A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Enhancing Design Inspiration through AI-Driven Keyword suggestions and Image Captioning

Frank Mohammed Tamer – 55684

Work project carried out under the supervision of:

Rongjiao Ji (*Nova SBE*)

Fransisco Tenente (*3DWays*)

01/02/2024

Abstract

This thesis explores AI's potential in enhancing design creativity, focusing on a machine learning platform that suggests visuals from keywords, using OpenAI and Meta models. It learns from interactions, encouraging exploration. Research compared image captioning models, identifying BLIP as superior for providing accurate, context-rich descriptions. The study shows AI's ability to bridge the gap between abstract ideas and visuals, promote exploration with relevant suggestions, and improve design tools with precise descriptions, underscoring the importance of collaboration and ongoing research in maximizing AI's impact in design.

Keywords: Product Development, Machine learning, Web Application, Image Segmentation, Image Inpainting, Multimodal AI Applications, Image Suggestions, Image Captioning, Recommender Systems

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

1 Introduction (group part)

1.1 Background

In a panorama where solutions exist for nearly every conceivable issue, Artificial Intelligence (AI) has played a significant role, with algorithms delivering time-saving results or even autonomously executing tasks beyond their initial training scope. However, there is still a big gap in the integrations AI can perform. Even most comprehensive models often fall short in combining multiple functionalities, which may involve different code realms, to form a complete tool tailored specifically to an organization's objectives (Cox 2023). The work produced in this thesis encompasses the creation of a unique solution addressing a creative and inspiration challenge identified by 3DWays while providing their services to customers.

3DWays is a Portuguese start-up specialized in product design, dedicated to assisting investors in industrializing new products and getting them to the market. Their services range from conceptualizing a new product to its final implementation and usage, aiming to be a "One-Stop Hardware Partner" (3DWays 2023a). The company's customers include other companies working in one of the sectors from product design to manufacture, as well as individual inventors looking to industrialize their creations.

3DWays mission, defined in the company's LinkedIn page:

"We perform user and market research, eco-design, prototyping, supply chain management and contract manufacturing. Through our partner network, we design business models and pricing strategies, we help get products certified and patented and we help them raise funds and hardware-related legal advisory." (3DWays 2023b)

Such a complete service package is a big help for other startups in avoiding risks associated with manufacturing hardware, since there is more time to refine the product's specifications that will differentiate itself from the rest of the market.

Besides operating at the prototyping and manufacturing levels, with the assistance of an advanced 3D printing facility, 3DWays is also specialized in mentoring companies designing products.

1.2 Objectives

The requirements stage is essential in any type of project. Companies like 3DWays should outline major milestones and agree with the client on the most important characteristics of the final product. Numerous challenges encountered during consulting jobs are found to be caused by weak planning in an early stage. To mitigate these issues, companies demand a novice tool to first help clients discover and define their ideas, and secondly help them visualize and express those designs.

For product design, this implies offering a visual survey for insights collection, instead of traditional survey with limited and inaccurate description by words. The web application developed helps users express and define the product concept, just like it fosters the elaboration of new ideas about that concept. A human brain can only create based on what it has seen before (Greisdorf and O'Connor 2002). Therefore, showing additional content related to users' searches, via word and image recommender systems, promotes the generation of new ideas and boosts creativity.

The product of this thesis will play a crucial role during the design requirements sessions phase, marking the initial interaction between 3DWays and any client. The combination of AI driven content generation with recommendation systems allows for practical and thorough understanding of user needs. Moreover, the final output, a mood board compiling user choices, provides a better context to each client's vision and expectations. Building upon the foundational work of Arabella Specker, conducted in a directed research internship at Nova SBE and analysed in the literature review chapter, this

research is extended to consider more sophisticated models for each component of the tool (Specker 2022). Such models include pre-trained algorithms from tech giants Open AI and Meta, and purpose-built recommender systems.

1.3 Thesis structure

This report starts with a literature review (chapter 2) where current state of the art in AI-driven product design is studied, alongside the challenges businesses face integrating AI in their workflows, plus an analysis of the work done in the previous thesis.

The following four chapters (chapter 3 to 6) delve into the distinct components developed for the tool, discussing each part's role, design, and implementation in the context of the project's goals. Most in-depth analysis of the models used and how they were adapted for the project's needs can be found here. A similar structure was used for all individual parts: Introduction, Literature Review, Methodology, Results, and Conclusion. Each chapter focuses on topics from the four main research areas: word recommender systems, image recommender systems, image modification models, and web application integration.

Chapter 7 contains a breakdown of the three main databases considered for the development of this project. Chapter 8 covers insights regarding the generation of the final mood board. Chapter 9 encloses tips and guidelines on how users and future developers should handle the web application.

Lastly, chapters 10, 11 and 12 are the results, discussion (limitations and future recommendations) and conclusion of the project, respectively.

1.4 Project flowchart – User journey

Figure 1 is a summary of the web applications' capabilities. Each functionality's inputs and outputs were coordinated to fit the rest of the pipeline. This overview provides a

reference for the reader to contextualize the upcoming chapters of the report.

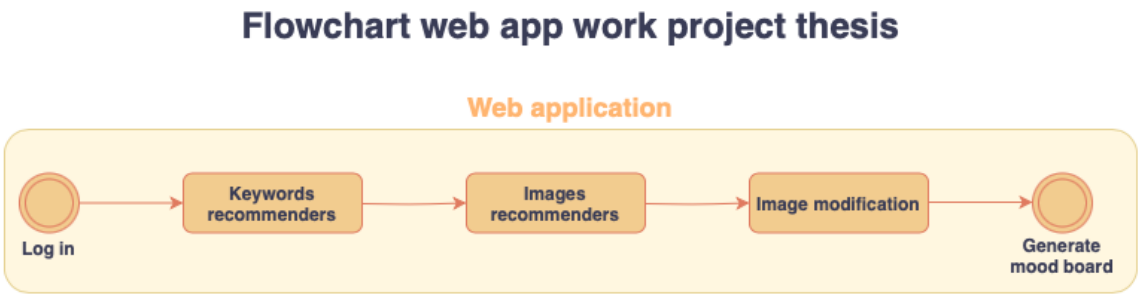

Figure 1 - Web Application flowchart

The chosen tool to merge all components was the library Streamlit, working as a web page generator for our tool. Users' first interaction with it is a login page, where a personal account will be created to store individual preferences and interactions (chapter 6.3.1). After entering the app, the first prompts required are keywords. These are essential for the next steps involving images or perfecting each user's keywords list. Two word-recommender systems will be used, one using Open AI GPT model (chapter 3), and another designed to check previously used related keywords (chapter 6.3.2).

With the selection of keywords completed, two image-recommender systems operate to present images to the user. The first model matches the inserted words to images from a database (chapter 4), while the second one tries to find patterns with previous images selected by other users (chapter 6.3.3).

The final section of the web application is dedicated to modifying images selected by users, to better align with their vision. This includes options for object segmentation and object removal (chapter 5).

The relevant outputs are finally combined in a real-time generated mood board (chapter 8). This mood board represents the automated solution that 3DWays sought to efficiently capture and understand client requirements.

2 Literature Review and current state of the art in AI-driven product design (group part)

2.1 AI in Conceptual Design and Product Development

In the evolving landscape of digitalization, AI has emerged as a central force in transforming the design and development of new products. The integration of AI in the design process represents a significant shift from traditional methods, preparing the way for more developed, efficient, and innovative design systems (Rosenthal and Niggemann 2022).

