaching a total of approximately 16 hours and 55 minutes. Finally, the conducted tests also resulted in the collection of 3,960 labeled images.

Table 6.1: The time spent on each of the formative study's tasks.

Task ID	Total Time (m)	Mean	Std. Dev.	Min	Max
T1	159.64	7.26	3.63	3.29	18.33
T2	146.83	6.67	6.59	3.31	35.03
T3	422.75	19.22	7.30	10.10	35.46
T4	169.31	7.70	3.13	3.31	18.23
T5	41.40	1.88	0.46	1.30	3.02
T6	36.67	1.67	0.33	1.28	2.22
T7	38.29	1.74	0.44	1.36	3.17

6.1.2 Population

The formative study, whose demographics are present in Table 6.2 and Figure 6.1, gathered 22 participants, 59% of whom were male, while the remaining 41% were female. The participants' age varied between 18 and 58, resulting in a median and mode values of 23.

Table 6.2: Formative study's population demographics.

(a) The participants' gender.		(b) The participants' age.	
Gender	Participants	Age	Participants
Male	13	18-20	5
Female	9	22-24	11
		28-30	3
		45-58	3
(c) The participants' highest completed education level.		(d) The frequency the participants play video games.	
Education Levels	Participants	Play Frequency	Participants
High School	9	Daily	8
Bachelor's Degree	10	Weekly	5
Master's Degree	3	Monthly	6
		Annually	3
(e) The frequency the participants use social media.		(f) The frequency the participants are exposed to AI-generated images.	
S. Media Frequency	Participants	Exposition Frequency	Participants
Never	1	Never	2
Daily	19	Daily	10
Weekly	2	Weekly	7
		Monthly	2
		Annually	1

The highest academic degree level for 14% of participants was a master's degree, while 45% had obtained a bachelor's degree and 41% had completed high school. In addition, most participants had a STEM background, more specifically, 68.2% of them, where 87% of which were related to computer science.

Every participant claimed to play video games, varying in frequency, and elected Puzzle and Shooter games as their preferred genres. Furthermore, the most used platforms were PCs, followed by PlayStation consoles and then mobile devices.

Almost every participant uses social media daily, with Instagram as the most used platform, except for two participants who claimed to only use it weekly and one who does not use any platform. According to their awareness, 45% of the participants claimed to be exposed to AI-generated images daily, 32% weekly, 9% monthly, and 5% yearly. The remaining 9% claimed to have never seen an AI-generated image.

Finally, the participants' self-evaluation regarding topics such as AI, image generative models, annotation tools, HCI, image editing tools, art, and photography presented median values of 3, 2, 2, 4, 3, 3, 3, respectively.

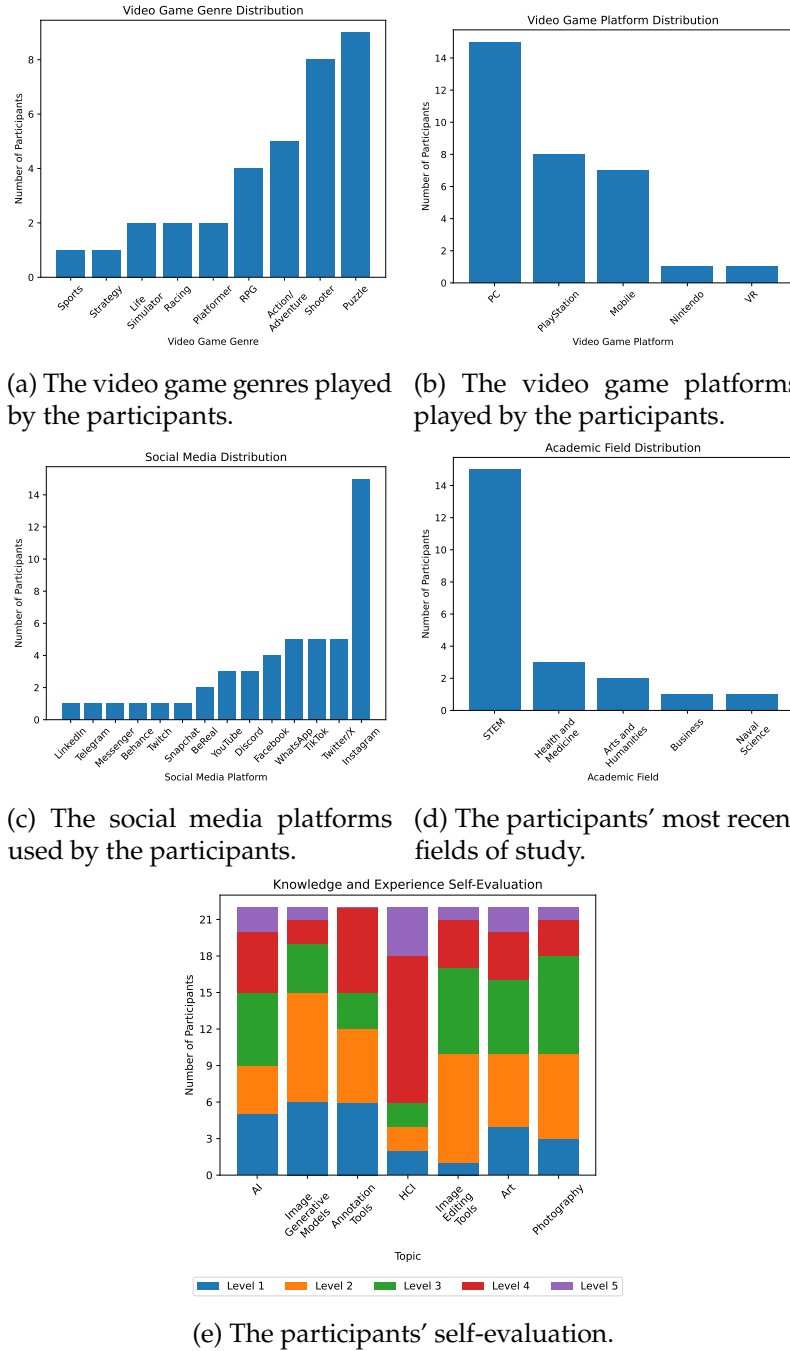


Figure 6.1: Formative study's population demographics.

6.1.3 Questionnaire Data Analysis

According to the collected data, presented in Table 6.3, the participants enjoyed *Adversarial Imagery's* different game modes, especially both versions of *Real or Fake* (G1 and TimedG1) and *Real vs Fake* (G2 and TimedG2), with each receiving a median value of 6. Despite presenting positive median values, the participants considered both versions of *Real and Fake* (G3 and TimedG3) to be less enjoyable, with G3 receiving a median value of 5.5 and TimedG3 a value of 4.5. The lower scores might be due to their structure and increased

difficulty, as the participants perceived them to be the most challenging, resulting in median values of 6 for both versions.

Table 6.3: The enjoyment level of *Adversarial Imagery's* game modes.

Game Mode ID	Median	Q1	Q3	Mode
G1	6	6	6.75	6
G2	6	6	6	6
G3	5.5	4.25	6	6
TimedG1	6	5	6.75	6
TimedG2	6	5	6	6
TimedG3	4.5	3	6	7

Both versions of *Real and Fake* (G3 and TimedG3) presented a median difficulty value of 6, as shown in Table 6.4. Additionally, the participants considered the timed game modes to be more difficult than their normal counterparts. Unlike our predictions, the participants perceived *Real or Fake* (G1 and TimedG1) to be less challenging than *Real vs Fake* (G2 and TimedG2), as G1 and TimedG1 received median values of 3 and 4, respectively, while G2 and TimedG2 received median values of 4 and 4.5.

Table 6.4: The difficulty level of *Adversarial Imagery's* game modes.

Game Mode ID	Median	Q1	Q3	Mode
G1	3	3	5.75	3
G2	4	3	5	4
G3	6	5	6	6
TimedG1	4	3	5.75	3
TimedG2	4.5	3.25	6	6
TimedG3	6	5	7	7

Table 6.5 reveals that the participants considered the game's entertainment value to be the most motivating factor to play more sessions, resulting in a median value of 6. In addition, the leaderboards and achievements also earned promising results, with both receiving median values of 5. Considering their motivation to improve their performance, the participants claimed the game's evaluation system, such as grades, scores, and leaderboard ranks, to be effective, resulting in a median value of 6. Also, although many participants did not achieve a perfect game and were unaware of the game's concept of perfect streaks, they envisioned it to be similarly effective.

Table 6.5: The impact of *Adversarial Imagery's* mechanics.

Game Mechanic Impact	Median	Q1	Q3	Mode
Impact of Fun on Play Time	6	5	6	6
Impact of Leaderboards on Play Time	5	4	6	5
Impact of Achievements on Play Time	5	4	6	5
Impact of Score/Grade/Rank on Performance	6	5	7	7
Impact of Max Perfect Streak on Performance	6	4.25	7	7

As Table 6.6 suggests, we asked the participants to rate how easy it is to distinguish between real and AI-generated images before and after their first game sessions. The first ratings presented a median value of 5, while the second presented a median value of 3. We also conducted a Wilcoxon signed-rank test, which revealed a significant difference between the two results ($p = 0.02$). Furthermore, the participants considered the AI-generated images to be realistic, earning a median value of 6. As such, the data implies a dissonance between the participants' expectations and the capabilities of current generative models, which can generate realistic images without resorting to prompt engineering.

Table 6.6: The ease of distinguishing between real images and AI-generated ones.

Time Instance	Median	Q1	Q3	Mode
Before Playthrough	5	4	5	5
After Playthrough	3	2	4.75	2

According to 73% of the participants, the most significant clue that helped them identify AI-generated images was the presence of human beings, as they often presented deformations, especially on their hands and faces. Furthermore, 32% searched for distorted text, while the same number of participants claimed to look for inconsistencies in lights, shadows, and reflections. Finally, 23% of the participants searched for distortions in animals and 18% claimed that images that looked too perfect were probably synthetic.

Regarding usability, *Adversarial Imagery* achieved a SUS score of 85.23 with a standard deviation of 9.13. To put the results into perspective, according to Bangor, Kortum, and Miller [5], the rating of "Excellent", the second highest rating in the presented scale with seven levels, corresponds to a mean SUS score value of 85.58 with a standard error of 9.473.

Finally, considering a potential future version of *Adversarial Imagery*, the participants shared some ideas for new game modes. For example, they suggested endless game modes, where players could keep guessing until they got one wrong or until time ran out, and a competitive version of the timed modes that would only consider the fastest guesses. The participants also suggested other features, such as adjustable timers for the timed game modes, depending on the players' accuracy, unlockable color themes for the menus, and a gallery containing the classified images as collectibles.

6.1.4 Gameplay Data Analysis

The data obtained from the participants' first game sessions of *Adversarial Imagery* does not entirely match their perception of difficulty, as illustrated in Figures 6.2 and 6.3. Despite being considered the easiest game mode, G1 presented a player accuracy of 70%, while G2 presented a higher value of 82%. Additionally, although both versions of *Real and Fake* (G3 and TimedG3) presented the same difficulty median values, G3 presented an accuracy value of 67%, the third highest, while TimedG3 presented the lowest accuracy of any game mode, 41%.

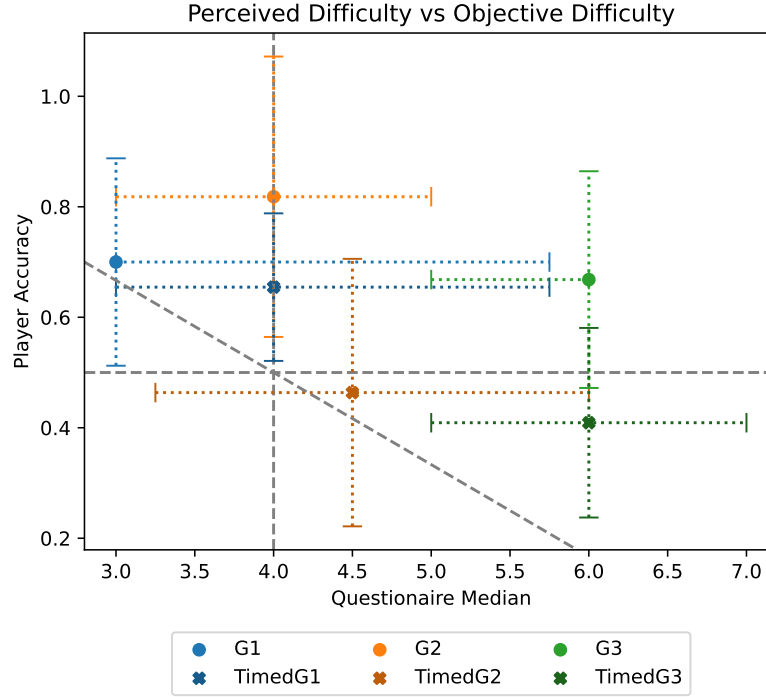


Figure 6.2: The correlation between the participants' accuracy and the perceived difficulty ratings assigned by them for each game mode. The vertical dotted lines represent the standard deviation of the computed accuracy values, while the horizontal dotted lines represent the Q1 and Q3 values of the perceived difficulty ratings.

Table 6.7: The player accuracy calculated from the participants' first and second game sessions of the normal game modes.

Game Mode ID	First Game Session	Second Game Session
G1	0.70	0.82
G2	0.82	0.92
G3	0.67	0.85

On their first attempts at the game's normal modes, the majority of the participants were unaware they could zoom in on the displayed images, as most did not pay attention to *Adversarial Imagery's* initial instructions. Their unawareness negatively affected their performance, especially in G3, considering the lower image size, which might have also influenced their opinions on the game's enjoyment and difficulty. As such, after informing them of that game mechanic, their accuracy improved on their second attempts, as evidenced in Table 6.7. It is worth noting that most participants avoided assigning empty labels to the images in G3, even in doubt, as there were only eight images with empty labels in the first attempt and three in the second attempt, considering that there were 220 classified images for each attempt.

In addition, two paired t-tests confirmed a significant difference between the resulting accuracies of the first and second attempts of G1 ($p = 0.02$) and G3 ($p = 0.00$). However, according to another t-test, G2's attempts did not present a significant difference ($p = 0.06$),

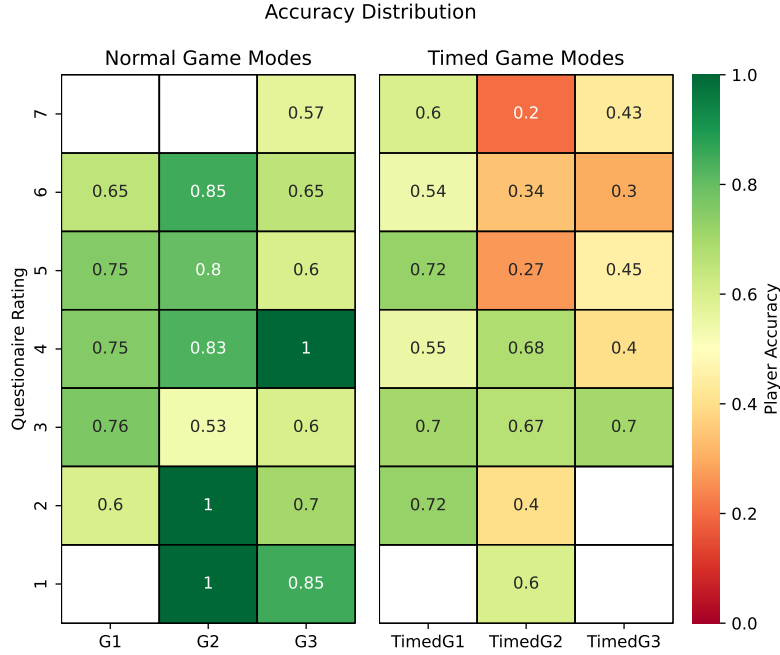


Figure 6.3: The accuracy distribution of the participants grouped by their perceived difficulty ratings for each game mode.

as the ability to compare the artificial images to real ones might be a sufficient advantage to aid the participants in successfully classifying the game’s images. The presented results motivated us to rethink our approach regarding *Adversarial Imagery’s* tutorials and revealed the impact of seemingly unimportant features on the data collected by a gamified data collection system. Furthermore, it also supported our decision to exclude participants with low English comprehension levels from the summative study.

As illustrated by Figure 6.4, accuracy improved when we increased the timed game modes’ round times from short to intermediate. However, the data does not show improvement with the longest round times, except for TimedG2, possibly because it provides less analysis time per image than the remaining game modes. The data implies the existence of a threshold, as the increase of time tends to improve accuracy until a certain point but begins to decrease and stabilize afterward. As the available time increased, the participants were less likely to run out of it, lowering the frequency of assigned empty labels. However, the extra time also provided them more opportunities to commit to their guesses, which were not always correct, increasing the frequency of incorrectly classified images and surpassing the benefits of the increased time limit.

We conducted three Tukey HSD tests to assess if the computed accuracies of each game mode were significantly different between the three round time variations. The time increase from short to intermediate resulted in a significant difference between the accuracy results for TimedG1 ($p = 0.04$) and TimedG2 ($p = 0.00$). However, there was no significant difference for TimedG3 ($p = 0.37$), as the extra time might not have been enough to significantly surpass the challenge caused by the small image size and inability

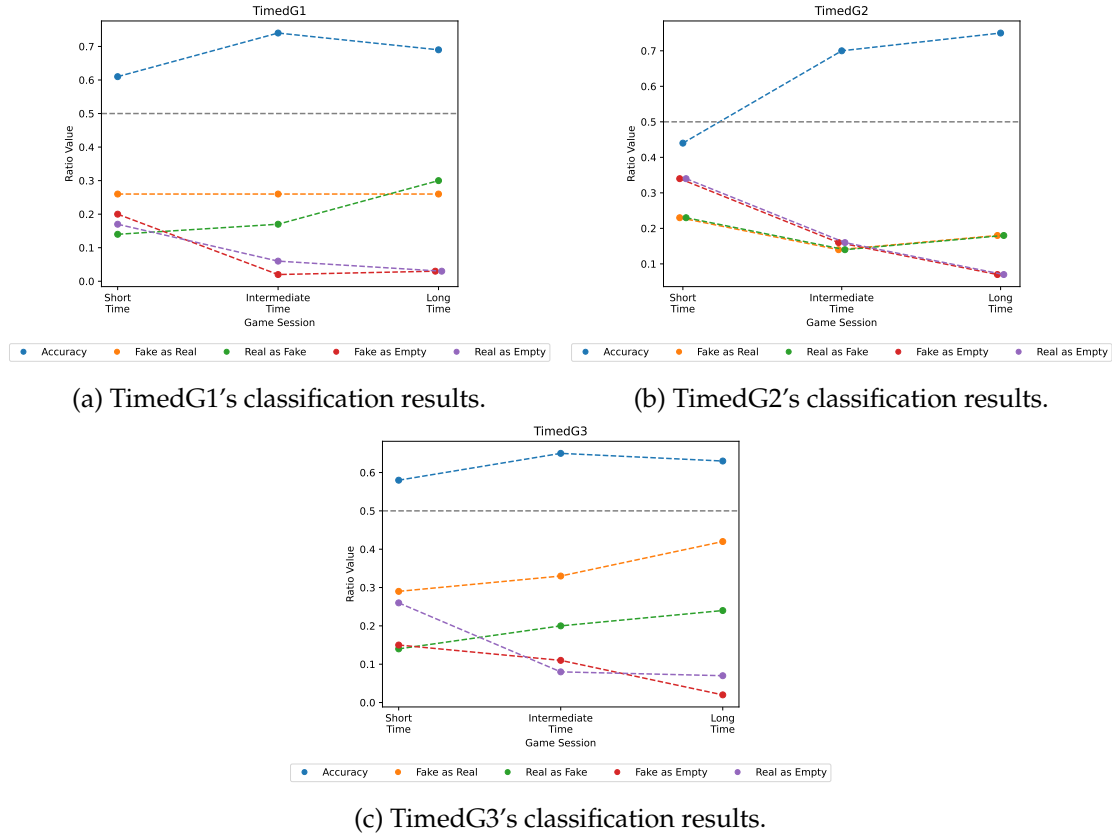


Figure 6.4: The classification accuracy and the rates of images classified with incorrect guesses during gameplay of the timed game modes.

to zoom in. Regarding the time increase from intermediate to long, there was no significant difference for TimedG1 ($p = 0.62$), TimedG2 ($p = 0.74$), and TimedG3 ($p = 0.96$). Lastly, the difference between the results obtained from the short and long sessions was not significant for TimedG1 ($p = 0.28$) and TimedG3 ($p = 0.52$), but it was for TimedG2 ($p = 0.00$).

It is worth noting that some participants commented that the timed game modes became easier the more they played, possibly due to the extra practice and exposition to recently classified images. Additionally, for each timed game mode, a linear regression confirmed a decrease of 0.03 points in the rates of assigned empty labels for each of the four attempts. Consequently, our observations supported the decision to prioritize unseen images to prevent any potential influence on the data. For the same reason, we also decided to increase the number of images for our summative study. Finally, one participant remembered the possibility of using the browser's developer tools, validating our initial concerns and implemented strategy.

6.2 Summative User Study

After analyzing the impact of *Adversarial Imagery*'s design on the collected data and implementing the necessary changes, we conducted a summative user study ($N = 129$) to benchmark the three selected generative models and assess their strengths and weaknesses. The new *Adversarial Imagery* version contains four hundred image pairs for each of the three datasets.

6.2.1 Protocol and Population

After its public distribution, *Adversarial Imagery* collected data for five weeks. We relied on social media and word-of-mouth to convince a substantial amount of players to participate in our study. As a result, *Adversarial Imagery* registered 129 player accounts. However, three players did not consent to the game's terms, while nine others did not meet the selection criteria, lowering the number of considered participants to 117.

As presented in Table 6.8, the considered participants spent 7 hours and 41 minutes playing the different game modes, resulting in the classification of 7,760 images, where 5,970 of the attributed labels were correct.

Table 6.8: General game statistics obtained from *Adversarial Imagery*.

Game Statistic	<i>Realistic Vision</i>	<i>DreamShaper</i>	<i>ReV Animated</i>	Total
Total Players	39	40	38	117
Total Time (h.)	02:10	02:51	02:40	07:41
Total Games	237	244	295	776
Total Guesses	2,370	2,440	2,950	7,760
Total Correct Guesses	1,667	1,807	2,496	5,970
Correct Guess Ratio	0.70	0.74	0.85	-

Finally, Figure 6.5 displays the distribution of *Adversarial Imagery*'s labeled images according to COCO's twelve supercategories. Similarly, Figure 6.6 displays the distribution of the labeled images that contain and do not contain the "Person" supercategory. It is worth mentioning that an image can present several elements of different supercategories. Furthermore, we assigned the same supercategories to the AI-generated images as their source images.

6.2.2 Results

Table 6.9 reveals that most participants of the summative study only played each game mode once, aside from some exceptions. Furthermore, not every player attempted the timed game modes after completing the normal ones. It is possible that they had already satisfied their curiosity, not feeling the need to play optional altered versions of the game modes they had recently completed. In addition, some players might have quit the game immediately after completing its main difficulty mode without realizing they had unlocked a new one. Finally, due to the lack of monitoring, it is difficult to conclude if the

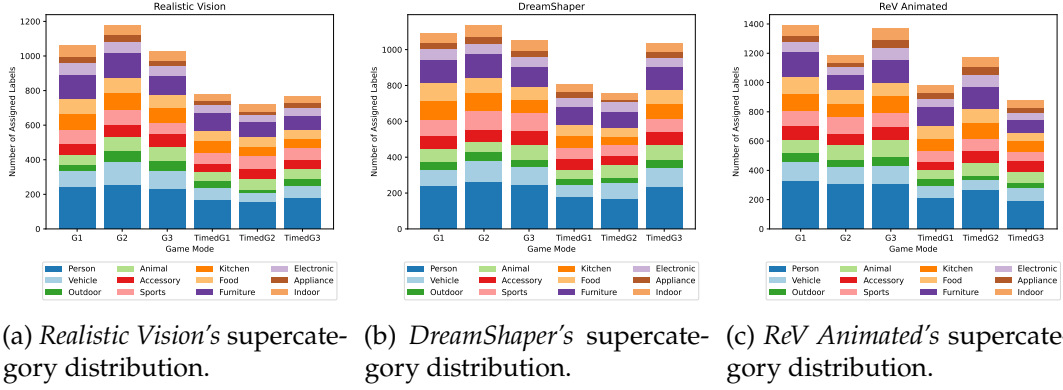


Figure 6.5: The distribution of *Adversarial Imagery*'s classified images according to COCO's twelve supercategories.

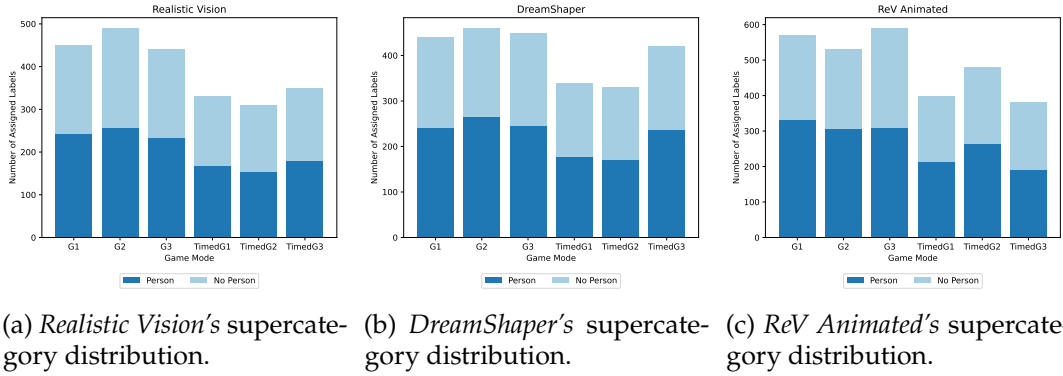


Figure 6.6: The distribution of *Adversarial Imagery*'s classified images that contain and do not contain the "Person" supercategory.

leaderboards and achievements were effective motivators despite the existence of players who played every game mode several times.

The data collected by *Adversarial Imagery* during the summative study supports our initial ranking predictions, as illustrated by Figure 6.7. The players mistook *Realistic Vision*'s AI-generated images as real more frequently than the images produced by the other models across every game mode. Furthermore, *Dreamshaper*'s multi-purpose model surpassed *ReV Animated*'s art-focused model.

Aligning with the discussed strategies of the participants of the formative study, according to the data illustrated in Figure 6.8, artificial images that contain people are generally easier to classify. Additionally, the previously discussed ranking remains consistent throughout every game mode, regardless of whether we only consider images that contain human beings or those without them. However, exceptionally, *DreamShaper* performed slightly better than *Realistic Vision* in TimedG3 when only considering images that did not contain the "Person" supercategory.

The data also reveals other details, such as the fact that *Realistic Vision* was capable of producing artificial images with humans more frequently convincing than *ReV Animated*'s images that did not contain them across every game mode, except for G1. Furthermore,

Table 6.9: The number of sessions played by each player of *Adversarial Imagery* grouped by their assigned databases.(a) *Realistic Vision's* game sessions.

Game Mode ID	Total	Median	Q1	Q3	Min	Max	Mode
G1	45	1	1	1	0	7	1
G2	49	1	1	1	0	11	1
G3	44	1	1	1	0	9	1
TimedG1	33	1	0	1	0	8	1
TimedG2	31	1	0	1	0	7	1
TimedG3	35	1	0	1	0	10	1

(b) *DreamShaper's* game sessions.

Game Mode ID	Total	Median	Q1	Q3	Min	Max	Mode
G1	44	1	1	1	0	3	1
G2	46	1	1	1	0	4	1
G3	45	1	1	1	0	4	1
TimedG1	34	0.5	0	1	0	5	0
TimedG2	33	1	0	1	0	4	0
TimedG3	42	0.5	0	1	0	5	0

(c) *ReV Animated's* game sessions.

Game Mode ID	Total	Median	Q1	Q3	Min	Max	Mode
G1	57	1	1	1	0	10	1
G2	53	1	1	1	0	10	1
G3	59	1	1	1	0	12	1
TimedG1	40	1	0	1	0	10	1
TimedG2	48	1	0	1	0	11	1
TimedG3	38	1	0	1	0	10	1

DreamShaper's artificial images that contain people present a higher fool rate than those without them produced by *ReV Animated* in both versions of *Real vs Fake* (G2 and TimedG2). Lastly, both of *DreamShaper's* artificial image sets present the same fool rate in TimedG2.

As illustrated by Figure 6.9, the model ranking generally follows the same placement across COCO's different supercategories with some exceptions. For example, in G1, *DreamShaper* surpasses *Realistic Vision* in supercategories such as "Indoor", "Electronic", and "Sports". Meanwhile, *ReV Animated* outperforms *DreamShaper* in supercategories such as "Appliance", "Kitchen", and "Animal".

It is worth mentioning that it is possible to use *Adversarial Imagery* to select the generative models' most realistic images according to the guesses of the players. For example, Figure 6.10 showcases the label distribution of the classified images in G1 from the *Realistic Vision* dataset, along with three AI-generated images frequently classified as real.

We stored the number of labels assigned to each image in a three-entry array, where each entry corresponds to the three possible classifications: real, fake, and empty. We represented the number of guesses assigned to the real images as positive integers and used negative integers to represent the labels of the AI-generated images, functioning

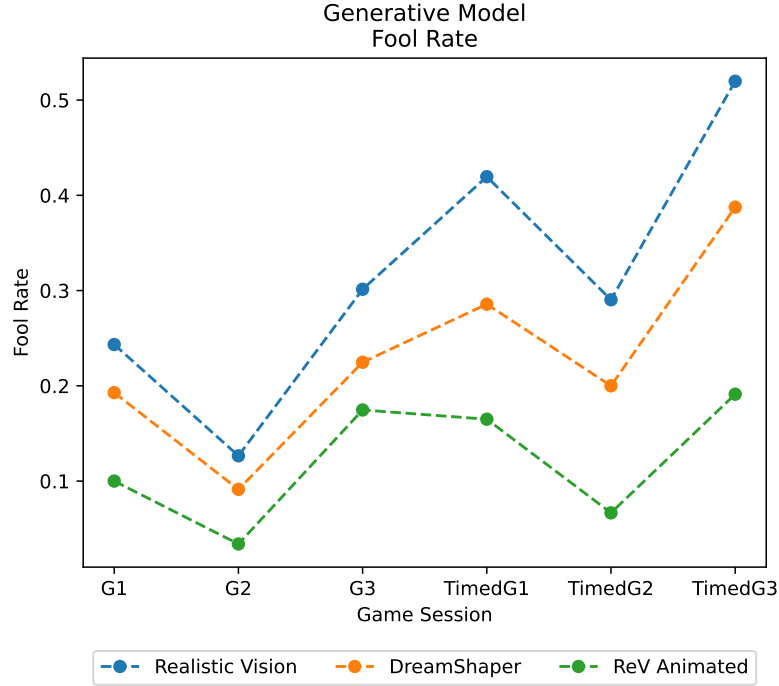


Figure 6.7: The rates of AI-generated images classified as real.

as scores. The plot displays the scores of the images normalized between -1.0 and 1.0. Additionally, it represents real classifications as green dots and fake classifications as red dots.

As such, the plot contains every real image in its green region, with scores above zero, mostly assigned with real labels, as seen by the larger number of green dots. On the other hand, the plot contains the AI-generated images in its red region, with scores below zero, mostly assigned with fake labels, as seen by the larger number of red dots. This visualization allowed us to select some of the most realistic artificial images according to the players by analyzing the green dots in the red region with the highest absolute score values. Finally, the density of green dots in the red region can also function as a visual metric, showcasing the achievable realism of the selected model and the frequency of players classifying AI-generated images as real.

6.3 Discussion

The data collected during the formative and summative studies allowed us to answer our research questions.

- Q1: Is it possible to use games as a benchmarking strategy for deep generative models of images?