Engineers and designers today face a challenging task of creating products that cater to the dynamic needs of millions of consumers. This demands a balance between design, engineering, manufacturing, and craftsmanship, all while following specific guidelines and expectations. AI technology has become instrumental in this regard, offering invaluable support in capturing the right data and guiding product design and development processes. A McKinsey survey from November 2020 highlights that over half of the organizations have adopted AI in at least one function, with many reporting a significant increase in revenue due to AI adoption, especially in manufacturing (McKinsey 2020).

The transformation of AI's role in product development has become apparent in recent years. AI, which was less prevalent in the field, has become a fundamental component. Major technology companies such as Google, IBM, and Amazon have been instrumental in advancing its use in various aspects of product development, including engineering. This shift is largely due to AI's ability to streamline complex processes, speed up the time to market for products, and encourage innovation in product design (Salierno, Leonardi and Giacomo Cabri 2021).

In the realm of conceptual design, especially for products not yet present in the market, AI offers considerable potential. AI-driven tools empower designers to quickly visualize and refine new product concepts, playing a key role in the early phases of product

development. These tools seem to improve the efficiency of the design process while also providing new opportunities for creativity and innovation, enhancing the overall design experience.

2.2 Challenges in AI integration for businesses

The integration of AI has become more prevalent within businesses, however remains a challenge due to various reasons. Companies face multiple obstacles, including the need for technical expertise, resistance to organizational change, financial limitations, data management difficulties, and ethical and legal concerns. These issues, ranging from internal training gaps to regulatory compliance, impact every aspect of AI implementation in businesses. The first challenge is the demand for technical expertise. AI technologies are evolving rapidly, requiring a depth of technical knowledge that many businesses find challenging to acquire (Russell and Norvig 2016). The adaptation of generic AI solutions to meet specific business needs often necessitates a bespoke approach, tailoring algorithms, data sets, and models to align with operational difficulties (Fox 2023). This situation has been seen by Munich Re, where the company tackled the technical expertise gap through internal training and strategic external collaborations (Fox 2023).

Another critical challenge is the resistance to change within organizations. The integration of AI often disrupts established processes and roles, bringing out a range of responses from employees, from passive non-compliance to active resistance (Kotter 1996). Overcoming this resistance requires comprehensive change management strategies. These include extensive staff training which focuses on the operational aspects of AI while also focusing on the rationale and benefits of these changes. Requiring a culture that embraces change, backed by leadership commitment and open communication, is crucial in this regard (Lewin 1951; Fox 2023). Financial constraints represent a third challenge, particularly seen

in small to medium sized enterprises (SMEs). The costs associated with AI integration especially for the initial setup, ongoing maintenance, necessary data storage and hiring specialized employees can be substantial. Addressing these financial challenges requires exploring various funding options such as government grants or venture capital, alongside strategic planning to ensure a favourable return on investment.

Data availability and quality is another obstacle. Effective AI applications depend heavily on the availability of high-quality data. Many businesses struggle with gathering sufficient data and concurrently ensuring its relevance and cleanliness (Davenport 2018). Addressing data quality issues is critical, as failure to do so can compromise the integrity and effectiveness of AI systems. Lastly, ethical and legal considerations present a complex terrain for businesses implementing AI. Issues such as algorithmic bias, data privacy, and compliance with evolving regulations laws like the GDPR in the European Union are at the forefront of these challenges (Bostrom and Yudkowsky 2014; ESPU 2018). An example is UNESCO's effort to minimize gender bias in AI search engines, highlighting the need for ethical audits and the development of frameworks for responsible AI use (Fox 2023).

2.3 Current state of the art – Building on Specker's work

Last year's paper, by Arabella Specker (2022), made progress with regards to integrating AI into the design process, specifically through the development of a mood board tool. This tool, designed to retrieve images from Pinterest using image similarity recommendations, incorporated web scraping, natural language processing, and the VGG16 convolutional neural network. Specker's approach to evaluating the tool's performance involved both qualitative assessments and quantitative measures, like cosine similarity metrics.

Specker's work underscores the potential of AI in enhancing design processes for

industrial designers. The tool effectively provided relevant images for inspiration, allowing designers to create and store mood boards efficiently. However, the study also highlighted certain limitations. The tool's reliance on Pinterest as the sole image source restricted the diversity of available images. Moreover, while the average cosine similarity score indicated a moderate level of image relevancy, the potential need for more diverse and dissimilar images for creative inspiration was noted.

The thesis represented a significant step in using AI for industrial design, particularly in the creation and utilization of mood boards. This tool streamlined the initial stages of the design process by providing designers with a novel way to gather inspiration, thus facilitating a quicker transition from concept to development.

Despite these advancements, Specker's research also highlighted several limitations, which open up avenues for further development. The first notable aspect is that the tool lacks a user-friendly interface, which could be a barrier for designers unfamiliar with the underlying technology, making it less accessible and potentially reducing its effectiveness in the design process. Moreover, the tool's dependence on a limited set of input words, coupled with the inability to include additional descriptive terms like adjectives, limited the spectrum and diversity of accessible design inspirations. This limitation could impact the creativity and relevance of the mood boards produced. There was also an absence of a guidance system within the tool, leaving users without support in selecting the most effective keywords for their search. This could lead to less effective search results and potentially hinder the design inspiration process. Lastly, the tool did not allow users to modify or edit the images it presented. This lack of interactivity could limit designers' ability to tailor the mood board to specific design needs, potentially impacting the overall utility of the tool in the creative process.

Building upon Specker's foundational work, the current research focuses on

enhancing the AI-driven design tool by addressing these limitations. Efforts include developing a more user-friendly interface to improve accessibility and ease of use. Expanding the range of input options will allow for more nuanced and tailored image results. Implementing a guidance system will assist users in effectively utilizing the tool for design inspiration. Moreover, incorporating features for image modification and editing will provide users with greater control over their mood boards, enhancing the tool's applicability in the creative design process.

3 Database and dataset overview (group part)

3.1 Open Images dataset V7

The selection of an appropriate dataset for advancing artificial intelligence models is crucial especially in tasks like image generation, captioning, and object selection. Numerous features play a role in the selection process and in this section, we will outline the characteristics of the open image V7.

3.1.1 Characteristics

Size: The Open Images dataset is extensive, incorporating precisely 9 million images across a diverse array of categories and contexts. This substantial size not only ensures an ample supply of data for training machine learning models but also aligns seamlessly with the company's requirement for a sizable dataset.

Accessibility: The Open Images V7 is publicly available and hosted by Google. This availability ensures that researchers, developers, and educators can easily download and integrate into their projects.

Data Structure: Open Images V7 is structured in multiple components catering to varied computer vision challenges (Ultralytics 2020). This structured dataset includes essential components such as:

- Bounding Boxes: With Over 16 million boxes that demarcate objects across 600 categories. These bounding boxes are made-up squares that serve as a guideline point for object recognition and create a collision box for that element (Tasq.ai 2023)
- Segmentation masks: With exact boundaries of 2.8M objects across 350 classes. These segmentation masks define the outline of objects, which characterizes their spatial extent to a much higher level of detail (Krasin et al 2017).

Labels and metadata: Each image in the dataset is annotated with detailed labels, providing valuable information about the objects and scenes depicted. This rich labelling facilitates various computer vision tasks, such as object recognition and segmentation. Additionally, the dataset includes metadata that offers various details about the images, enhancing its usability for diverse applications in machine learning and research.