The results of the previously discussed studies demonstrate the effectiveness and potential of our methodology, as it achieved our initial goals. It succeeded in collecting

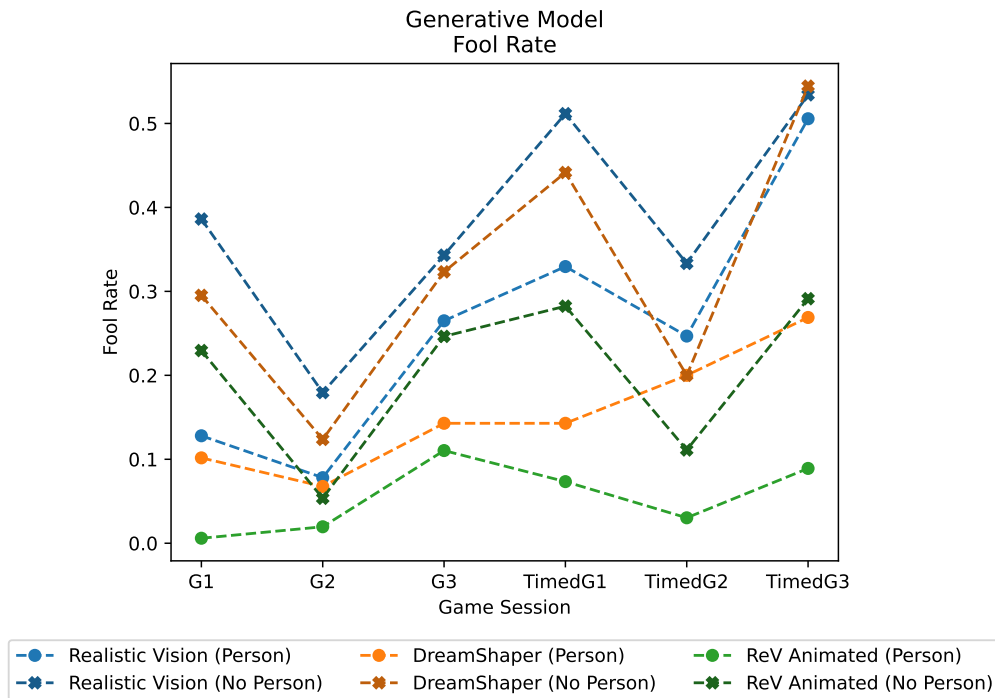


Figure 6.8: The rates of AI-generated images that contain and do not contain the "Person" supercategory classified as real.

a substantial amount of data from several players, resulting in the classification of 7,760 images by 117 players, which enabled us to benchmark the three selected generative models, coinciding with our ranking predictions. Furthermore, it allowed us to assess the models' strengths and weaknesses according to the elements present in the generated images, such as people and objects, under different contextual scenarios.

- Q2: What are deep generative models of images capable of, and what are their limits?

Despite not relying on prompt engineering, the three chosen models were capable of generating realistic images convincing enough to fool the players of *Adversarial Imagery*. However, each model performed differently, as they were suitable for different tasks. Although the participants of the formative study considered the artificial images to be realistic, they frequently reported issues regarding human beings' anatomy. In addition, they also found distortions and deformations in animals and letters, considered some lights, shadows, and reflections to be inconsistent, and were suspicious of images that looked too perfect. Lastly, the summative study revealed how often the participants labeled AI-generated images with certain elements as real, revealing that the artificial images that did not contain human beings were more convincing than those with them.

- Q3: How easy is it to fool a human being with AI-generated images?

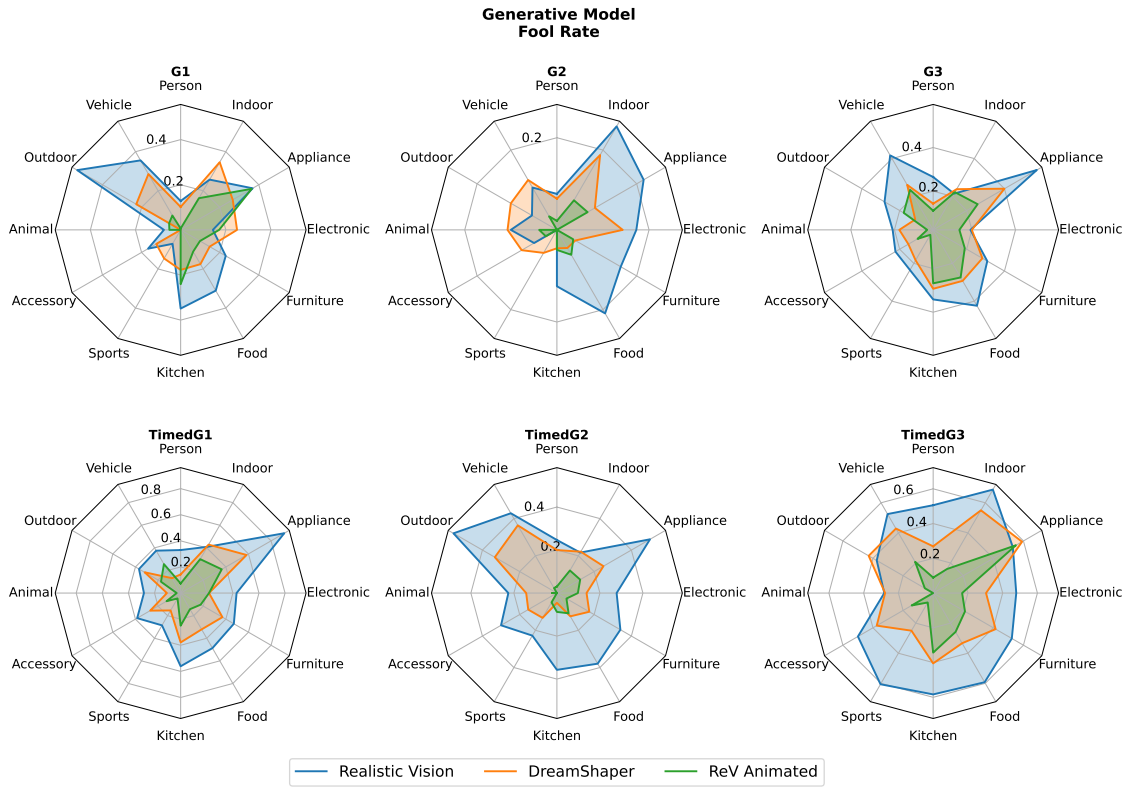


Figure 6.9: The rates of AI-generated images classified as real according to COCO’s twelve supercategories.

As discussed in the previous answer, distortions, especially those present in human beings, decrease the likelihood of people classifying AI-generated images as real, depending on their subtlety. The fool rate depends on the chosen generative model and the context in which the images are present, as supported by the different results obtained from the different game modes of *Adversarial Imagery*. It also depends on the image size, as evidenced by the difference between the game sessions where the participants did and did not zoom in on them.

Additionally, the data obtained from the timed game modes suggest that the success rate of the generative models depends on the classifiers’ attention and motivation. As such, AI-generated images might convince people more frequently on social media platforms, where users can frequently find them, especially considering the low amount of time spent on each image. It is also worth noting that the participants of the formative study considered distinguishing between real and AI-generated images to be more difficult than they anticipated. Consequently, it implies that people’s perception of AI-generated images might not match the capabilities of current technology.

- Q4: What are the challenges of designing a game meant to collect valuable and trustworthy data for research?

One of the biggest challenges of designing a game such as *Adversarial Imagery* is to

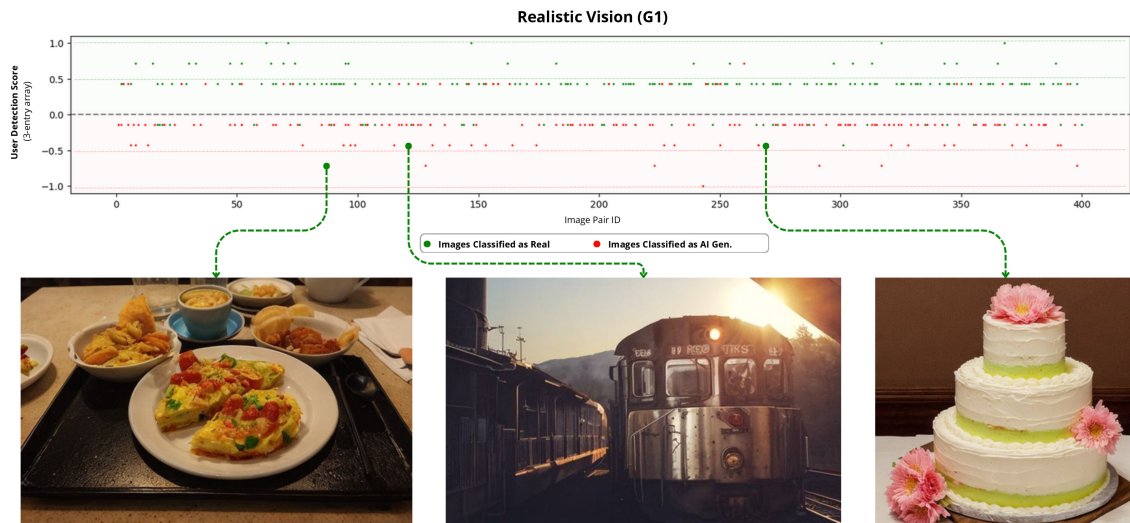


Figure 6.10: The normalized image scores of *Realistic Vision*'s dataset according to the assigned labels obtained from G1. The green region contains real images, and the red region contains AI-generated images. The green dots represent real classifications, and the red dots represent fake classifications. The most realistic images correspond to the green dots in the red region with the highest absolute score value, as they represent the AI-generated images more frequently classified as real than the rest, exemplified by the images with the following image pair IDs: 87, 121, and 269.

achieve a balance between a fun game, to ensure several players keep playing the game, and a reliable data collection system, to ensure that the collected data is valuable and trustworthy. It is crucial to carefully analyze every design decision, as some features designed to contribute towards one of the system's two goals might hurt the other, as exemplified by the presence of player accounts.

Furthermore, the design of certain features, which in a regular video game would seem unimportant, such as the game's tutorials, can have a massive impact on the collected data. Due to the inability to monitor and guide the players, the game's design must compensate for that to avoid cheating and other adversarial behavior that could affect the collected data. Finally, it is difficult to predict the game's total number of players and the amount of time they are willing to play it, which can affect the amount of collected data and certain decisions, as demonstrated by the number of available images.

- Q5: How can the design of a gamified data collection system influence its collected data?

As previously discussed with *Adversarial Imagery*'s player accounts and tutorials, the design of gamification elements can impact the amount and reliability of the collected data. Although the different game modes presented the same objective of classifying real and AI-generated images, their designs originated different results. Features such as the ability to zoom in significantly affected the collected data, especially considering

the different default image sizes for each game mode. As a result, it is essential to ensure players understand what to do and what they can do.

It is also important to motivate the players to take their tasks seriously, whether by classifying their performance with a scoring system or giving them a sense of progression through achievements, considering the positive ratings attributed by the participants of the formative study. Finally, it is worth noting that limiting the amount of time and preventing players from zooming in on the images reduced their accuracy. However, increasing the available time did not always improve their performance, especially when the images were too small to analyze.

6.4 Summary

After implementing the first version of *Adversarial Imagery*, we conducted a formative study to evaluate the impact of the design of the different game modes and features, such as leaderboards and achievements, on the collected data. We also examined the visual clues that aided the participants in classifying the game's images. Afterward, according to our observations and feedback, we updated the game. Then, we made it publicly available as a means to convince several players to participate in our summative study and gather a substantial amount of data. With the collected data, we ranked *Realistic Vision* as the best model, followed by *DreamShaper*, and then *ReV Animated*. Finally, we assessed the capabilities of the selected models, analyzed the success of our methodology, and answered the research questions proposed by this thesis.

CONCLUSION

Deep generative models are tools capable of producing high-quality images whose realism has the potential to fool people into thinking they are human-made. The abundance of generative models offers a considerable amount of choices to potential users. However, finding the most appropriate model for the task at hand can be challenging, especially when inexperienced users lack the knowledge to harness the models' full potential.

As such, aiming to find a solution for the need for benchmarking, in Chapter 1, we established our goals and motivations. In Chapter 2, we analyzed the implementation of many deep generative models and their applications. Furthermore, we analyzed the concept of the Turing test and the design of certain video games whose main objective was to draw in as many players as possible to complete certain tasks and collect a substantial amount of reliable data.

With the acquired knowledge, in Chapter 3, we proposed a gamified benchmarking approach to evaluate the realism of AI-generated images produced by three models. We also presented an overview of our methodology's architecture, which included two main components: an image generation pipeline and a game named *Adversarial Imagery*.

In Chapter 4, we discussed the implementation of the image generation pipeline and analyzed the three chosen models and the image dataset used to condition the generation of the AI-generated images. In Chapter 5, we analyzed the design of *Adversarial Imagery*, including its game modes inspired by the Turing tests, where we tasked players to identify the real and AI-generated images. We also discussed the importance of certain features, such as the game's leaderboards and achievements, designed to motivate players to keep playing and take their tasks seriously.

Finally, in Chapter 6, we presented the results of two studies, which enabled us to analyze the impact of *Adversarial Imagery*'s design on the collected data, benchmark the three selected generative models and assess the success of our methodology.

Adversarial Imagery revealed great potential as a gamified benchmarking approach for generative models of images, as it collected a large quantity of valuable data, accomplishing our objectives. The game successfully gathered 117 players, resulting in the collection of 7,760 image labels, which allowed us to rank three generative models according to the

realism of their generated images. We were also able to assess the capabilities of current technology and how often AI-generated images could fool human beings under different circumstances and highlighted the challenges of generating certain elements, such as realistic human beings. Additionally, we demonstrated the challenges of designing a gamified data collection system, especially considering the impact of the design of certain features, such as the game’s tutorials and the presence of time constraints, on the collected data.

7.1 Future Work

It could be valuable to update *Adversarial Imagery* considering the ideas suggested during the formative study and the discussed unimplemented game modes. For example, the introduction of multiplayer game modes and the subsequent analysis regarding their effect on the collected data, player count, and player retention compared to the existing single-player game modes could benefit projects that follow a similar concept.

Furthermore, we consider that *Adversarial Imagery* serves as a proof of concept. As such, a similar methodology might be useful to evaluate the realism of other AI-generated content, such as audio or video, and other studies that require massive amounts of data.

It could also be worthwhile to investigate if the knowledge that certain content was AI-generated affects people’s preferences, depending on the context of its usage and the demographics of the questioned participants. Since the topic can invoke mixed reactions, monitoring the evolution of society’s feelings and opinions might be important to progress.

BIBLIOGRAPHY

- [1] L. Ahn. “Games with a Purpose”. In: *Computer* 39 (2006-07), pp. 92–94. DOI: [10.1109/MC.2006.196](https://doi.org/10.1109/MC.2006.196) (cit. on p. 27).
- [2] L. von Ahn and L. Dabbish. *Designing games with a purpose*. 2008-08. DOI: [10.1145/1378704.1378719](https://doi.org/10.1145/1378704.1378719) (cit. on pp. 2, 23, 24, 26).
- [3] L. V. Ahn and L. Dabbish. *Labeling Images with a Computer Game*. 2004 (cit. on p. 24).
- [4] L. V. Ahn, M. Kedia, and M. Blum. *Verbosity: A Game for Collecting Common-Sense Facts*. 2006 (cit. on p. 25).
- [5] A. Bangor, P. T. Kortum, and J. T. Miller. “An Empirical Evaluation of the System Usability Scale”. In: *International Journal of Human–Computer Interaction* 24.6 (2008), pp. 574–594. DOI: [10.1080/10447310802205776](https://doi.org/10.1080/10447310802205776) (cit. on p. 68).
- [6] D. Bank, N. Koenigstein, and R. Giryes. “Autoencoders”. In: *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*. Ed. by L. Rokach, O. Maimon, and E. Shmueli. Cham: Springer International Publishing, 2023, pp. 353–354. ISBN: 978-3-031-24628-9. DOI: [10.1007/978-3-031-24628-9_16](https://doi.org/10.1007/978-3-031-24628-9_16) (cit. on p. 11).
- [7] C. Bassett. “The computational therapeutic: exploring Weizenbaum’s ELIZA as a history of the present”. In: *AI and Society* 34 (4 2019-12), pp. 803–812. ISSN: 14355655. DOI: [10.1007/s00146-018-0825-9](https://doi.org/10.1007/s00146-018-0825-9) (cit. on pp. 19, 20).
- [8] Y. Bitton, N. Guetta, R. Yosef, Y. Elovici, M. Bansal, G. Stanovsky, and R. Schwartz. *WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models*. 2022-07. DOI: [10.48550/arXiv.2207.12576](https://doi.org/10.48550/arXiv.2207.12576) (cit. on p. 47).
- [9] L. Blair. “The Use Of Video Game Achievements To Enhance Player Performance, Self-efficacy, And Motivation”. In: (2011) (cit. on p. 52).
- [10] B. J. Copeland. “The Turing Test*”. In: *Minds and Machines* 10.4 (2000), pp. 519–539 (cit. on pp. 6, 17, 20).
- [11] R. B. Dannenberg, E. L. M. Law, L. V. Ahn, and M. Crawford. *TagATune: A Game for Music and Sound Annotation*. 2007 (cit. on p. 26).