Applications: As highlighted by Ultralytics (2020), Open-image-dataset 7 is a cornerstone for training and evaluating state-of-the-art models in various computer vision tasks.

Flexibility: The Open Images Dataset offers considerable flexibility in sample selection. This adaptability proves invaluable for testing and refining machine learning models.

To sum up, these characteristics collectively affirm the suitability of the Open Images dataset for our project, which involves the generation of images, creation of captions, and object selection.

3.1.2 Implementation

The integration of FiftyOne in downloading the Open Images dataset offers a comprehensive solution for computer vision developers. FiftyOne, as a python library and web application provide an interactive tool for data exploration, model evaluation, and iterative development. Additionally, the seamless integration of FiftyOne with the Open Images dataset streamlines tasks such as debugging, error analysis, and collaborative exploration, making it an invaluable resource for teams engaged in advanced computer vision projects.

In conclusion, this seamless implementation confirms it as the optimal choice for the development of our web application.

3.1.3 Database structure

After the implementation of the Open Image Dataset, the structure will be explained in more detail below.

Table 3. Open Images dataset structure

Name	Open-image-v7-validation-10
Media Type	Image
Num samples	50
Persistent	False
Tags	[]
Same fields	id: ObjectIdField
	filepath: StringField
	tags: ListField
	metadata: EmbeddedDocumentField (ImageMetadata)
	positive labels: EmbeddedDocumentField (Classifications)
	negative labels: EmbeddedDocumentField (Classifications)
	detections: EmbeddedDocumentField (Detections)
	segmentations: EmbeddedDocumentField (Detections)
	relationships: EmbeddedDocumentField (Detections)

Id (ObjectIdField): is a unique identifier for each sample in the dataset. This ObjectIdField is an auto-generated identifier that makes us differentiate between samples, allowing easy referencing during analysis and manipulation.

Filepath (StringField): contains the path to the image file associated with each sample. This information is important, especially when unable to access or load a special image for further processing.

Tags (List Field (StringField)): The tags field is a list that contains user-defined

tags associated with each sample.

Metadata (EmbeddedDocumentField(ImageMetadata)): The metadata field is an embedded document containing metadata associated with the image. Image metadata include information such as Licence, Author, Original size, OriginalLandingURL, and any other relevant details that provide context about the image's origin and characteristics.

Positive_labels (EmbeddedDocumentField(Classifications)): The positive labels field is an embedded document that contains information about labels associated with the image. In the context of the Open Images dataset, it includes Label Name, Confidence, logits, and imageID.

Negative_labels (EmbeddedDocumentField(Classifications)): It is similar to positive_labels, However, the negative_labels field is an embedded document that contains information about negative classifications or labels. Negative labels represent the absence of certain labels in a specific image. It also includes: LabelName, Confidence,logits and imageID.

Detections (EmbeddedDocumentField(Detections)): The detections field is an embedded document that includes information about object detections within the image. In the Open image Dataset, it includes: ImageID, label,bounding_box, mask,index,Confidence, IsOccluded,IsTruncated, IsGroupOf,IsDepiction,IsInside

Segmentations (EmbeddedDocumentField(Detections)): The segmentations field, like detections, is an embedded document that contains information related to image segmentation like Mask Path, ImageID, attributes, label,bounding_box, confidence, index

relationships (EmbeddedDocumentField(Detections)): The relationships field is an embedded document that contains information about spatial or semantic relationships between objects within the image. This includes relationships like Imaged, Label1, Label2, bouding_box, index, confidence, label, mask, attributes, label, and tags.

To sum up, the Open Images Dataset's sample fields offer a wealth of information about every image, including file specifics, tags, metadata, object labels, detections, segmentations, and relationships. These fields can be used by academics and industry professionals to carry out different types of analysis, develop machine learning models, and acquire an understanding of the context and content of the images in the dataset.

3.2 Flickr 8K - Image Captioning dataset

3.2.1 Criteria for Dataset Selection in Image Captioning Research

The selection of an appropriate dataset is a critical step in the advancement of image captioning models within the field of artificial intelligence (Paullada et al. 2021). It forms the backbone for both training and evaluating the performance and robustness of these models. The criteria set forth for this selection are rooted in ensuring that the dataset not only serves the immediate requirements of model training and evaluation but also aligns with the broader research objectives and ethical standards of the field (Fenza et al. 2021).

Size and Diversity: A dataset of adequate size is fundamental to confine a wide range of image contexts and scenarios, facilitating comprehensive model training and testing. Diversity in imagery, including variations in scenes, objects, and activities, is crucial to assess the model's versatility and ability to generalize across different visual contexts.

Quality and Relevance of Captions: The chosen dataset must contain captions that are accurate in describing the image and also being relevant to the application of the model. High-quality captions are essential for using models and evaluating pre-trained models that can generate precise and contextually appropriate descriptions.

Balance and Representation: Ensuring a balanced representation of various subjects and themes within the dataset is vital to avoid biases in model training. A well-balanced dataset contributes to the development of a model that is fair and effective across diverse design contexts.

Accessibility and Ethical Considerations: The dataset's accessibility for academic research and compliance with ethical standards, including the appropriate licensing of images and captions, is a critical aspect of the selection process. This ensures the responsible use of data in line with academic and research integrity.

These criteria were instrumental in guiding the selection of a dataset that not only meets the technical requirements of the image captioning functionality but also aligns with the ethical and practical considerations of this thesis. In evaluating these criteria against available datasets, the Flickr 8K dataset published on Kaggle by Aditya (2020) appeared as a particularly suitable dataset. The following section will provide an overview of the Flickr 8K dataset, detailing its structure and fields, and discuss how it meets the outlined criteria.

3.2.2 Overview

The Flickr 8K dataset is widely recognized in the field of image captioning research for its suitability in developing and evaluating image captioning models, matching well with the established criteria. Flickr 8K is characterized by its manageable size, containing a total of 8,092 images. This moderate scale makes the dataset particularly suited for research scenarios where computational resources are limited, such as training models on low-end laptops or desktops. Additionally, the dataset is freely available, which aligns with the accessibility criterion, making it a popular choice for academic and research purposes.

Labeling and Data Structure: Each image in the Flickr 8K dataset has five distinct captions, providing textual context for each visual scene. This multiple-caption structure enhances the dataset's utility for training and evaluating image captioning models, as it offers diverse interpretations of the same visual content. The dataset is divided into specific subsets, with 6,000 images designated for training, 1,000 for testing, and 1,000 for development purposes. This clear segmentation facilitates the systematic training, testing, and validation of models.

The Flickr 8K dataset comes with the necessary textual files that describe the training and test sets. The Flickr8k.token.txt file, in particular, contains a comprehensive list of the 40,460 captions, offering a pool of data for model training and testing.

3.2.3 Significance and broader application of the dataset

Given its structure and features, the Flickr8K dataset fulfils the criteria set for selecting an image captioning dataset. The dataset's size and diversity make it an ideal choice for comprehensive model training without necessitating extensive computational resources. The quality and relevance of the captions in the dataset are conducive to developing models capable of generating accurate and contextually appropriate descriptions.

The multiple caption per image structure of Flickr 8K allows for a more nuanced evaluation of the model's performance, as the diversity in captions can offer insights into how well the model captures various aspects of an image. Furthermore, the straightforward and well-organized structure of the dataset simplifies the processes of data pre-processing and model training, making it a practical choice for research.

In the broader context of this thesis, the Flickr 8K dataset provides a solid foundation for cross-comparison and collaborative analysis. Its use in different individual components of the project, such as training distinct models or evaluating various image captioning techniques, enables a cohesive and comprehensive research approach. The dataset's versatility and alignment with the project's objectives underscore its significance as a valuable resource within this collective research.

In summary, the Flickr 8K dataset, with its manageable size, well-labelled data, and accessible format, presents itself as an optimal choice for our image captioning research. Its selection is justified not only by its technical and structural features but also by its alignment with the broader goals and ethical considerations of this thesis.

4 Mood board PDF (group part)

Mood boards have long been an essential tool in the creative process, serving as a visual representation of ideas, themes, and inspirations. Their origins can be traced back to the early 20th century, where they were used in fashion and interior design. Initially, these boards were physical collages of images, text, and samples of objects in a composition. Over time, their use has expanded across various fields including graphic design, film production, and even in digital media planning, illustrating their versatility and adaptability to different creative needs.

By their very nature, mood boards are designed to evoke a certain style or mood, providing a tangible way for designers to communicate their concepts and visions. They act as a springboard for creativity, allowing for a free-flowing exchange of ideas and serving as a reference point throughout the design process (Rieuf, Bouchard and Aoussat 2015).

In the context of this project, the mood board takes on a digital form, evolving from its traditional physical manifestation. The digital mood board, particularly in the form of a PDF, becomes a crucial component in the design cycle. It offers a unique way to capture and present the essence of a user's design journey, encapsulating their selections, preferences, and inspirations. For example, in the realm of web design, digital mood boards can be used to visually communicate the look and feel of a website before any code is written. Similarly, in fashion, they can collate various fabric textures, colour palettes, and accessory choices to form a coherent style vision. This digital format not only aligns with the modern, tech-driven approach of the design tool but also offers greater accessibility and ease of sharing, crucial in collaborative design environments (Chipambwa and Chikwanya 2022).

4.1 Requirements

The mood board PDF, as conceptualized within the scope of the project by 3DWays, was designed with specific requirements to ensure both functionality and aesthetic appeal.

Central to these requirements was the stipulation that the mood board be confined to a single page. This single-page constraint is pivotal as it compels the user to refine their creative vision into a concise, yet powerful visual narrative. The challenge is to focus on what is essential and most expressive of their project's theme. This limitation not only enhances the clarity and impact of the mood board but also aligns with the principles of simplicity and succinctness in design.

In terms of image selection, the guidelines set a minimum of five and a maximum of ten images. This range was carefully chosen to allow for sufficient flexibility in showcasing a variety of elements while maintaining a clear and uncluttered layout. The decision to limit the mood board to six images was made to optimize the balance between visual diversity and coherence. This number ensures that each image can be presented with enough detail to be meaningful, yet not so many that they overwhelm the viewer or dilute the mood board's overall impact.

Regarding the size of the images, while 3DWays did not specify exact dimensions, the requirement was that they should be clearly visible. To meet this criterion, a decision was made to standardize the images at 50 pixels by 50 pixels. This uniformity in size aids in creating a harmonious and aesthetically pleasing grid layout, ensuring that each image is visible and contributes equally to the mood board's narrative. The choice of image size and the decision to go with six images were made to ensure that the final mood board is not only visually appealing but also practical in terms of visibility and comprehension.

These carefully considered requirements and decisions reflect a deep understanding of design principles and user experience. By setting these parameters, 3DWays aims to guide users in creating mood boards that are not only visually striking but also effectively communicate the essence of their design projects. The process of adhering to these

constraints is an exercise in creativity and critical thinking, pushing users to select images that are most representative of their ideas and that work together to tell a cohesive story.

In summary, the mood board PDF's design, dictated by the requirements set by 3DWays and the subsequent decisions made within the project's scope, is a testament to the careful consideration of how best to visually communicate design concepts. The single-page format, the specific number and size of images, and the overall layout are all elements that come together to create a tool that is both useful and inspiring for users, aiding them in bringing their creative visions to life.

4.2 Implementation

The implementation of the Mood board PDF within the project was a nuanced process, heavily reliant on the capabilities of the PyFPDF library, a Python tool known for its efficacy in PDF generation (PyFPDF 2023). The integration of this library was a critical step in the final stage of the dashboard process. As users navigated through the dashboard, selecting images that resonated with their design vision along with the generated captions, these selections were systematically stored in our database. This data storage was crucial for the creation of the mood board PDF.

A unique aspect of the implementation was the handling of image modifications. When users altered an image, the modified version, rather than the original, was designated for inclusion in the mood board. This approach ensured that the mood board accurately reflected the user's creative input and final vision. Additionally, for these modified images, the keywords associated with the images were utilized instead of traditional captions. This decision was driven by the workflow of the application, as outlined in the Streamlit part of the thesis, where image captions are generated prior to any modifications. Consequently, the modified images lacked specific captions, necessitating the use of keywords as a descriptive alternative.

The challenge of implementing the Mood board PDF was further compounded by PyFPDF's reliance on absolute positioning. This required precise placement of every image and text element at specific coordinates within the PDF file that is being created. To address this, a process was developed to standardize the dimensions of each image to 50 pixels by 50 pixels. This standardization was crucial in maintaining a uniform and aesthetically pleasing layout, accommodating the diverse range of images selected by users.

Given the constraint of fitting at least six images onto a single page, alongside their respective keywords, the layout design required careful planning and precision. A dynamic template was devised that could adjust to the varying images and text, ensuring each element was given adequate space and prominence. This template was not only visually appealing but also enhanced the readability and impact of the keywords, which were pulled directly from the database entries.

In conclusion, the Mood board is a fusion of traditional design principles with contemporary digital functionalities. Defined by the requirements set by 3DWays, it effectively encapsulates a user's design process within a single, coherent page. The tool adheres to specific requirements regarding the number and size of images, ensuring a focused narrative that accurately reflects the user's creative process. The inclusion of modified images and associated keywords adds a layer of personalization to the mood board, making it a true reflection of the user's vision and modifications.

The implementation, utilizing the PyFPDF library, navigated the technical challenges of absolute positioning and image standardization. This approach resulted in a layout that balances functionality with the need to convey a clear narrative. It offers users a distinctive tool to visualize and share their design inspirations, bridging the gap between concept and presentation.

5 User Guidelines (group part)

Effective navigation and utilization of a web application are essential, particularly in the context of design and creative ideation. This section is dedicated to providing users with clear and detailed guidelines for interacting with our web application, a tool specifically crafted for design inspiration. These guidelines serve as a practical framework, enabling users to efficiently leverage the application's capabilities to transform their design ideas into visual representations. The application offers a range of features that assist in assembling a digital mood board, an integral component in the visualization of design concepts. By adhering to these guidelines, users will be able to proficiently use the application, facilitating the creation of a mood board that effectively summarizes their design inspirations and ideas. This approach ensures that users can fully engage with the tool's functionalities, making the most of its potential to aid in the design process.

The initial step in accessing the inspiration tool involves account creation, a prerequisite for using the application. This process requires users to set up a username and a password, with the password needing to be a minimum of five characters for security purposes. It is imperative for users to remember these credentials, as they are essential for subsequent access to the application.