-
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: [10.1109/CVPR.2009.5206848](https://doi.org/10.1109/CVPR.2009.5206848) (cit. on p. 15).
 - [13] P. Dhariwal and A. Nichol. “Diffusion Models Beat GANs on Image Synthesis”. In: *Advances in neural information processing systems* 34 (2021), pp. 8780–8794 (cit. on p. 10).
 - [14] J. Feigenbaum, D. P. Gross, D. Acemoglu, D. Autor, E. Glaeser, S. Greenstein, B. Kerr, F. Levy, B. Margo, B. Beach, W. Hanlon, and P. Restrepo. *Automation and the Fate of Young Workers: Evidence from Telephone Operation in the Early 20th Century*. 2020 (cit. on p. 2).
 - [15] K. Finstad. “Response Interpolation and Scale Sensitivity: Evidence Against 5-Point Scales”. In: *J. Usability Studies* 5.3 (2010-05), pp. 104–110 (cit. on p. 64).
 - [16] F. García-Peñalvo and A. Vázquez-Ingelmo. “What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI”. In: *International Journal of Interactive Multimedia and Artificial Intelligence* In Press (In Press 2023), p. 1. DOI: [10.9781/ijimai.2023.07.006](https://doi.org/10.9781/ijimai.2023.07.006) (cit. on pp. 1, 9).
 - [17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. *Generative Adversarial Nets*. 2014. arXiv: [1406.2661](https://arxiv.org/abs/1406.2661) [stat.ML] (cit. on pp. 3, 10).
 - [18] M. Haenlein and A. Kaplan. “A brief history of artificial intelligence: On the past, present, and future of artificial intelligence”. In: *California Management Review* 61 (4 2019-08), pp. 5–14. ISSN: 21628564. DOI: [10.1177/0008125619864925](https://doi.org/10.1177/0008125619864925) (cit. on p. 6).
 - [19] P. Hämäläinen, M. Tavast, and A. Kunnari. “Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3580688](https://doi.org/10.1145/3544548.3580688) (cit. on p. 56).
 - [20] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. 2018. arXiv: [1706.08500](https://arxiv.org/abs/1706.08500) [cs.LG] (cit. on p. 12).
 - [21] J. Ho, A. Jain, and P. Abbeel. “Denoising Diffusion Probabilistic Models”. In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851 (cit. on p. 11).
 - [22] S. Jain and D. C. Parkes. “A game-theoretic analysis of the ESP game”. In: vol. 1. Association for Computing Machinery, 2013-01. DOI: [10.1145/2399187.2399190](https://doi.org/10.1145/2399187.2399190) (cit. on p. 25).
 - [23] D. Jannai, A. Meron, B. Lenz, Y. Levine, and Y. Shoham. “Human or Not? A Gamified Approach to the Turing Test”. In: (2023-05) (cit. on pp. 20, 21).

- [24] R. Jesus, D. Goncalves, A. J. Abrantes, and N. Correia. “Playing Games as a Way to Improve Automatic Image Annotation”. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2008, pp. 1–8. DOI: [10.1109/CVPRW.2008.4563045](https://doi.org/10.1109/CVPRW.2008.4563045) (cit. on p. 47).
- [25] M. Kumar, D. Weissenborn, and N. Kalchbrenner. *Colorization Transformer*. 2021. arXiv: [2102.04432](https://arxiv.org/abs/2102.04432) [cs.CV] (cit. on p. 15).
- [26] R. N. Landers and A. K. Landers. “An Empirical Test of the Theory of Gamified Learning: The Effect of Leaderboards on Time-on-Task and Academic Performance”. In: *Simul. Gaming* 45.6 (2014-12), pp. 769–785. ISSN: 1046-8781. DOI: [10.1177/1046878114563662](https://doi.org/10.1177/1046878114563662) (cit. on p. 52).
- [27] E. Law and L. Ahn. “Input-agreement: A New Mechanism for Collecting Data Using Human Computation Games”. In: 2009-10, pp. 1197–1206. DOI: [10.1145/1518701.1518881](https://doi.org/10.1145/1518701.1518881) (cit. on p. 26).
- [28] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie. “Evaluation of algorithms using games: The case of music tagging.” In: *ISMIR*. Citeseer. 2009, p. 388 (cit. on p. 27).
- [29] Y. Lecun, Y. Bengio, and G. Hinton. *Deep learning*. 2015-05. DOI: [10.1038/nature14539](https://doi.org/10.1038/nature14539) (cit. on p. 7).
- [30] W. Li and C. Zhang. “Markov Chain Analysis”. In: *International Encyclopedia of Human Geography*. Ed. by R. Kitchin and N. Thrift. Oxford: Elsevier, 2009, pp. 455–460. ISBN: 978-0-08-044910-4. DOI: <https://doi.org/10.1016/B978-008044910-4.00470-3> (cit. on p. 11).
- [31] J. M. Lourenço. *The NOVAthesis L^AT_EX Template User’s Manual*. NOVA University Lisbon. 2021. URL: <https://github.com/joaomlourenco/novathesis/raw/main/template.pdf> (cit. on pp. i, iii).
- [32] M. Malheiros and S. Preibusch. “Sign-Up or Give-Up: Exploring User Drop-Out in Web Service Registration”. In: *Symposium on Usable Privacy and Security (SOUPS)*. 2013 (cit. on p. 47).
- [33] G. Martínez, L. Watson, P. Reviriego, J. A. Hernández, M. Juárez, and R. Sarkar. *Combining Generative Artificial Intelligence (AI) and the Internet: Heading towards Evolution or Degradation?* 2023. arXiv: [2303.01255](https://arxiv.org/abs/2303.01255) [cs.CV] (cit. on pp. 16, 18).
- [34] M. L. Mauldin. “Chatterbots, Tnymuds, and the Turing Test Entering the Loebner Prize Competition”. In: *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*. AAAI’94. Seattle, Washington: AAAI Press, 1994, pp. 16–21 (cit. on p. 19).
- [35] S. O. Mohammadi and A. Kalhor. *Smart Fashion: A Review of AI Applications in the Fashion & Apparel Industry*. 2021. arXiv: [2111.00905](https://arxiv.org/abs/2111.00905) [cs.CV] (cit. on p. 45).

-
- [36] A. Newell, J. C. Shaw, and H. A. Simon. "Report on a general problem solving program". In: *IFIP congress*. Vol. 256. Pittsburgh, PA. 1959, p. 64 (cit. on p. 7).
- [37] M. A. Nielsen. *Neural networks and deep learning*. Vol. 25. Determination press San Francisco, CA, USA, 2015. Chap. 1 (cit. on pp. 7, 8).
- [38] A. M. Noll. "Computers and the visual arts". In: *Design Quarterly* 66/67 (1966), p. 70 (cit. on p. 23).
- [39] A. M. Noll. *Human or Machine: A Subjective Comparison of Piet Mondrian's "Composition with Lines" (1917) and a Computer-Generated Picture*. 1966 (cit. on p. 22).
- [40] A. M. Noll. "The beginnings of computer art in the United States: A memoir". In: *Leonardo* (1994), pp. 39–44 (cit. on p. 23).
- [41] M. Pandelaere, B. Briers, and C. Lembregts. "How to Make a 29% Increase Look Bigger: The Unit Effect in Option Comparisons". In: *Journal of Consumer Research* 38.2 (2011), pp. 308–322. ISSN: 00935301, 15375277 (cit. on p. 52).
- [42] Y. Pang, J. Lin, T. Qin, and Z. Chen. *Image-to-Image Translation: Methods and Applications*. 2021. arXiv: [2101.08629 \[cs.CV\]](#) (cit. on p. 12).
- [43] J. Ploennigs and M. Berger. "AI art in architecture". In: *AI in Civil Engineering* 2.1 (2023-08). ISSN: 2730-5392. DOI: [10.1007/s43503-023-00018-y](#) (cit. on pp. 2, 14).
- [44] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: [2103.00020 \[cs.CV\]](#) (cit. on p. 12).
- [45] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al. "Improving language understanding by generative pre-training". In: (2018) (cit. on p. 8).
- [46] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: [2204.06125 \[cs.CV\]](#) (cit. on pp. 11–13).
- [47] 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*. 2022, pp. 10684–10695 (cit. on p. 11).
- [48] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*. 2023. arXiv: [2208.12242 \[cs.CV\]](#) (cit. on pp. 13, 14).
- [49] L. Ruthotto and E. Haber. *An Introduction to Deep Generative Modeling*. 2021. arXiv: [2103.05180 \[cs.LG\]](#) (cit. on p. 10).

- [50] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. “Palette: Image-to-Image Diffusion Models”. In: *ACM SIGGRAPH 2022 Conference Proceedings*. SIGGRAPH ’22. Vancouver, BC, Canada: Association for Computing Machinery, 2022. ISBN: 9781450393379. DOI: [10.1145/3528233.3530757](https://doi.org/10.1145/3528233.3530757) (cit. on pp. 2, 14, 16, 22).
- [51] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. arXiv: [2205.11487](https://arxiv.org/abs/2205.11487) [cs.CV] (cit. on p. 2).
- [52] H. Sasaki, C. G. Willcocks, and T. P. Breckon. “Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models”. In: *arXiv preprint arXiv:2104.05358* (2021) (cit. on p. 13).
- [53] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. “Mastering the game of Go with deep neural networks and tree search”. In: *Nature* 529 (7587 2016-01), pp. 484–489. ISSN: 14764687. DOI: [10.1038/nature16961](https://doi.org/10.1038/nature16961) (cit. on p. 7).
- [54] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. “Deep unsupervised learning using nonequilibrium thermodynamics”. In: *International conference on machine learning*. PMLR. 2015, pp. 2256–2265 (cit. on p. 11).
- [55] B. Trabucco, K. Doherty, M. Gurinas, and R. Salakhutdinov. *Effective Data Augmentation With Diffusion Models*. 2023. arXiv: [2302.07944](https://arxiv.org/abs/2302.07944) [cs.CV] (cit. on pp. 2, 15, 17).
- [56] A. M. Turing. “Computing Machinery and Intelligence”. In: *Mind* 59.236 (1950-10), pp. 433–460. ISSN: 0026-4423. DOI: [10.1093/mind/LIX.236.433](https://doi.org/10.1093/mind/LIX.236.433) (cit. on pp. 3, 6, 17).
- [57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017 (cit. on pp. 8–10).
- [58] K. Warwick and H. Shah. “Can machines think? A report on Turing test experiments at the Royal Society”. In: *Journal of Experimental and Theoretical Artificial Intelligence* 28 (6 2016-11), pp. 989–1007. ISSN: 13623079. DOI: [10.1080/0952813X.2015.1055826](https://doi.org/10.1080/0952813X.2015.1055826) (cit. on p. 19).

- [59] H. Yan, H. Zhang, L. Liu, D. Zhou, X. Xu, Z. Zhang, and S. Yan. “Toward intelligent design: An ai-based fashion designer using generative adversarial networks aided by sketch and rendering generators”. In: *IEEE Transactions on Multimedia* (2022) (cit. on pp. 2, 14).
- [60] D. Yi, C. Guo, and T. Bai. “Exploring painting synthesis with diffusion models”. In: Institute of Electrical and Electronics Engineers Inc., 2021-07, pp. 332–335. ISBN: 9781665433372. DOI: [10.1109/DTP152967.2021.9540115](https://doi.org/10.1109/DTP152967.2021.9540115) (cit. on pp. 22, 23).
- [61] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang. “Why Johnny Can’t Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts”. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI ’23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: [10.1145/3544548.3581388](https://doi.org/10.1145/3544548.3581388) (cit. on p. 35).
- [62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric*. 2018. arXiv: [1801.03924](https://arxiv.org/abs/1801.03924) [cs.CV] (cit. on p. 14).

FIGURE ANNEX



(a) Description and Highlights: "x The Beatles intarsia-knit coat; black/white, organic cotton, mix print, intarsia knit, drop shoulder, no fastening, open front, long sleeves, long length, When buying this unisex item, keep in mind that it is graded in standard men's sizing., POSITIVELY CONSCIOUS: We've partnered with Good On You 2014 an ethical agency that rates how sustainably brands perform across various people, planet and animal welfare criteria. This product comes from a Conscious brand that performs well in relation to their impact on animals."



(b) Description and Highlights: "mid-rise drop-crotch tapered jeans; light blue, cotton, tapered design, contrast stitching, belt loops, front button and zip fastening, two front welt pockets, drop crotch."

Figure I.1: Farfetch images and their associated product descriptions and highlights.

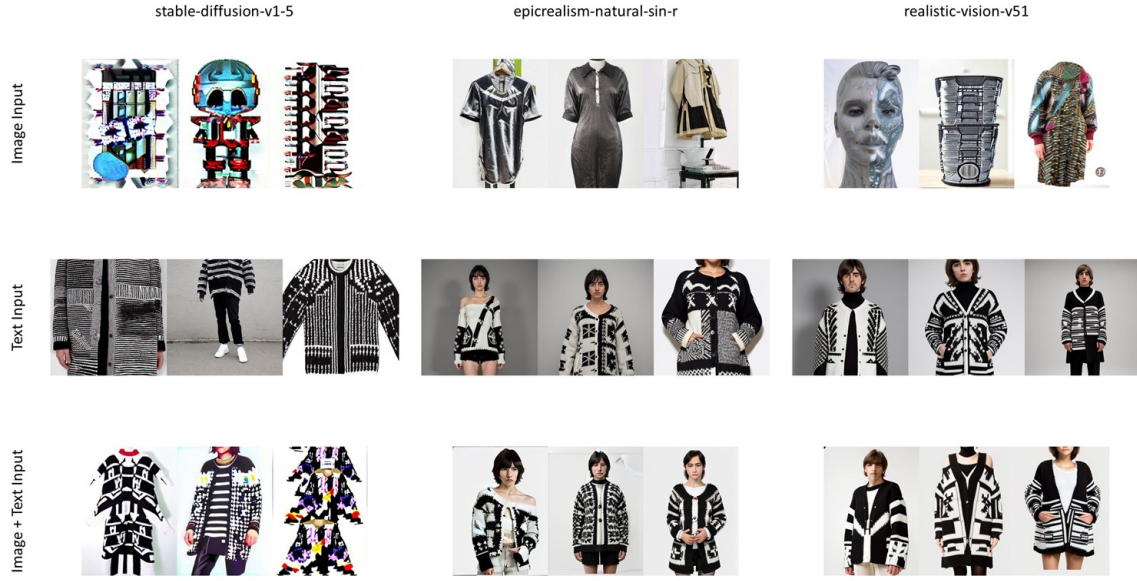


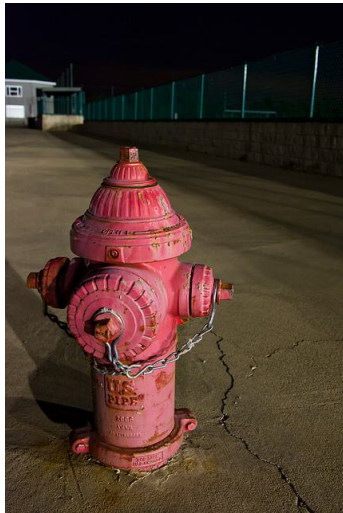
Figure I.2: Images generated based on Figure I.1a, by *Stable Diffusion* (left column), *epiCRealism* (center column), and *Realistic Vision* (right column). The first row corresponds to generation conditioned on a product's image, the second row corresponds to generation conditioned on the product's description and highlights, and the third row corresponds to generation conditioned on both (*strength* = 0.8, *num_inference_steps* = 50, *guidance_scale* = 7.5).



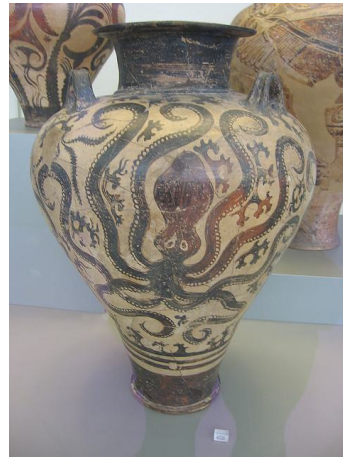
Figure I.3: Images generated based on Figure I.1b, by *Stable Diffusion* (left column), *epiCRealism* (center column), and *Realistic Vision* (right column). The first row corresponds to generation conditioned on a product's image, the second row corresponds to generation conditioned on the product's description and highlights, and the third row corresponds to generation conditioned on both (*strength* = 0.8, *num_inference_steps* = 50, *guidance_scale* = 7.5).



Figure I.4: Manually selected images generated by *Realistic Vision*, where each row is based on Figures I.1a and I.1b ($strength = 0.25$, $num_inference_steps = 100$, $guidance_scale = 7$).