Access to the login page is provided on the left side of the computer screen. Within the *Navigation* section, a *Go To* option is available, accompanied by a dropdown button. This button presents two options: *Sign Up* and *Log In*. Selecting *Log In* directs users to the appropriate page for entering their login details.

Users should be aware that the first entry into the application might involve an extended loading time. This delay is attributed to the initialization phase, where necessary libraries are loaded to ensure the proper functioning of the application. This setup is a one-

time process, vital for the activation and readiness of all features and tools within the application.

Upon reaching the main page, users are presented with various interactive buttons. The application requires an input of five keywords to begin. While these buttons can be explored in any order, it is recommended to start with the *Generate Suggestions* button for an efficient experience. Subsequent to this step, users are encouraged to interact with other features, further enhancing their engagement with the application.

In the application's interface, users are prompted to input keywords in a designated text box, located above the *Generate Suggestions* button. The interface instructs users to *'Enter at least 5 keywords (separated by spaces):'*. It is crucial for users to note that the application's programming requires these keywords to be separated by spaces for accurate processing. In cases where users intend to input multi-word phrases, such as *'Minimalist Camera'*, the application interprets each word separately. To ensure the phrase is treated as a single entity, users should employ an underscore, inputting it as *'Minimalist_Camera'*. This notation is essential for maintaining the integrity of compound terms, including colour-specific items like *'black_bottle'*, to ensure the application accurately interprets the user's intent.

Upon pressing the *Generate Suggestions* button, two additional buttons become accessible. These buttons are designed to provide users with expanded suggestions of keywords, fostering creativity and offering new design ideas. Should users wish to incorporate these new keywords into their search, they must enter them into the text box above the *Generate Suggestions* button. Furthermore, the *Meaning of the word* button requires users to input a word into the corresponding text box to retrieve its definition. In instances where the application does not recognize a word, it will display a message: *'No definition found for word'*. Users seeking further clarification may resort to external sources

such as Google for definitions. Additionally, the *Generate Images* feature presents each image alongside a *Select/Drop this image* button. This functionality allows users to save images they find appealing for further modification. If a user decides against an image, they can deselect it by pressing the button again. Selected images are displayed on the website below the image gallery. Similar to the image selection process, pressing the *More Images* button will display additional images, following the same logic for selection and deselection.

The operational guidelines for the *Modify images selected* button and its associated sub-buttons are outlined as follows:

Image Selection for Modification: Upon clicking the buttons below, a pop-up window will display the selected image. Users are prompted to select points on the image to delineate the object they wish to isolate. Optionally, they can also define a bounding box to specify the area containing the object.

Defining Object Position: Users can interact with the image using mouse clicks to specify areas of interest. A left mouse click is used to mark positive input points – areas to be included in the object isolation. Conversely, a right mouse click is used to mark negative input points – areas the model should ignore. To exit the pop-up window and finalize selections, users should press the *q* key.

Bounding Box Specification: To define a bounding box around the object, users should select two points with a left mouse click. The first click establishes the top-left corner of the box, while the second click sets the bottom-right corner. The pop-up window will automatically close after the second point is selected.

Post-modification, each edited image is saved under a unique filename. Users have the flexibility to edit an image multiple times, with each version being saved. However, it is important to note that only the last modification of an image is retained in the database. Consequently, the application tracks only the final version of each modified image for each

user. Should users desire multiple modifications of the same image, they are advised to repeat the entire modification process from the beginning for each desired version.

Furthermore, it is important to note that modified images are not immediately visible within the application. The modifications made by the user will only be displayed at the conclusion of their interaction with the inspiration tool. Therefore, users are encouraged to carefully follow the given instructions during the image modification process to ensure their desired outcomes are achieved.

In conclusion, adhering to these user guidelines is instrumental in maximizing the effectiveness of the web application. By following the outlined steps and recommendations, users can ensure a seamless and productive experience. The guidelines are designed to empower users, enabling them to fully engage with the application's capabilities and to express their creative visions effectively.

6 Results (group part)

After the integration of all functionalities, users are now capable of generating a mood board. All previous Result sections provide insights for the complete tool. To complement them, this chapter focus on presenting the final output. This personalized creation is a culmination of the choices users make throughout their journey on the web application.

At the end of the design process, users can access the "Generate Mood Board" feature, in order to create a cohesive display of the selected images. These images are indicated with a message showing their respective identification numbers (Figure 31). Additionally, the user is prompted with a button to "*Start Mood Generation*" and a hyperlink to download the PDF locally.

This process starts with a filter applied to the 'data' database, retrieving all saved history of the specific user in the web application. Another selection process is needed to choose the final six images which will be compiled in the PDF file. Finally, each image is plotted, supplemented by the descriptive caption generated by the system previously described, to provide additional context to the mood board. These captions are only available to images coming from the image database, as only them are inserted in the caption model. Images suggested at a later stage by the recommender system based on previous interactions are stored differently in the database. Therefore, only the keyword label is presented in the mood board.

An example of a mood board is available in Figure 33 in the appendix.

7 Discussion (group part)

Solution impact 3Dways

Previously to the solution, the process behind product design was archaic, filled with meetings and complex conversations on the topic, without a clear visualisation or sometimes even an idea of how the product should look. Information also needed to be recovered during development as it travelled from mouth to mouth. With the implementation of the application, the process is expected to be cut short. With the help of computation, ideas will flow naturally to the user with visual examples that incentivise the creative procedure. The aid of visual examples looks to maintain the process as true to the idealisation of the client.

7.1 Limitation

While the application holds promise for various advantages, it is crucial to recognise and rectify specific limitations. Acknowledging and addressing these constraints is imperative, as they can impact the application's overall functionality and user satisfaction.

The first limitation is the inheritance of biases in the training data of the models within the image captioning and keyword suggestion functionalities. These biases in the training data could affect the accuracy and neutrality of their outputs. Moreover, the functionality of the keyword suggestion generator is heavily reliant on the continuous availability of the OpenAI API. In instances where the API experiences downtime or is not accessible, the tool's ability to generate keyword suggestions is significantly hindered. This dependency poses a risk, as any interruption in the API service directly impacts the user experience and the overall reliability of the tool. An alternative perspective highlights that the application includes an independent keyword suggestion feature based on user history, functioning without reliance on third-party services such as OpenAI. The effectiveness of the keyword suggestion generator relies on the initial user input, which may limit its utility if users provide vague keywords that the AI cannot give good keyword suggestions in

return.

Besides this, it is often not easy to objectively evaluate the output and relevance of the image captions. Some images can be interpreted differently by different people. An example: Consider an image showcasing a rainy street scene with people carrying umbrellas. While the AI model might generate a caption like "People walking on a wet street with umbrellas," different viewers might interpret the scene differently. One viewer might see it as depicting gloomy, rainy weather affecting city life, while another might interpret it as a cosy, atmospheric scene typical of urban life. This variation in human perception highlights the challenge of ensuring that AI-generated captions align with every user's subjective interpretation and the image's intended emotional or thematic context.

In the development of the CLIP model, due to low computational power, the sample size was restricted in order to manage challenges posed by time and memory constraints. As a consequence of the small sample size, a more restricted collection of images limited the model's ability to generalise results in some cases. The similarity scores are low due to the quality of the dataset. It did not abide the case that CLIP's zero-shot classifiers can be sensitive to wording or phrasing. Another limitation is that the model is unable to count the number of objects in an image or classify their distance, making it unreliable in images with a large number of objects.