(a) Caption: "A red fire hydrant with faded paint sits on the concrete."



(b) Caption: "An artistic vase with a depiction of an octopus on it."



(c) Caption: "A barbecue sandwich is sitting on a plate with some greens and some kind of yellow sauce."



(d) Caption: "A cow laying down in a grass field."

Figure I.5: COCO images and their chosen captions.



Figure I.6: Images generated based on Figure I.5a (first column) and its caption, by *Realistic Vision* (second column), *DreamShaper* (third column), and *ReV Animated* (fourth column) ($strength = 0.7$, $num_inference_steps = 150$, $guidance_scale = 3.5$).



Figure I.7: Images generated based on Figure I.5b (first column) and its caption, by *Realistic Vision* (second column), *DreamShaper* (third column), and *ReV Animated* (fourth column) ($strength = 0.7$, $num_inference_steps = 150$, $guidance_scale = 3.5$).



Figure I.8: Images generated based on Figure I.5c (first column) and its caption, by *Realistic Vision* (second column), *DreamShaper* (third column), and *ReV Animated* (fourth column) ($strength = 0.7$, $num_inference_steps = 150$, $guidance_scale = 3.5$).



Figure I.9: Images generated based on Figure I.5d (first column) and its caption, by *Realistic Vision* (second column), *DreamShaper* (third column), and *ReV Animated* (fourth column) (*strength* = 0.7, *num_inference_steps* = 150, *guidance_scale* = 3.5).



Figure I.10: An early sketch of the *Real or Fake* (G1) game mode, where players are tasked with classifying each image individually as Real or Fake (AI-generated).

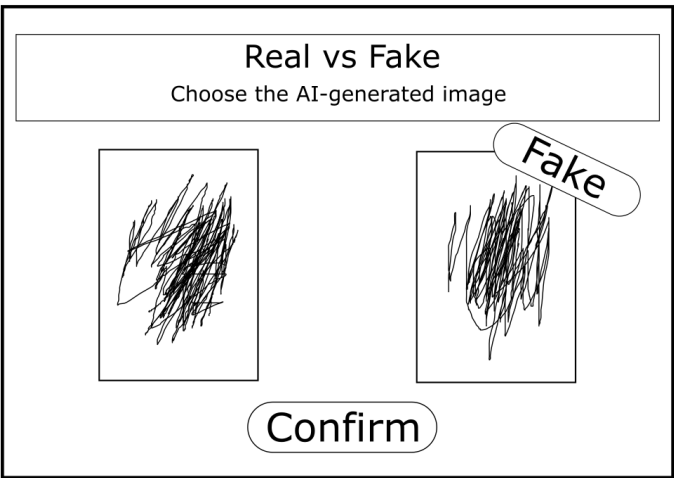


Figure I.11: An early sketch of the *Real vs Fake* (G2) game mode, where players are tasked with identifying the Fake (AI-generated) image from each pair.

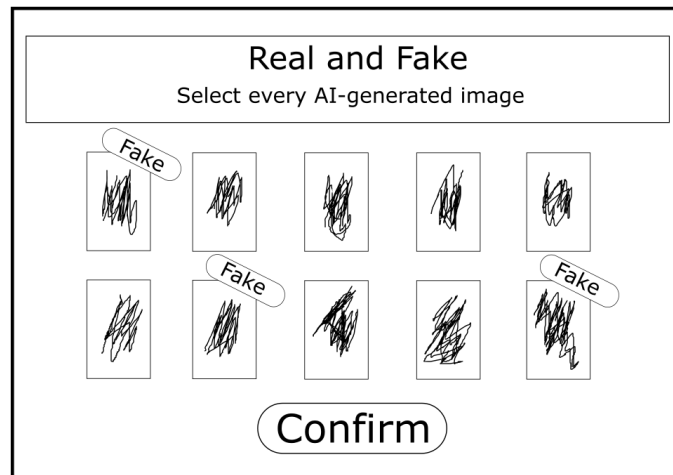


Figure I.12: An early sketch of the *Real and Fake* (G3) game mode, where players are tasked with classifying every image from the group as Real or Fake (AI-generated).

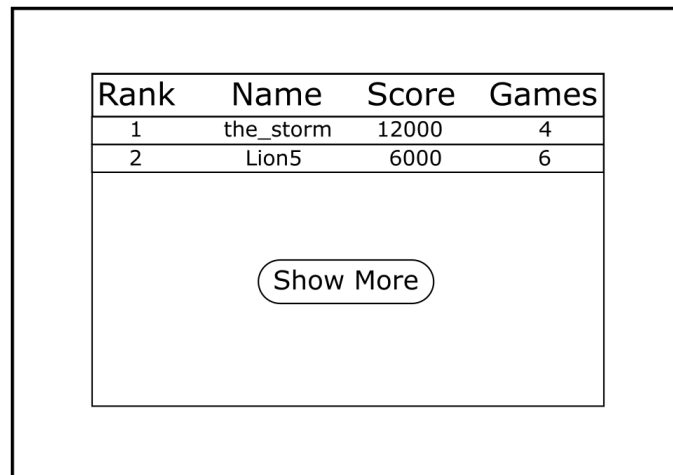


Figure I.13: An early sketch of a leaderboard for *Adversarial Imagery* depicting players' ranks, names, scores, and number of played games.

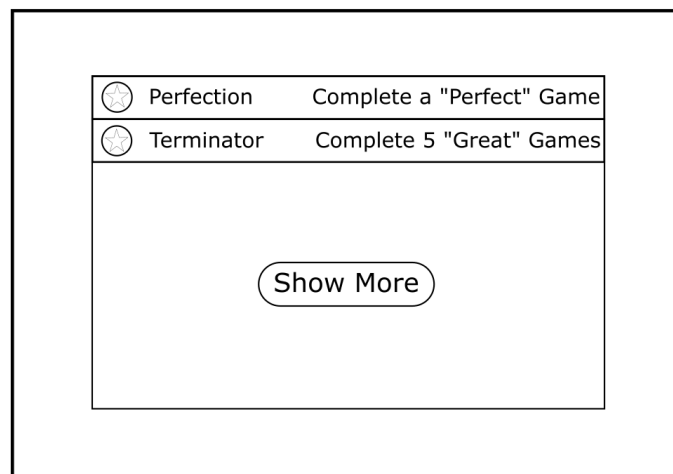


Figure I.14: An early sketch of *Adversarial Imagery's* achievements list to be unlocked through skillful gameplay.

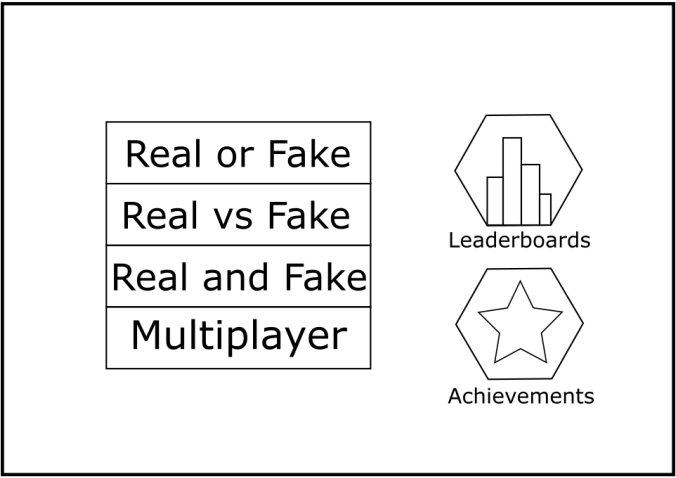


Figure I.15: An early sketch of *Adversarial Imagery*'s main menu showcasing the buttons to access different game modes, leaderboards, and achievements.



Figure I.16: A screenshot of *Adversarial Imagery*'s "Log In" screen.



Figure I.17: A screenshot of *Adversarial Imagery*'s "Sign Up" screen.

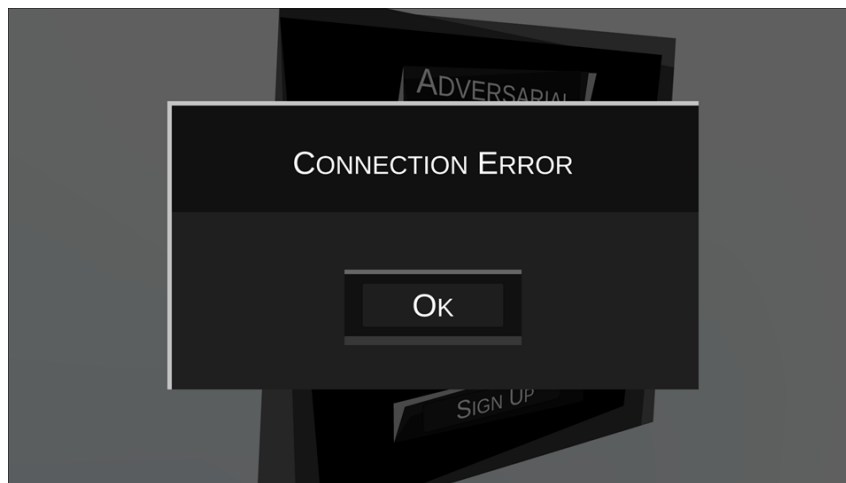


Figure I.18: A screenshot of a warning displayed to players of *Adversarial Imagery* after an internet connection error.

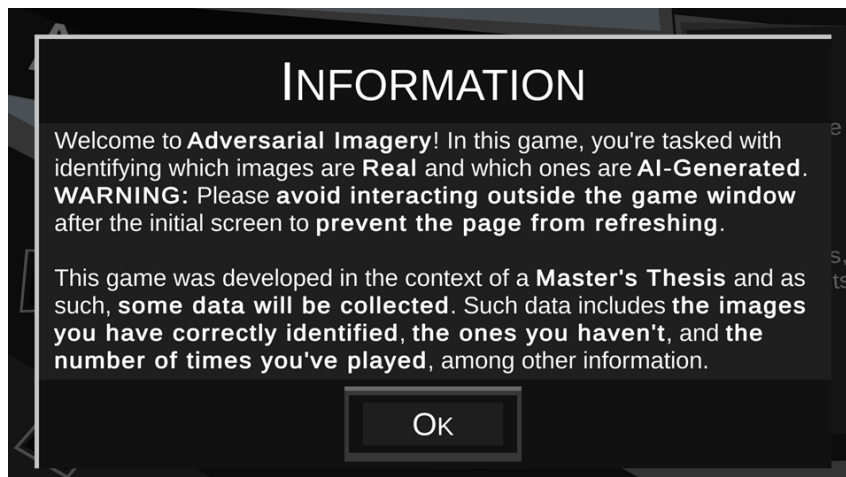


Figure I.19: A screenshot of *Adversarial Imagery*'s consent form.



Figure I.20: A screenshot of *Adversarial Imagery*'s player selection question regarding the player's English comprehension level.

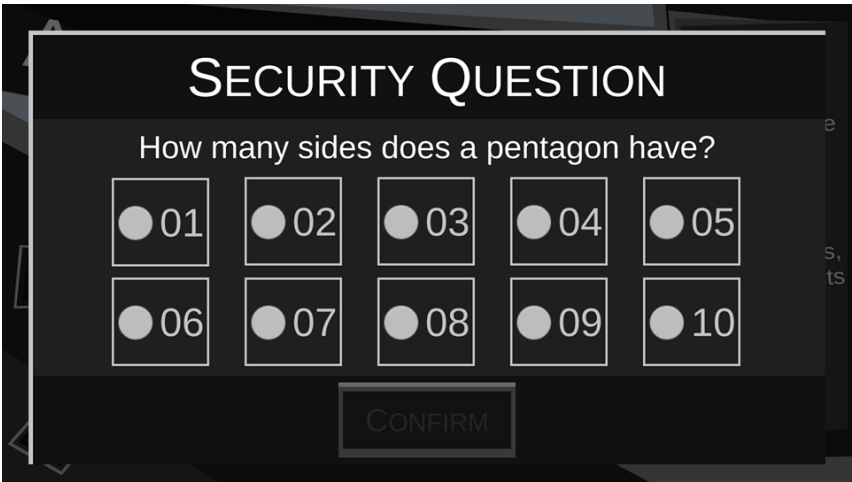
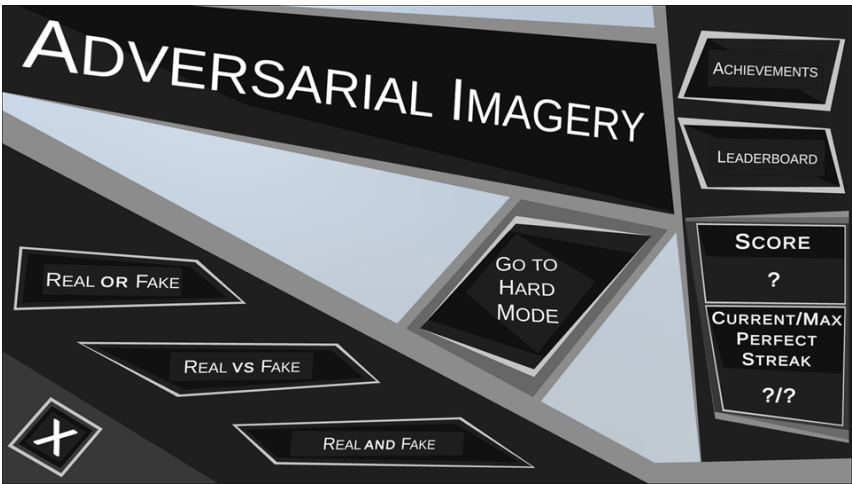
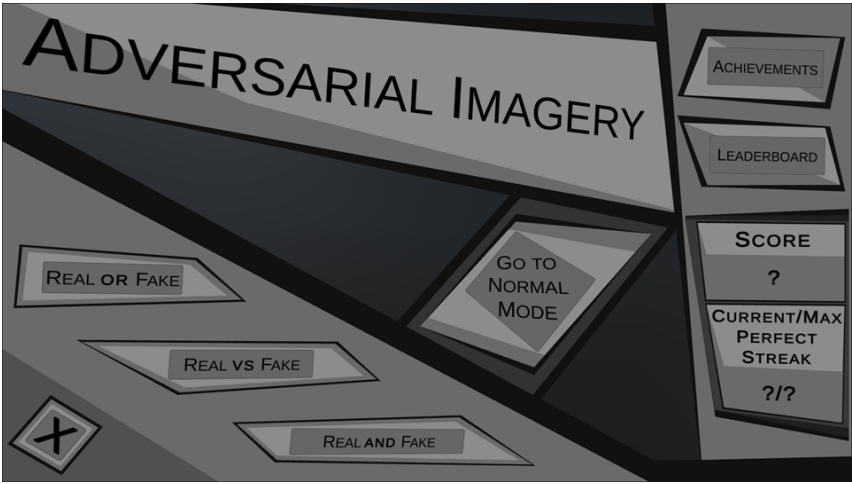


Figure I.21: A screenshot of *Adversarial Imagery*'s player selection question regarding the player's cognitive capabilities.

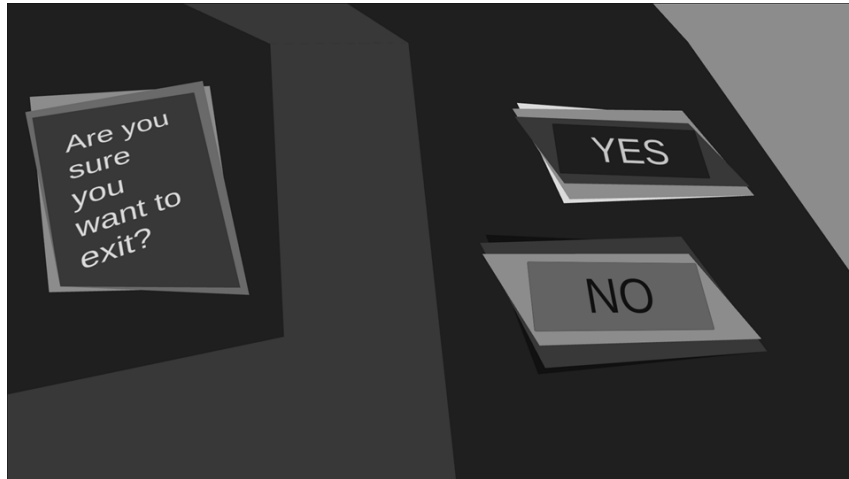


(a) A screenshot of the normal mode's main menu from *Adversarial Imagery*.

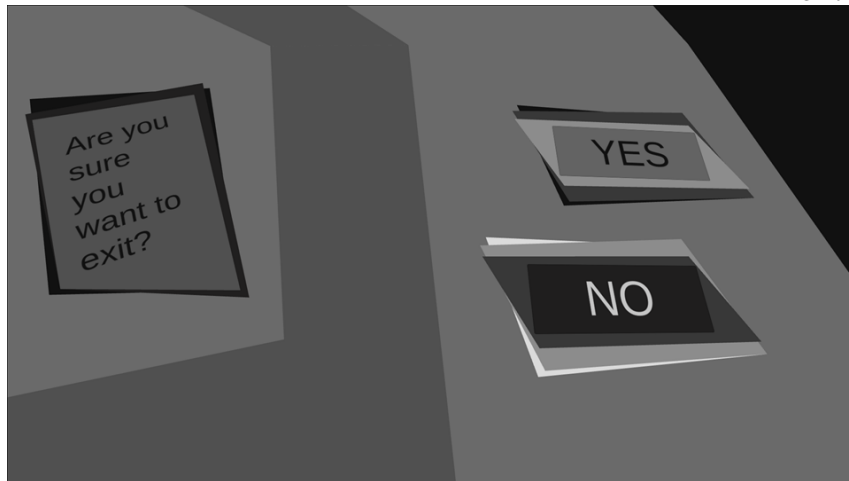


(b) A screenshot of the hard mode's main menu from *Adversarial Imagery*.

Figure I.22: Screenshots of *Adversarial Imagery*'s main menus for each difficulty mode.



(a) A screenshot of the normal mode's exit menu from *Adversarial Imagery*.



(b) A screenshot of the hard mode's exit menu from *Adversarial Imagery*.

Figure I.23: Screenshots of *Adversarial Imagery*'s exit menus for each difficulty mode.

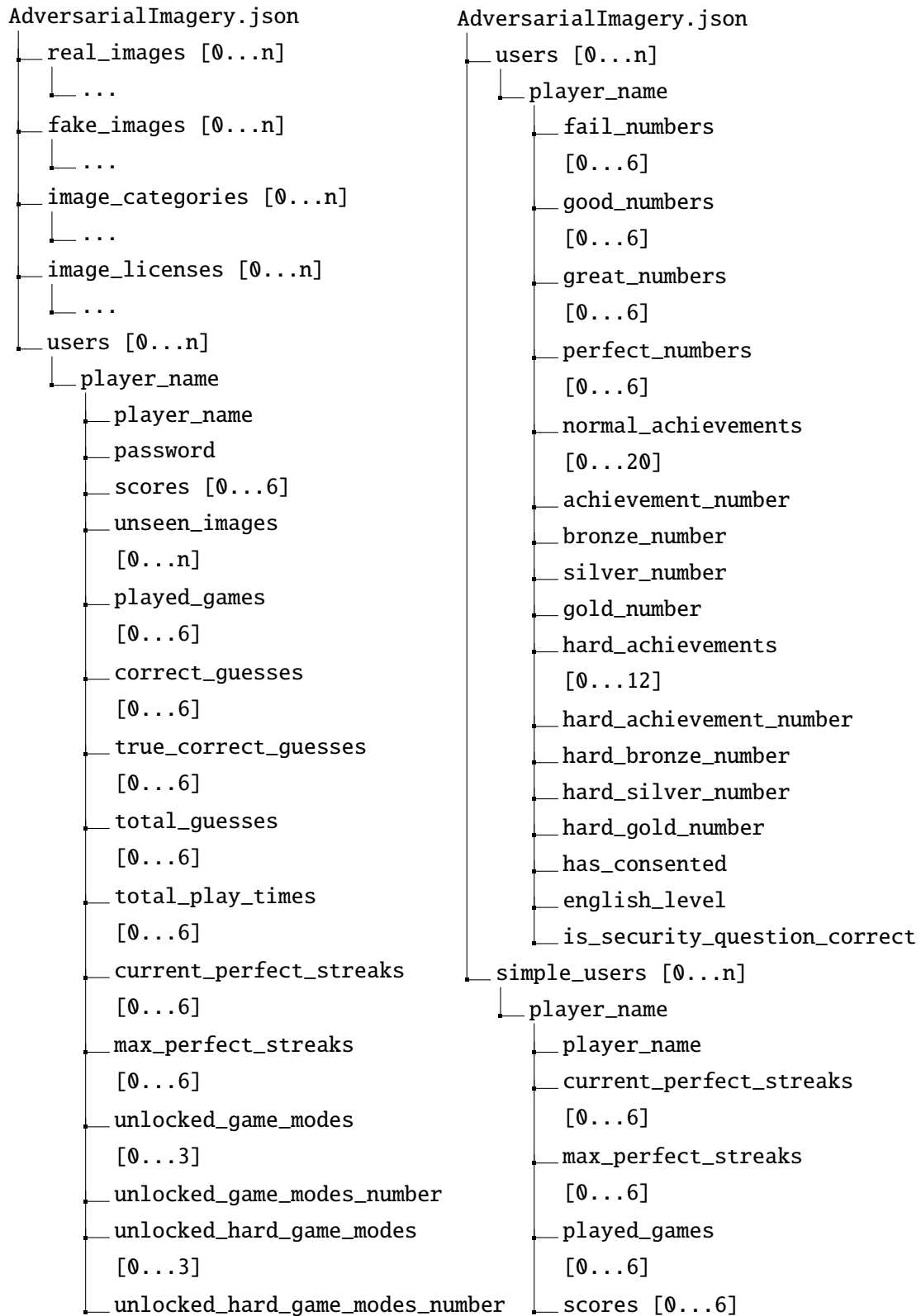


Figure I.24: The complete structure of *Adversarial Imagery*'s resulting data, including the initial image metadata.

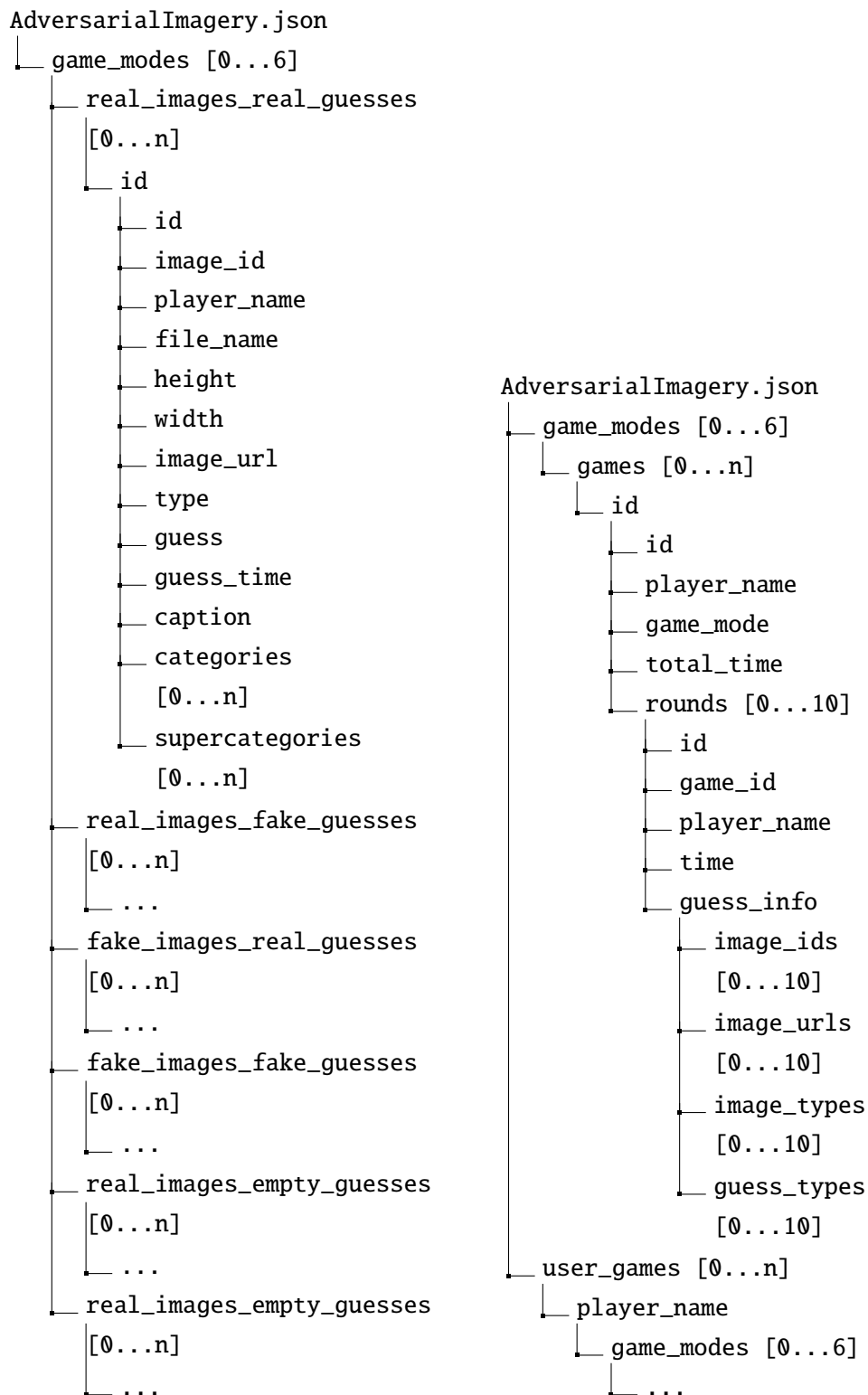
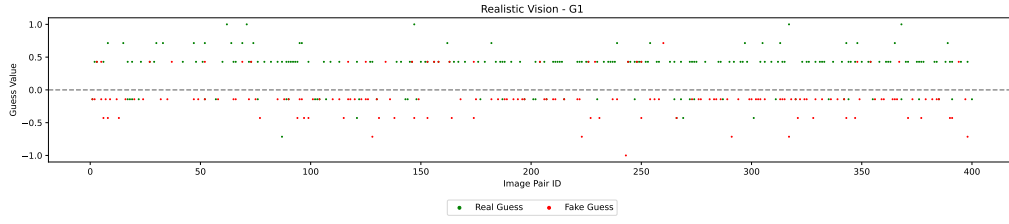
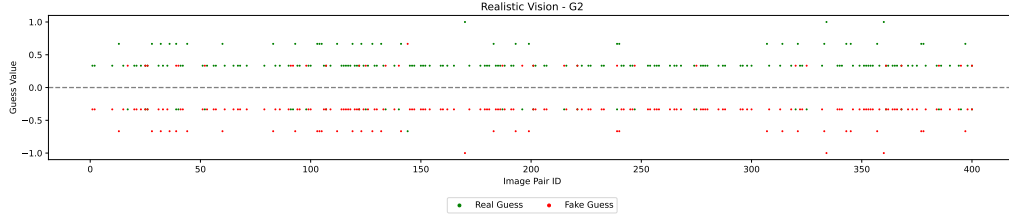
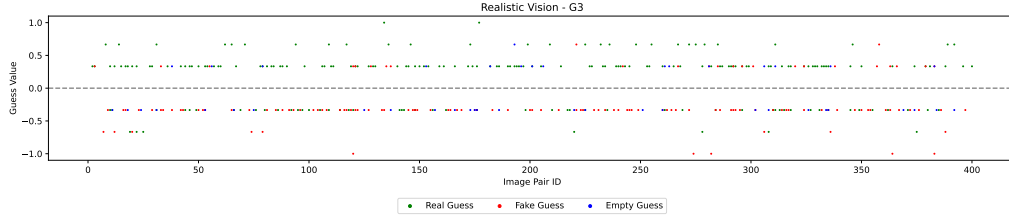
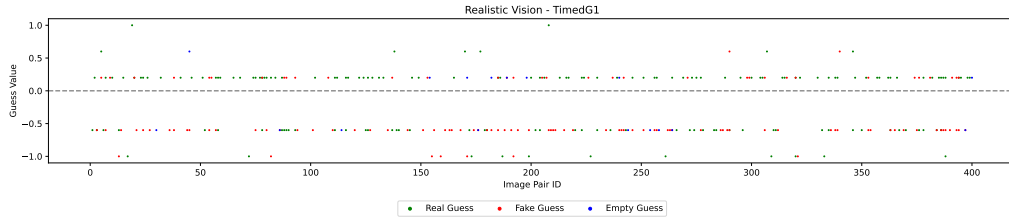
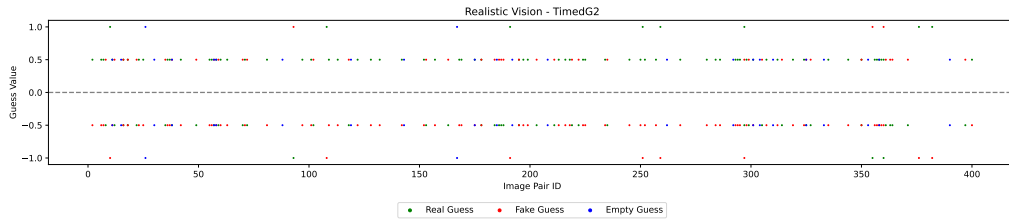
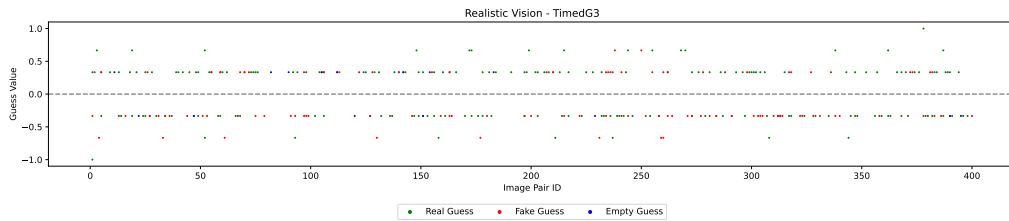
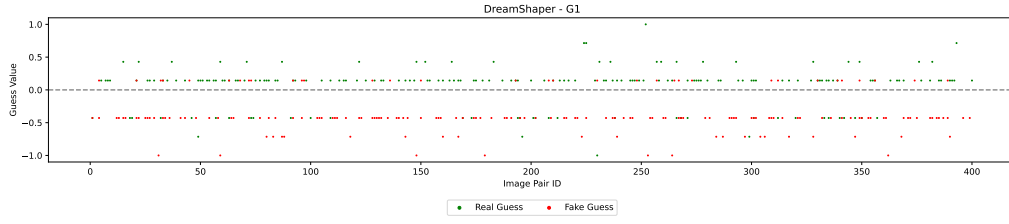
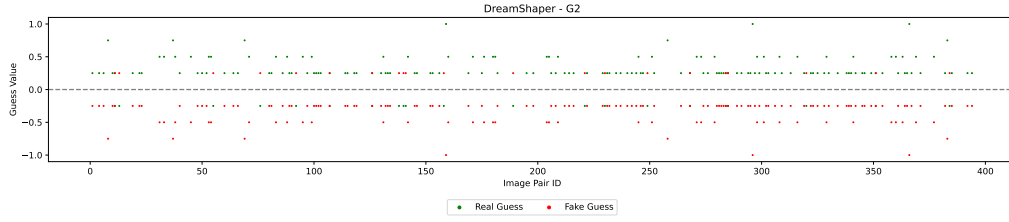


Figure I.24: The complete structure of *Adversarial Imagery's* resulting data, including the initial image metadata (cont.)

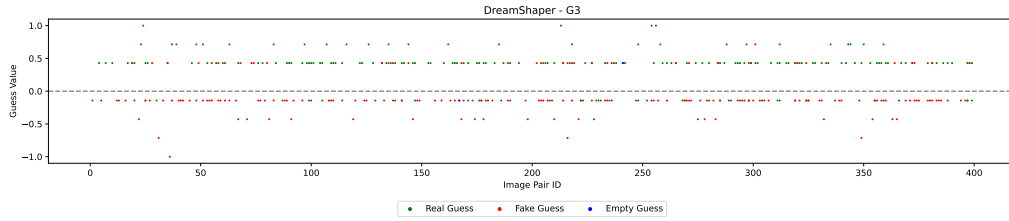
(a) The distribution of *Realistic Vision*'s dataset labels obtained from G1.(b) The distribution of *Realistic Vision*'s dataset labels obtained from G2.(c) The distribution of *Realistic Vision*'s dataset labels obtained from G3.(d) The distribution of *Realistic Vision*'s dataset labels obtained from TimedG1.(e) The distribution of *Realistic Vision*'s dataset labels obtained from TimedG2.(f) The distribution of *Realistic Vision*'s dataset labels obtained from TimedG3.Figure I.25: The distribution of *Realistic Vision*'s dataset labels obtained from *Adversarial Imagery*'s different game modes.