Meanwhile, for the image modification tool, the implication was that the performance depends on the quality of the input images. Lower resolution or poorly detailed images may not yield optimal results. Inserting an image in the image encoder takes around 1.5 to 2 minutes, using the processing power available in the development phase. This could create a considerable wait time for the user but also enable real-time processing capabilities once all images are processed due to the rest of the system being lightweight.

Finally, some bugs were detected within the segmentation process while integrating

the image segmentation code with the Windows operating system. These were only found in the final stage, as the development process was performed in MacOS. This is something to consider when deploying the solution to the client. With limited time on this deliverable, the web application needs clear improvement. Users have to follow the specific steps referenced in the guidelines for the application to work; otherwise, it can crash, causing the program to restart. Modifying images is a time-consuming effort, attributed to the prolonged processing time for the set-up code of the segmentation procedure.

7.2 Recommendations for future work

A general advice for future improvements includes a fine tune of the employed AI models to combat biases and overfitting that might occur. With increasing utilisation of the tool, new user information (text and image) will be created, which can be used to better test the models.

A possible solution for the last limitation mentioned is modifying the web application button that feeds selected images into the SAM model to consider only images the user intends to modify, rather than all images selected for the mood board. Such modification could significantly reduce wait times when the images requiring modification are fewer than those included in the mood board.

When developing the SAM project, Meta created a pre-trained image dataset, SA-1B, including more than one billion masks over 11 million images. If it fulfils all the image requirements, this dataset can be considered to accelerate the segmentation process. In this project, the decision was made to prioritise flexibility by enabling the use of any image database and allowing users to create custom segmentations through points and box prompt inputs rather than limiting the scope to predefined parameters.

By implementing an AI-powered tool to analyse project data and generate a mood board based on shared elements and themes, the tool could effectively capture the project's

essence and provide a unified reference point for all involved parties. This would eliminate the need for individual users to select images and would ensure that the mood board accurately reflects the collective vision of the project team. Moreover, adopting a project-level approach would foster collaboration and promote alignment among different entities within the project. The improved mood board would facilitate better communication and understanding of the project's overall aesthetic direction by providing a shared visual reference point. Overall, transitioning from a user-level to a project-level mood board would significantly advance the tool's capabilities. It would enhance collaboration, streamline communication, and ensure the mood board accurately reflects the project's unified visual identity.

Regarding the application, the interface could be improved from the customer's perspective, making it more appealing and user friendly. Another addition should take into account multiple users for a project, creating a specific ID that connects every subject in a project.

8 Conclusion (group part)

The final product of this thesis accomplishes the business objectives defined in the introduction section. Still subject to testing from 3DWays, the tool improves the comprehension of clients' ideas and refines the requirements for a typical 3DWays project. Given the short development period, there were many limitations and recommendations identified. Nonetheless, we are confident that this is a strong foundation to work on and improve to create a final deliverable tool, as evidenced by the results section.

The platform is compatible with different image databases, contributing to the versatility of the web application. Different variations can be created for specific product design types, each featuring different images available and perhaps additional functionalities.

Overall, the potential and scope of a tool such as this one is substantial, providing an innovative approach to the initial stages of project management, a methodology that can be adapted to other businesses as well. The Python code underpinning the tool holds potential beyond its current application. Each implemented functionality can be tailored to serve different purposes, showcasing the tool's flexibility. Furthermore, with the rapid development and release of AI models, newer, more efficient, and effective versions of the algorithms used can be integrated into this tool to keep it up to date.

9 Enhancing Design Inspiration through AI-Driven Keyword suggestions and Image Captioning (Frank Tamer)

9.1 Introduction

9.1.1 Background

The beneficial functions of image captioning and keyword suggestion generators in

artificial intelligence significantly impact different aspects of design processes in different areas. In the realm of image captioning, the capacity to automatically generate descriptive text for images proves transformative. For instance, in graphic design, designers leverage image captioning to swiftly create alt text for website images, enhancing accessibility for users with visual impairments and streamlining the design workflow (Vinyals et al. 2015). Simultaneously, keyword suggestion generators play a crucial role in guiding designers toward effective content tagging. In web design, these generators assist designers in identifying relevant keywords for metadata and content labels. This streamlines content organization while also optimizing the website's search engine visibility, contributing to a more seamless user experience (Hulth 2003).

The evolution of image captioning, driven by advancements in Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has further amplified the impact on design processes. In augmented reality (AR) design, automatic image captioning facilitates the creation of immersive experiences by providing real-time and contextually rich descriptions of AR elements. This application enhances the user experience and accelerates the design iteration process (Vinyals et al. 2015; Devlin et al. 2018). In the context of keyword suggestion generators, their integration into a design inspiration tool can be exemplified by a scenario where a graphic designer receives suggestions for relevant keywords as they work on a project. This assists the designer in maintaining consistency and coherence throughout the design, ultimately contributing to a more polished and professional outcome (Hulth 2003).

9.1.2 Objectives

The primary objectives of this chapter are twofold:

1. To develop and evaluate a keyword suggestion generator utilizing prompt engineering techniques, aiming to assist designers in generating a diverse and

contextually relevant set of keywords during the design process. The focus is on enhancing the creative process by providing designers with effective keyword suggestions.

2. To implement and assess image captioning using pre-trained models, to generate informative and engaging descriptions for selected images. These descriptions will serve the dual purpose of improving the machine's comprehension of a user's design preferences and enhancing the final mood board, which plays a crucial role in the design cycle.

The proposed objectives contribute to the development of a comprehensive design inspiration tool by addressing the critical aspects of keyword suggestion generation and image captioning. The enhanced keyword suggestions provided by the generator will facilitate ideation and concept refinement while the informative image caption will enrich the design process and enhance mood boards.

9.1.3 Structure

This chapter is organized to facilitate a coherent exploration of two core elements: image captioning and keyword suggestion generators. The exploration begins with a comprehensive literature review and related work, establishing the foundation by examining the current state of physical product design, keyword suggestion implementations, the field of image captioning, and the importance of machine learning models. Following this, the Methodology outlines the methods employed for the development of the keyword suggestion generator and the implementation of image captioning. In the results, the findings of this research are presented, while the Conclusion and Discussion summarize key findings and their implications.

9.2 Literature review and related work

9.2.1 Current state of physical product design

This research builds upon the existing state of a physical product for design inspiration, which explored the implementation of machine learning tools for design inspiration as conducted by Specker (2022). However, critical functionalities that are currently lacking in the existing product have been identified, specifically in the generation of images based on user input and the provision of meaningful descriptions for these images. The primary limitation lies in the reliance on user-provided keywords. Users are expected to enter keywords they believe are relevant to their project, however this can be a challenging task, especially when contextual guidance is needed. Users may struggle to come up with the most appropriate and creative keywords, hindering the design process.

To address these limitations, this research leverages the capabilities of artificial intelligence, particularly machine learning models like *GPT-3.5*. The proposed solution involves understanding the context of user input words and providing intelligent suggestions that the user may not think of when using the product. This will aid designers in generating a more diverse and contextually relevant set of keywords, thereby enhancing the creative process (Filippi 2023).

Another aspect of design inspiration tools that demands attention is the absence of meaningful descriptions or captions for generated images. While this may not be problematic for simple images, it becomes increasingly important when considering the creation of a final product (Kulkarni et al. 2013). Users benefit from having a textual description that encapsulates the essence of an image, providing additional context and understanding.

9.2.2 Keyword suggestion with AI

Keyword suggestion tools have evolved from simple statistical models to

sophisticated AI-driven systems. Early tools relied on frequency counts; the integration of machine learning has enabled the analysis of large datasets, leading to predictions of contextually relevant and targeted keywords (Wang 2022). Natural Language Processing (NLP) has further refined keyword suggestions, allowing for an understanding of language that encompasses syntax and context, moving beyond keyword density to sentiment and thematic content (Manning et al. 2014). This has been particularly enhanced by the development of deep learning models that can process and generate language-based predictions.

Prompt engineering with pre-trained language models like GPT has become a cornerstone for generating relevant keyword suggestions. By crafting effective prompts, these models can produce creative and contextually appropriate outputs. Studies have shown that transformer-based models improve the relevance of keyword suggestions for various applications, including digital marketing and content creation (Brown et al. 2020). However, challenges persist, particularly in ensuring the balance between relevance and creativity and addressing inherent biases within AI systems (Bolukbasi et al. 2016). Despite these challenges, the trajectory of keyword suggestion tools is towards more personalized and adaptive AI that can leverage individual user data for enhanced suggestion accuracy.

9.2.3 Image Captioning

Image captioning has become a critical intersection of computer vision and natural language processing, aiming to generate descriptive text for images automatically. The field has rapidly advanced with the introduction of deep learning techniques, particularly the use of Convolutional Neural Networks (CNNs) for image recognition and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks for generating descriptive text (Vinyals et al. 2015). The synergy of CNNs and RNNs has led to the development of models which recognizes elements within an image while also

understanding the context and relationships between these elements to generate coherent and relevant captions (Karpathy and Fei-Fei 2015). This dual approach has been pivotal in creating systems that can approach human-level descriptions of visual content.

Recent studies have focused on improving the accuracy and richness of the generated captions. Attention mechanisms, for instance, allow models to focus on specific parts of an image when generating text, leading to more detailed and accurate descriptions (Xu et al. 2015). Additionally, the integration of semantic information and object recognition has further enhanced the descriptive capabilities of captioning models (You et al. 2016). Despite progress, generating captions that are both accurate and contextually rich remains a challenge. The complexity of natural language and the subtleties of visual cues often mean that captions may lack nuance or fail to capture less obvious aspects of images. Moreover, the training data used for these models can influence the output, sometimes leading to biased or inaccurate descriptions if the data is not diverse or comprehensive enough. In summary, image captioning models have shown significant promise in generating descriptive text for images, with applications ranging from aiding visually impaired users to enhancing content discoverability. The ongoing research aims to refine these models for greater accuracy and contextual relevance, ensuring that the captions are as informative and inclusive as possible.

9.2.4 Scientific relevance

The scientific relevance of this research lies in its contribution to the fields of machine learning and design. By advancing keyword suggestion and image captioning tools, this work intersects with significant areas of AI research, including natural language processing, computer vision, and human-computer interaction. The creation of a keyword suggestion tool, which leverages prompt engineering in AI models such as GPT, marks an innovative use of language models in the realm of design. This could lay the foundation for future AI-driven creative tools. In parallel, the improvement of image captioning systems

through the use of pre-trained models is a step towards attaining a deeper AI comprehension of visual elements. This research not only has the potential to streamline the design process for practitioners, as well as contributing to the academic discourse by addressing the challenges of contextuality and creativity in AI-generated content. The findings may offer insights into how AI can better understand and assist with human creative tasks, a core interest in AI research.

9.3 Methodology

9.3.1 Keyword suggestion generator

The development of the keyword suggestion generator is a pivotal component of this thesis, aiming to enhance the creative process for designers by providing them with a diverse set of contextually relevant keywords. Prompt engineering is the first critical step in the interaction with language models. It involves crafting a prompt that effectively guides AI to produce the desired output (Miller 2023). Through iterative testing, it was determined that a prompt that is concise yet direct yields the most relevant suggestions. The chosen prompt format was: "Please suggest 4 alternative or similar keywords for each of the following keywords: user inputs". This format was found to strike a balance between brevity and specificity, leading to high-quality keyword suggestions.

The OpenAI API, particularly the text-davinci-003 engine, was utilized for this project. This engine, part of the GPT-3.5 Turbo model, was selected for its balance of performance and cost-efficiency (OpenAI Platform 2023). The API request is configured to return a raw string of suggestions, with the max tokens' parameter set to 200 to constrain the response length and manage API usage costs. The default temperature parameter of 1.0 was initially retained to maintain a balance between randomness and predictability in the suggestions. Future iterations may explore the adjustment of this parameter to evaluate its

impact on the diversity and creativity of the keyword suggestions.

The Python implementation for generating keyword suggestions is straightforward. A user inputs keywords, which are then passed to a function that returns keyword suggestions. This function constructs the API prompt and handles the response. The response is processed to strip newline characters and split the text into individual keyword suggestions, which are then deduplicated and formatted for presentation to the user. To aid understanding, especially for less common keywords, another function was implemented to provide dictionary definitions. Utilizing the wordnet library from the Natural Language Toolkit (NLTK), the function retrieves the definition of a given word (Bird, Klein, and Loper 2009). This feature enhances the user experience by allowing designers to quickly understand the context of the suggested keywords.

9.3.2 Pre-trained models for image captioning

In this subsection of the thesis, the focus centres on leveraging pre-trained models for the task of image captioning. Pre-trained models are machine learning models that have been previously trained on a large dataset and are available for use with or without further fine-tuning. The use of pre-trained models offers significant advantages, especially in terms of resource efficiency and model robustness (Han et al. 2021). Training a model from scratch requires extensive computational resources, a large and diverse dataset, and significant time for training to achieve high accuracy. Pre-trained models, on the other hand, offer a head start as they have already been trained on extensive datasets (Choi and Choi 2022). This approach saves resources while also leveraging the model's pre-existing knowledge, often leading to better performance, especially in tasks like image captioning where understanding context and nuances is crucial.

The decision to utilize open-source pre-trained models is driven by their accessibility, community support, and the continuous improvements they undergo. Open-

source models often represent the cutting edge in their field due to contributions from a diverse set of researchers and developers (Han et al. 2023). Additionally, using these models avoids the need for the extensive infrastructure required to train large-scale models.

ViT GPT-2 Image Captioning

The first pre-trained model that has been utilized in this project is the ViT GPT-2 model. This image captioning model leverages the complementary functionalities of two AI models, Vision Transformer (ViT) and GPT-2, to generate detailed and contextually relevant captions for images. ViT excels in extracting meaningful features from images, while GPT-2 excels in generating natural and grammatically correct captions (NLP Connect 2022). ViT's ability to directly process images as matrices of pixels, unlike traditional convolutional neural networks (CNNs) that analyse images as sequences of patches, plays a crucial role in capturing long-range dependencies within images. This enables ViT to grasp the overall structure and context of an image. GPT-2's language generation capabilities are used to translate ViT's extracted image features into descriptive captions. GPT-2's ability to process language at a granular level, understanding syntax, semantics, and context, ensures that generated captions are descriptive, grammatically correct and lastly contextually relevant.