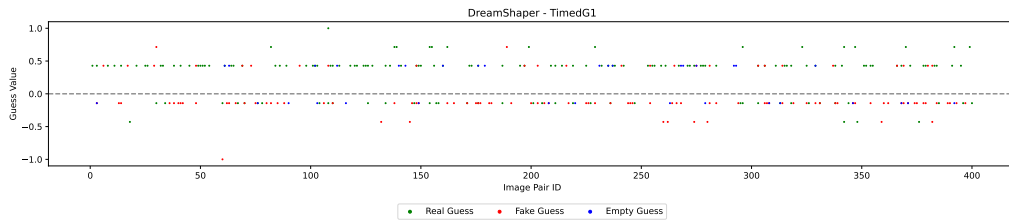
(a) The distribution of *DreamShapers's* dataset labels obtained from G1.



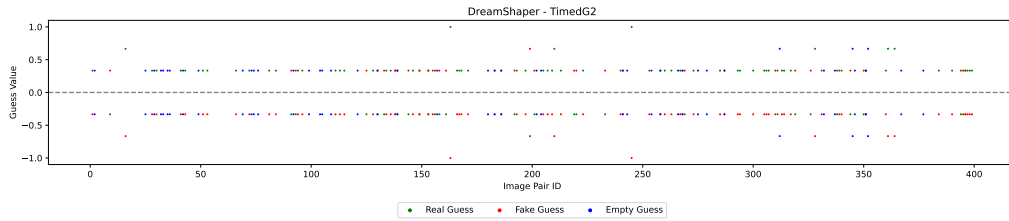
(b) The distribution of *DreamShapers's* dataset labels obtained from G2.



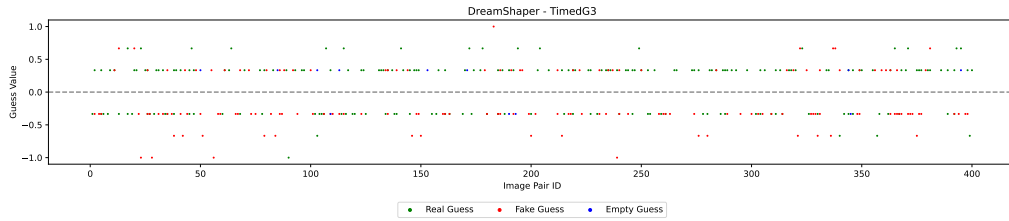
(c) The distribution of *DreamShapers's* dataset labels obtained from G3.



(d) The distribution of *DreamShapers's* dataset labels obtained from TimedG1.

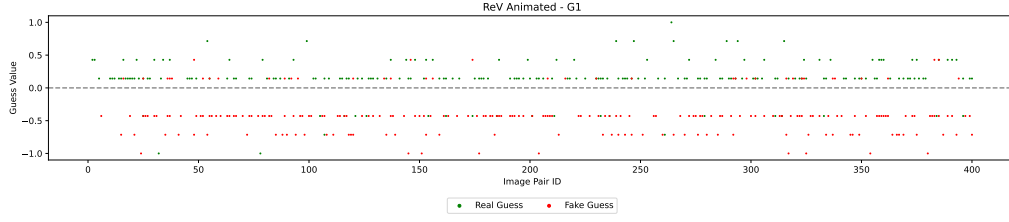
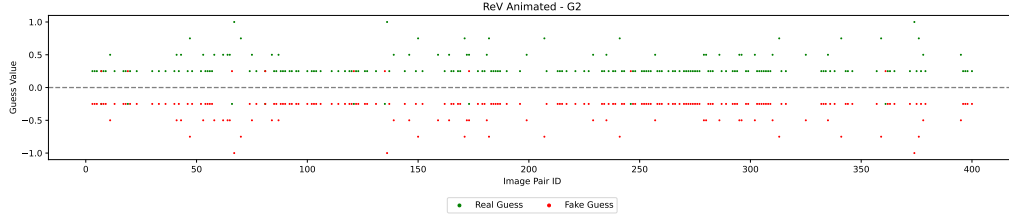
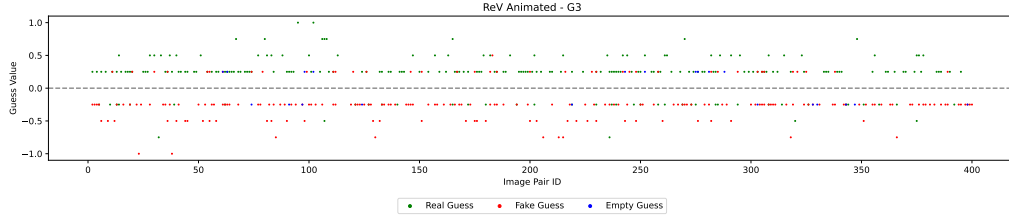
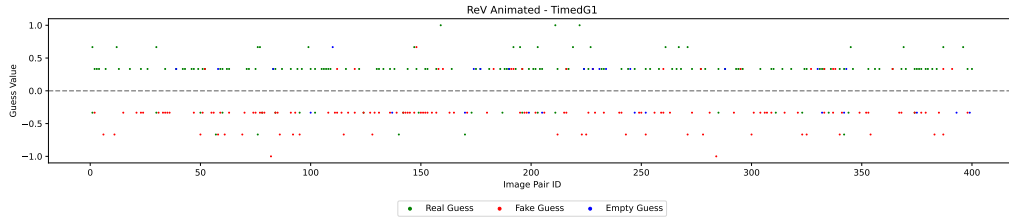
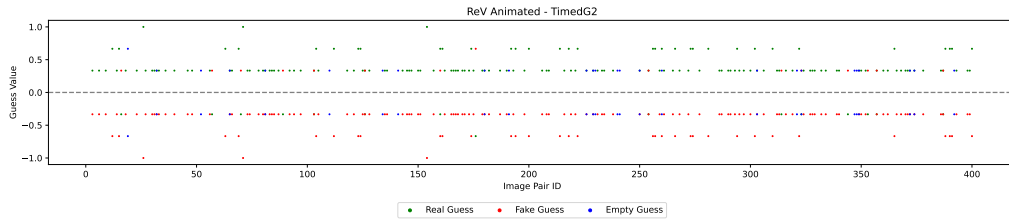
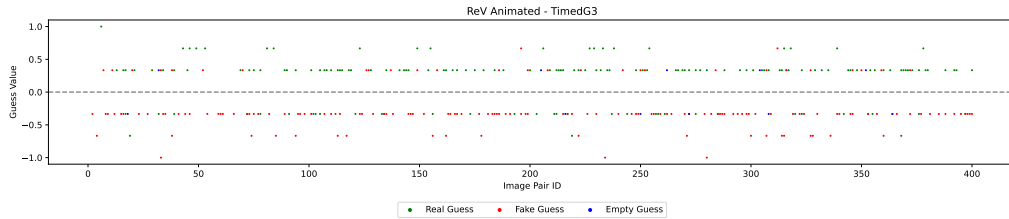


(e) The distribution of *DreamShapers's* dataset labels obtained from TimedG2.



(f) The distribution of *DreamShapers's* dataset labels obtained from TimedG3.

Figure I.26: The distribution of *DreamShapers's* dataset labels obtained from *Adversarial Imagery's* different game modes.

(a) The distribution of *ReV Animated*'s dataset labels obtained from G1.(b) The distribution of *ReV Animated*'s dataset labels obtained from G2.(c) The distribution of *ReV Animated*'s dataset labels obtained from G3.(d) The distribution of *ReV Animated*'s dataset labels obtained from TimedG1.(e) The distribution of *ReV Animated*'s dataset labels obtained from TimedG2.(f) The distribution of *ReV Animated*'s dataset labels obtained from TimedG3.Figure I.27: The distribution of *ReV Animated*'s dataset labels obtained from *Adversarial Imagery*'s different game modes.



(a) Caption: "A building with a parking lot has a field across the street." (*Realistic Vision*; ID = 25).



(b) Caption: "Rolls sit on a plate of oil and Parmesan cheese." (*Realistic Vision*; ID = 220).



(c) Caption: "A teddy bear is scrunched down on the shelf" (*DreamShaper*; ID = 49).



(d) Caption: "Colorful kites with long tails against a mountain backdrop." (*DreamShaper*; ID = 230).



(e) Caption: "a glass of wine and a wine bottle on a table" (*ReV Animated*; ID = 32).



(f) Caption: "A close up picture of green succulents and their leaves." (*ReV Animated*; ID = 320).

Figure I.28: Manually picked AI-generated images based on *Adversarial Imagery*'s label distribution.

II

TABLE ANNEX

Table II.1: Image statistics obtained and computed from players' first playthrough of *Adversarial Imagery* during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	85	38	28	69	0	0	220
G2	90	20	20	90	0	0	220
G3	81	30	35	66	5	3	220
TimedG1	80	19	34	64	14	9	220
TimedG2	51	27	27	51	32	32	220
TimedG3	53	21	37	37	36	36	220

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.75	0.69	0.72
G2	0.82	0.82	0.82
G3	0.70	0.70	0.70
TimedG1	0.70	0.71	0.70
TimedG2	0.65	0.46	0.54
TimedG3	0.59	0.48	0.53

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.64	0.71	0.68
G2	0.82	0.82	0.82
G3	0.69	0.63	0.66
TimedG1	0.77	0.60	0.67
TimedG2	0.65	0.46	0.54
TimedG3	0.64	0.34	0.44

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.70	0.29	0.31	0.00	0.00
G2	0.82	0.18	0.18	0.00	0.00
G3	0.67	0.34	0.26	0.03	0.04
TimedG1	0.65	0.32	0.17	0.08	0.12
TimedG2	0.46	0.25	0.25	0.29	0.29
TimedG3	0.41	0.34	0.19	0.33	0.33

Table II.2: Image statistics obtained and computed from players' second playthrough of *Adversarial Imagery's* normal game modes, aware of the game's instructions, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	93	22	18	87	0	0	220
G2	101	9	9	101	0	0	220
G3	101	17	13	86	2	1	220

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.84	0.81	0.82
G2	0.92	0.92	0.92
G3	0.89	0.84	0.86

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.80	0.83	0.81
G2	0.92	0.92	0.92
G3	0.83	0.86	0.85

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.82	0.17	0.19	0.00	0.00
G2	0.92	0.08	0.08	0.00	0.00
G3	0.85	0.13	0.14	0.01	0.02

Table II.3: Image statistics obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with short round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	71	15	30	63	18	23	220
TimedG2	48	25	25	48	37	37	220
TimedG3	68	16	31	59	30	16	220

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.70	0.68	0.69
TimedG2	0.66	0.44	0.52
TimedG3	0.69	0.60	0.64

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.81	0.54	0.65
TimedG2	0.66	0.44	0.52
TimedG3	0.79	0.56	0.65

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.61	0.26	0.14	0.20	0.17
TimedG2	0.44	0.23	0.23	0.34	0.34
TimedG3	0.58	0.29	0.14	0.15	0.26

Table II.4: Image statistics obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with intermediate round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	75	17	32	88	6	2	220
TimedG2	77	15	15	77	18	18	220
TimedG3	88	25	32	54	10	11	220

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.70	0.77	0.73
TimedG2	0.84	0.70	0.76
TimedG3	0.73	0.72	0.72

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.84	0.72	0.78
TimedG2	0.84	0.70	0.76
TimedG3	0.68	0.56	0.61

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.74	0.26	0.17	0.02	0.06
TimedG2	0.70	0.14	0.14	0.16	0.16
TimedG3	0.65	0.33	0.20	0.11	0.08

Table II.5: Image statistics obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with long round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	69	31	30	83	3	4	220
TimedG2	82	20	20	82	8	8	220
TimedG3	82	28	43	57	8	2	220

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.70	0.67	0.68
TimedG2	0.80	0.75	0.77
TimedG3	0.66	0.69	0.67

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.73	0.71	0.72
TimedG2	0.80	0.75	0.77
TimedG3	0.67	0.56	0.61

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.69	0.26	0.30	0.03	0.03
TimedG2	0.75	0.18	0.18	0.07	0.07
TimedG3	0.63	0.42	0.24	0.02	0.07

Table II.6: The statistics of every image with the "Person" supercategory obtained and computed from players' first playthrough of *Adversarial Imagery* during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	46	15	7	38	0	0	106
G2	47	6	6	47	0	0	106
G3	45	15	17	36	1	2	116
TimedG1	38	4	16	34	8	4	104
TimedG2	28	17	17	28	15	15	120
TimedG3	25	11	16	19	17	15	103

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.87	0.75	0.81
G2	0.89	0.89	0.89
G3	0.73	0.74	0.73
TimedG1	0.70	0.76	0.73
TimedG2	0.62	0.47	0.53
TimedG3	0.61	0.47	0.53

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.72	0.84	0.78
G2	0.89	0.89	0.89
G3	0.71	0.65	0.68
TimedG1	0.89	0.63	0.74
TimedG2	0.62	0.47	0.53
TimedG3	0.63	0.38	0.48

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.79	0.16	0.25	0.00	0.00
G2	0.89	0.11	0.11	0.00	0.00
G3	0.70	0.31	0.25	0.04	0.02
TimedG1	0.69	0.30	0.08	0.07	0.16
TimedG2	0.47	0.28	0.28	0.25	0.25
TimedG3	0.43	0.32	0.21	0.30	0.32

Table II.7: The statistics of every image with the "Person" supercategory obtained and computed from players' second playthrough of *Adversarial Imagery's* normal game modes, aware of the game's instructions, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	49	14	7	48	0	0	118
G2	52	1	1	52	0	0	106
G3	59	4	4	42	1	0	110

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.88	0.78	0.82
G2	0.98	0.98	0.98
G3	0.94	0.92	0.93

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.77	0.87	0.82
G2	0.98	0.98	0.98
G3	0.91	0.91	0.91

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.82	0.13	0.22	0.00	0.00
G2	0.98	0.02	0.02	0.00	0.00
G3	0.92	0.09	0.06	0.00	0.02

Table II.8: The statistics of every image with the "Person" supercategory obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with short round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	31	8	10	32	9	13	103
TimedG2	26	9	9	26	20	20	110
TimedG3	36	7	12	32	14	6	107

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.76	0.65	0.70
TimedG2	0.74	0.47	0.58
TimedG3	0.75	0.63	0.69

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.80	0.58	0.67
TimedG2	0.74	0.47	0.58
TimedG3	0.82	0.64	0.72

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.61	0.18	0.17	0.24	0.19
TimedG2	0.47	0.16	0.16	0.36	0.36
TimedG3	0.64	0.24	0.12	0.12	0.25

Table II.9: The statistics of every image with the "Person" supercategory obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with intermediate round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	38	6	14	55	5	1	119
TimedG2	44	6	6	44	6	6	112
TimedG3	52	11	11	25	3	7	109

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.73	0.78	0.75
TimedG2	0.88	0.79	0.83
TimedG3	0.83	0.79	0.81

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.90	0.79	0.84
TimedG2	0.88	0.79	0.83
TimedG3	0.69	0.58	0.63

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.78	0.20	0.12	0.01	0.10
TimedG2	0.79	0.11	0.11	0.11	0.11
TimedG3	0.71	0.26	0.17	0.16	0.05

Table II.10: The statistics of every image with the "Person" supercategory obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with long round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	25	13	7	43	0	1	89
TimedG2	46	8	8	46	6	6	120
TimedG3	39	16	18	25	3	1	102

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.78	0.66	0.71
TimedG2	0.85	0.77	0.81
TimedG3	0.68	0.67	0.68

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.77	0.84	0.80
TimedG2	0.85	0.77	0.81
TimedG3	0.61	0.57	0.59

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.76	0.14	0.34	0.02	0.00
TimedG2	0.77	0.13	0.13	0.10	0.10
TimedG3	0.63	0.41	0.28	0.02	0.05

Table II.11: The statistics of every image without the "Person" supercategory obtained and computed from players' first playthrough of *Adversarial Imagery* during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	39	23	21	31	0	0	114
G2	43	14	14	43	0	0	114
G3	36	15	18	30	4	1	104
TimedG1	42	15	18	30	6	5	116
TimedG2	23	10	10	23	17	17	100
TimedG3	28	10	21	18	19	21	117

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.65	0.63	0.64
G2	0.75	0.75	0.75
G3	0.67	0.65	0.66
TimedG1	0.70	0.67	0.68
TimedG2	0.70	0.46	0.55
TimedG3	0.57	0.49	0.53

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.57	0.60	0.58
G2	0.75	0.75	0.75
G3	0.67	0.61	0.64
TimedG1	0.67	0.57	0.61
TimedG2	0.70	0.46	0.55
TimedG3	0.64	0.30	0.41

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.61	0.40	0.37	0.00	0.00
G2	0.75	0.25	0.25	0.00	0.00
G3	0.63	0.37	0.27	0.02	0.07
TimedG1	0.62	0.34	0.24	0.09	0.10
TimedG2	0.46	0.20	0.20	0.34	0.34
TimedG3	0.39	0.35	0.18	0.35	0.33

Table II.12: The statistics of every image without the "Person" supercategory obtained and computed from players' second playthrough of *Adversarial Imagery's* normal game modes, aware of the game's instructions, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	44	8	11	39	0	0	102
G2	49	8	8	49	0	0	114
G3	42	13	9	44	1	1	110

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.80	0.85	0.82
G2	0.86	0.86	0.86
G3	0.82	0.75	0.79

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.83	0.78	0.80
G2	0.86	0.86	0.86
G3	0.77	0.81	0.79

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.81	0.22	0.15	0.00	0.00
G2	0.86	0.14	0.14	0.00	0.00
G3	0.78	0.17	0.23	0.02	0.02

Table II.13: The statistics of every image without the "Person" supercategory obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with short round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	40	7	20	31	9	10	117
TimedG2	22	16	16	22	17	17	110
TimedG3	32	9	19	27	16	10	113

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.67	0.71	0.69
TimedG2	0.58	0.40	0.47
TimedG3	0.63	0.56	0.59

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.82	0.51	0.63
TimedG2	0.58	0.40	0.47
TimedG3	0.75	0.48	0.59

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.61	0.33	0.13	0.16	0.16
TimedG2	0.40	0.29	0.29	0.31	0.31
TimedG3	0.52	0.34	0.16	0.18	0.28

Table II.14: The statistics of every image without the "Person" supercategory obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with intermediate round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	37	11	18	33	1	1	101
TimedG2	33	9	9	33	12	12	108
TimedG3	36	14	21	29	7	4	111

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.67	0.76	0.71
TimedG2	0.79	0.61	0.69
TimedG3	0.63	0.63	0.63

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.75	0.63	0.69
TimedG2	0.79	0.61	0.69
TimedG3	0.67	0.54	0.60

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.69	0.35	0.22	0.02	0.02
TimedG2	0.61	0.17	0.17	0.22	0.22
TimedG3	0.59	0.39	0.25	0.07	0.12

Table II.15: The statistics of every image without the "Person" supercategory obtained and computed from players' playthrough of *Adversarial Imagery's* hard game modes, with long round times, during the formative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
TimedG1	44	18	23	40	3	3	131
TimedG2	36	12	12	36	2	2	100
TimedG3	43	12	25	32	5	1	118

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
TimedG1	0.66	0.68	0.67
TimedG2	0.75	0.72	0.73
TimedG3	0.63	0.72	0.67

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
TimedG1	0.69	0.61	0.65
TimedG2	0.75	0.72	0.73
TimedG3	0.73	0.55	0.63

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
TimedG1	0.64	0.35	0.28	0.05	0.05
TimedG2	0.72	0.24	0.24	0.04	0.04
TimedG3	0.64	0.43	0.20	0.02	0.08

Table II.16: Unfiltered general statistics obtained from *Adversarial Imagery*.

Game Statistic	Realistic Vision	DreamShaper	ReV Animated	Total
Total Players	44	43	42	129
Total Time (h.)	02:29	02:54	02:42	08:05
Total Games	252	247	299	798
Total Guesses	2,520	2,470	2,990	7,980
Total Correct Guesses	1,789	1,834	2,521	6,144
Correct Guess Ratio	0.71	0.74	0.84	-

Table II.17: The unfiltered number of sessions played by each player of *Adversarial Imagery* grouped by their assigned databases.(a) *Realistic Vision's* game sessions.

Game Mode ID	Total	Median	Q1	Q3	Min	Max	Mode
G1	50	1	1	1	0	7	1
G2	52	1	1	1	0	11	1
G3	48	1	1	1	0	9	1
TimedG1	34	1	0	1	0	8	1
TimedG2	32	1	0	1	0	7	1
TimedG3	36	1	0	1	0	10	1

(b) *DreamShaper's* game sessions.

Game Mode ID	Total	Median	Q1	Q3	Min	Max	Mode
G1	45	1	1	1	0	3	1
G2	47	1	1	1	0	4	1
G3	46	1	1	1	0	4	1
TimedG1	34	0	0	1	0	5	0
TimedG2	33	1	0	1	0	4	0
TimedG3	42	0	0	1	0	5	0

(c) *ReV Animated's* game sessions.

Game Mode ID	Total	Median	Q1	Q3	Min	Max	Mode
G1	59	1	1	1	0	10	1
G2	53	1	1	1	0	10	1
G3	61	1	1	1	0	12	1
TimedG1	40	1	0	1	0	10	0
TimedG2	48	1	0	1	0	11	0
TimedG3	38	1	0	1	0	10	0

Table II.18: Realistic Vision’s image statistics obtained and computed from players’ playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	196	28	55	171	0	0	450
G2	214	31	31	214	0	0	490
G3	177	27	66	125	17	28	440
TimedG1	109	38	73	92	9	9	330
TimedG2	83	45	45	83	27	27	310
TimedG3	125	40	92	78	8	7	350

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.78	0.88	0.83
G2	0.87	0.87	0.87
G3	0.73	0.80	0.76
TimedG1	0.60	0.70	0.64
TimedG2	0.65	0.54	0.59
TimedG3	0.58	0.72	0.64

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.86	0.76	0.80
G2	0.87	0.87	0.87
G3	0.82	0.57	0.67
TimedG1	0.71	0.53	0.61
TimedG2	0.65	0.54	0.59
TimedG3	0.66	0.44	0.53

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.82	0.24	0.13	0.00	0.00
G2	0.87	0.13	0.13	0.00	0.00
G3	0.69	0.30	0.12	0.13	0.08
TimedG1	0.61	0.42	0.24	0.05	0.06
TimedG2	0.54	0.29	0.29	0.17	0.17
TimedG3	0.58	0.52	0.23	0.04	0.05

Table II.19: DreamShaper’s image statistics obtained and computed from players’ playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	184	33	43	180	0	0	440
G2	209	21	21	209	0	0	460
G3	190	41	49	168	1	1	450
TimedG1	132	29	46	99	18	16	340
TimedG2	84	33	33	84	48	48	330
TimedG3	144	59	81	124	8	4	420

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.81	0.85	0.83
G2	0.91	0.91	0.91
G3	0.79	0.82	0.81
TimedG1	0.74	0.74	0.74
TimedG2	0.72	0.51	0.60
TimedG3	0.64	0.68	0.66

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.85	0.81	0.83
G2	0.91	0.91	0.91
G3	0.80	0.77	0.79
TimedG1	0.77	0.61	0.69
TimedG2	0.72	0.51	0.60
TimedG3	0.68	0.59	0.63

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.83	0.19	0.15	0.00	0.00
G2	0.91	0.09	0.09	0.00	0.00
G3	0.80	0.22	0.18	0.00	0.00
TimedG1	0.68	0.29	0.16	0.10	0.10
TimedG2	0.51	0.20	0.20	0.29	0.29
TimedG3	0.64	0.39	0.28	0.02	0.04

Table II.20: ReV Animated’s image statistics obtained and computed from players’ playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	236	44	29	261	0	0	570
G2	256	9	9	256	0	0	530
G3	247	41	51	230	10	11	590
TimedG1	167	23	32	150	16	12	400
TimedG2	197	16	16	197	27	27	480
TimedG3	140	31	39	159	5	6	380

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.89	0.84	0.87
G2	0.97	0.97	0.97
G3	0.83	0.83	0.83
TimedG1	0.84	0.81	0.82
TimedG2	0.92	0.82	0.87
TimedG3	0.78	0.80	0.79

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.86	0.90	0.88
G2	0.97	0.97	0.97
G3	0.85	0.79	0.82
TimedG1	0.87	0.77	0.82
TimedG2	0.92	0.82	0.87
TimedG3	0.84	0.78	0.81

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.87	0.10	0.16	0.00	0.00
G2	0.97	0.03	0.03	0.00	0.00
G3	0.81	0.17	0.14	0.04	0.03
TimedG1	0.79	0.16	0.11	0.06	0.08
TimedG2	0.82	0.07	0.07	0.11	0.11
TimedG3	0.79	0.19	0.18	0.03	0.03

Table II.21: Realistic Vision's statistics of every image with the "Person" supercategory obtained and computed from players' playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	104	13	16	109	0	0	242
G2	118	10	10	118	0	0	256
G3	94	13	31	67	8	19	232
TimedG1	58	18	29	56	4	3	168
TimedG2	47	19	19	47	11	11	154
TimedG3	66	20	45	42	6	2	181

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.87	0.89	0.88
G2	0.92	0.92	0.92
G3	0.75	0.82	0.78
TimedG1	0.67	0.73	0.69
TimedG2	0.71	0.61	0.66
TimedG3	0.59	0.72	0.65

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.89	0.87	0.88
G2	0.92	0.92	0.92
G3	0.84	0.57	0.68
TimedG1	0.76	0.64	0.69
TimedG2	0.71	0.61	0.66
TimedG3	0.68	0.47	0.56

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.88	0.13	0.11	0.00	0.00
G2	0.92	0.08	0.08	0.00	0.00
G3	0.69	0.26	0.11	0.16	0.07
TimedG1	0.68	0.33	0.23	0.03	0.05
TimedG2	0.61	0.25	0.25	0.14	0.14
TimedG3	0.60	0.51	0.22	0.02	0.07

Table II.22: DreamShaper’s statistics of every image with the "Person" supercategory obtained and computed from players’ playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	106	18	12	106	0	0	242
G2	124	9	9	124	0	0	266
G3	103	23	17	102	1	0	246
TimedG1	67	17	12	64	10	8	178
TimedG2	48	17	17	48	20	20	170
TimedG3	78	35	32	85	4	2	236

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.90	0.85	0.88
G2	0.93	0.93	0.93
G3	0.86	0.81	0.83
TimedG1	0.85	0.71	0.77
TimedG2	0.74	0.56	0.64
TimedG3	0.71	0.67	0.69

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.85	0.90	0.88
G2	0.93	0.93	0.93
G3	0.82	0.86	0.84
TimedG1	0.79	0.76	0.78
TimedG2	0.74	0.56	0.64
TimedG3	0.71	0.71	0.71

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.88	0.10	0.15	0.00	0.00
G2	0.93	0.07	0.07	0.00	0.00
G3	0.83	0.14	0.18	0.00	0.01
TimedG1	0.74	0.14	0.18	0.10	0.11
TimedG2	0.56	0.20	0.20	0.24	0.24
TimedG3	0.69	0.27	0.30	0.02	0.03

Table II.23: ReV Animated's statistics of every image with the "Person" supercategory obtained and computed from players' playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	144	19	1	167	0	0	331
G2	150	3	3	150	0	0	306
G3	129	22	17	129	4	8	309
TimedG1	81	14	8	94	10	7	214
TimedG2	113	4	4	113	15	15	264
TimedG3	73	15	9	89	2	3	191

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.99	0.88	0.94
G2	0.98	0.98	0.98
G3	0.88	0.83	0.86
TimedG1	0.91	0.77	0.84
TimedG2	0.97	0.86	0.91
TimedG3	0.89	0.81	0.85

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.90	0.99	0.94
G2	0.98	0.98	0.98
G3	0.85	0.84	0.85
TimedG1	0.87	0.86	0.87
TimedG2	0.97	0.86	0.91
TimedG3	0.86	0.88	0.87

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.94	0.01	0.12	0.00	0.00
G2	0.98	0.02	0.02	0.00	0.00
G3	0.83	0.11	0.14	0.05	0.03
TimedG1	0.82	0.07	0.13	0.06	0.10
TimedG2	0.86	0.03	0.03	0.11	0.11
TimedG3	0.85	0.09	0.17	0.03	0.02

Table II.24: Realistic Vision’s statistics of every image without the "Person" supercategory obtained and computed from players’ playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	92	15	39	62	0	0	208
G2	96	21	21	96	0	0	234
G3	83	14	35	58	9	9	208
TimedG1	51	20	44	36	5	6	162
TimedG2	36	26	26	36	16	16	156
TimedG3	59	20	47	36	2	5	169

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.70	0.86	0.77
G2	0.82	0.82	0.82
G3	0.70	0.78	0.74
TimedG1	0.54	0.67	0.60
TimedG2	0.58	0.46	0.51
TimedG3	0.56	0.73	0.63

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.81	0.61	0.70
G2	0.82	0.82	0.82
G3	0.81	0.57	0.67
TimedG1	0.64	0.42	0.51
TimedG2	0.58	0.46	0.51
TimedG3	0.64	0.41	0.50

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.74	0.39	0.14	0.00	0.00
G2	0.82	0.18	0.18	0.00	0.00
G3	0.68	0.34	0.13	0.09	0.08
TimedG1	0.54	0.51	0.26	0.07	0.07
TimedG2	0.46	0.33	0.33	0.21	0.21
TimedG3	0.56	0.53	0.25	0.06	0.02

Table II.25: DreamShaper’s statistics of every image without the "Person" supercategory obtained and computed from players’ playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	78	15	31	74	0	0	198
G2	85	12	12	85	0	0	194
G3	87	18	32	66	0	1	204
TimedG1	65	12	34	35	8	8	162
TimedG2	36	16	16	36	28	28	160
TimedG3	66	24	49	39	4	2	184

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.72	0.84	0.77
G2	0.88	0.88	0.88
G3	0.73	0.83	0.78
TimedG1	0.66	0.76	0.71
TimedG2	0.69	0.45	0.55
TimedG3	0.57	0.70	0.63

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.83	0.70	0.76
G2	0.88	0.88	0.88
G3	0.79	0.67	0.72
TimedG1	0.74	0.45	0.56
TimedG2	0.69	0.45	0.55
TimedG3	0.62	0.43	0.51

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.77	0.30	0.16	0.00	0.00
G2	0.88	0.12	0.12	0.00	0.00
G3	0.75	0.32	0.17	0.01	0.00
TimedG1	0.62	0.44	0.14	0.10	0.09
TimedG2	0.45	0.20	0.20	0.35	0.35
TimedG3	0.57	0.54	0.26	0.02	0.04

Table II.26: ReV Animated’s statistics of every image without the "Person" supercategory obtained and computed from players’ playthroughs of *Adversarial Imagery* during the summative study.

(a) The number of real images classified as real (RR), the number of real images classified as fake (RF), the number of fake images classified as real (FR), the number of fake images classified as fake (FF), the number of real images without classification (RE), the number of fake images without classification (FE).

Game Mode ID	RR	RF	FR	FF	RE	FE	Total Images
G1	92	25	28	94	0	0	239
G2	106	6	6	106	0	0	224
G3	118	19	34	101	6	3	281
TimedG1	86	9	24	56	6	5	186
TimedG2	84	12	12	84	12	12	216
TimedG3	67	16	30	70	3	3	189

(b) The precision, recall, and F1 scores of real images.

Game Mode ID	Precision (Real)	Recall (Real)	F1 (Real)
G1	0.77	0.79	0.78
G2	0.95	0.95	0.95
G3	0.78	0.83	0.80
TimedG1	0.78	0.85	0.82
TimedG2	0.88	0.78	0.82
TimedG3	0.69	0.78	0.73

(c) The precision, recall, and F1 scores of fake images.

Game Mode ID	Precision (Fake)	Recall (Fake)	F1 (Fake)
G1	0.79	0.77	0.78
G2	0.95	0.95	0.95
G3	0.84	0.73	0.78
TimedG1	0.86	0.66	0.75
TimedG2	0.88	0.78	0.82
TimedG3	0.81	0.68	0.74

(d) The classification accuracy and the rates of images classified with incorrect guesses.

Game Mode ID	Accuracy	F as R	R as F	F as E	R as E
G1	0.78	0.23	0.21	0.00	0.00
G2	0.95	0.05	0.05	0.00	0.00
G3	0.78	0.25	0.13	0.02	0.04
TimedG1	0.76	0.28	0.09	0.06	0.06
TimedG2	0.78	0.11	0.11	0.11	0.11
TimedG3	0.72	0.29	0.19	0.03	0.03



2025 Adversarial Imagery: The Potential of Gamified Benchmarking in Evaluating Deep Generative Models of Images

João Domingues