The overall implementation of the ViT GPT-2 image captioning model involves an integration of ViT and GPT-2. After the image undergoes initial preprocessing to convert it into a format compatible with ViT, the pre-processed image is fed into the ViT encoder to extract a comprehensive representation of its visual features. These extracted image features are then passed through a cross-modal attention mechanism, enabling them to interact with the language representations generated by GPT-2. This interaction facilitates the exchange of information between the visual and linguistic domains, allowing the model to refine the generated captions and ensure they accurately capture the essence of the image. Finally, the GPT-2 caption generation module receives the fused visual and linguistic information and

produces a descriptive, accurate, and evocative caption based on its understanding of the image and the provided text prompt. The generated caption undergoes post-processing steps to further enhance its clarity and accuracy, removing extraneous punctuation or correcting grammatical errors.

BLIP Image Captioning

The second image captioning model is BLIP, which works similarly as the previous model, differentiating in certain areas. The BLIP architecture combines a unidirectional encoder for processing images with a bidirectional encoder for processing text. This approach enables BLIP to effectively capture the context of both the image and the text, leading to more informative and context-aware captions. The unidirectional image encoder enables BLIP to effectively process images as sequential data, while the bidirectional text encoder allows it to capture the temporal relationships between words in a sentence (Li et al. 2022). Similarly, to ViT GPT2, this model also utilizes a cross-modal attention mechanism to exchange information between the image features extracted by the unidirectional image encoder and the language representations generated by the Transformer-based decoder. This interaction allows BLIP to refine the captions, ensuring they accurately capture the essence of the image.

The implementation of the BLIP image captioning model involves a sequential process. Initially the image undergoes preprocessing to convert it into a format compatible with BLIP. By converting an image to tensor format and resizing, the image becomes available for use within the model. The image is then fed into the unidirectional image encoder, which extracts a comprehensive representation of the image's visual features. Next, the extracted image features and language representations are combined to form a comprehensive representation of both the image and the prompt. This fused representation is then fed into the Transformer-based decoder. Finally, the Transformer-based decoder

generates a caption based on its understanding of the fused visual and linguistic information. The generated caption undergoes post-processing steps, such as removing unnecessary punctuation or correcting grammatical errors, to ensure its overall clarity and accuracy.

9.3.3 Evaluation Techniques and Model Performance Determination

To evaluate the performance of the ViT GPT-2 and BLIP image captioning models, a systematic approach was adopted using the *Flickr8K* dataset (Aditya 2020). This dataset comprises 8,000 images, each accompanied by five different captions, providing a rich source for assessing the accuracy and relevance of generated captions. The Flickr8K dataset was imported and transformed into a Dataframe for ease of use in a *JupyterLab* environment. This transformation facilitated the streamlined processing of images and their corresponding ground-truth captions, essential for the evaluation process.

Evaluation Metrics

Two functions were developed to calculate text similarity, providing a quantitative measure of the captions generated by the pre-trained models against the dataset's ground-truth captions.

Cosine similarity function: This function computes the cosine similarity between two text strings. Cosine similarity measures the cosine of the angle between two vectors, in this case, the vector representations of the text strings. It's a widely used metric for gauging text similarity, with a score closer to 1 indicating higher similarity (Gunawan, Sembiring and Andri Budiman 2018).

BERT-Based similarity function: Utilizing the BERT-based text model from Reimers and Gurevych (2019), this function calculates text similarity based on sentence embeddings. BERT's contextual embedding capabilities provide a nuanced understanding of sentence similarity, making it a robust choice for comparing the semantic similarity of captions.

Evaluation Process

The evaluation process involved the following steps:

1. *Sample Selection:* A random sample of images was selected from the Flickr8K dataset. Initially, a smaller subset of 20 images was used, followed by a larger set of 200 images, to ensure a comprehensive evaluation.
2. *Model Testing:* Both the ViT GPT-2 and BLIP models were run on these samples. Each model generated captions for the selected images.
3. *Similarity Scoring:* For each generated caption, similarity scores were calculated using both the cosine similarity function and the BERT-based similarity function. These scores were compared against the ground-truth captions from the dataset.
4. *Repetition and Averaging:* To ensure reliability, this process was repeated three times for each sample set. The average similarity scores from these repetitions were used to assess the performance of each model.

The model with the highest average similarity scores across both metrics (cosine and BERT-based similarity) was considered superior. This approach evaluated the models' ability to generate syntactically correct captions and simultaneously their effectiveness in capturing the semantic essence of the images, which is crucial in the context of design inspiration. By employing these evaluation techniques, the research aims to identify the most effective pre-trained model for image captioning, ensuring that the final tool provides designers with accurate image descriptions that describe the images for better understanding of the mood board.

9.4 Results

The following section provides an overview of the main results of this project. These include the outcome of implementing the keyword suggestion generator, the presentation of results from the model evaluation, and their implications for the image captioning

implementation. To initiate the keyword suggestion process, users are prompted to input a minimum of five keywords, a prerequisite for generating pertinent suggestions. This input is transmitted through the prompt to the API, which returns four suggestions per user-entered keyword. Although the output initially displays six keyword suggestions, it remains dynamic, allowing users to shuffle through alternatives and receive six new suggestions. The implementation of this shuffling mechanism, as outlined in the methodology, was initially designed to reduce the amount of API requests in order to reduce the costs. Figure 2 in the appendix provides an illustrative example of the output from the keyword generator.

The keyword suggestion generator serves as a valuable tool within the design process dashboard in a way that by providing users with dynamic lists of suggestions, the necessity for extensive keyword research is mitigated, saving designers valuable time and effort. This can be particularly beneficial for novice designers who may not have the expertise or experience to identify relevant keywords. Moreover, the dynamic nature of the suggestions encourages experimentation and exploration, potentially leading to unexpected and creative insights.

Based on the evaluation process described in chapter *3.3.3*, the ViT GPT-2 and BLIP image captioning models were implemented and evaluated against the Flickr 8K dataset. The results of this evaluation were decisive in determining which pre-trained model would be chosen for implementation in the web application and used for the design inspiration dashboard. Figure 3 presents a clear comparison of the average accuracy scores for both models, ranging from 0 to 100%, across two evaluation methods: cosine similarity and BERT-based evaluation. The scores indicate the percentage of accuracy for each image captioning model compared to the ground-truth captions found in the dataset. This comparison of accuracy scores highlights the performance of each model across three independent runs on both large and small samples of the dataset.

The results provide evidence that the BLIP model outperforms the ViT GPT-2 model on average, both in small and large sample sizes. Specifically, the BLIP model shows higher accuracy in both the Cosine and BERT-based evaluations. For instance, within the small sample size, the BLIP model achieved Cosine scores ranging from 64% to 68% accuracy, and BERT scores from 72% to 76%. Meanwhile the ViT GPT-2 model's performance was notably lower, with its highest Cosine and BERT scores in the small sample being 58% and 67%, respectively. The difference in performance is also consistent in the larger sample size. The BLIP model maintained a higher level of accuracy and relevance in its image captioning, as indicated by its scores. For example, in the big sample dataset, the BLIP model's lowest BERT score (65%) still surpasses the highest score achieved by the ViT GPT-2 model (63%).

Another notable finding within the evaluation is the consistently higher scores achieved by the BERT-based text similarity compared to Cosine-based similarity measures. This suggests that BERT's advanced understanding of context plays a significant role in its performance. For example, consider the text strings "*A dog is running through a field"* and *"A dog is walking through the grass"*. While a Cosine similarity approach might penalize these strings more heavily due to the different words used, BERT recognizes the contextual similarity between *'running'* and *'walking'* and between *'field'* and *'grass'*. This understanding of nuanced differences is crucial for accurate image captioning, as it reflects a more sophisticated comprehension of the content and context of images.

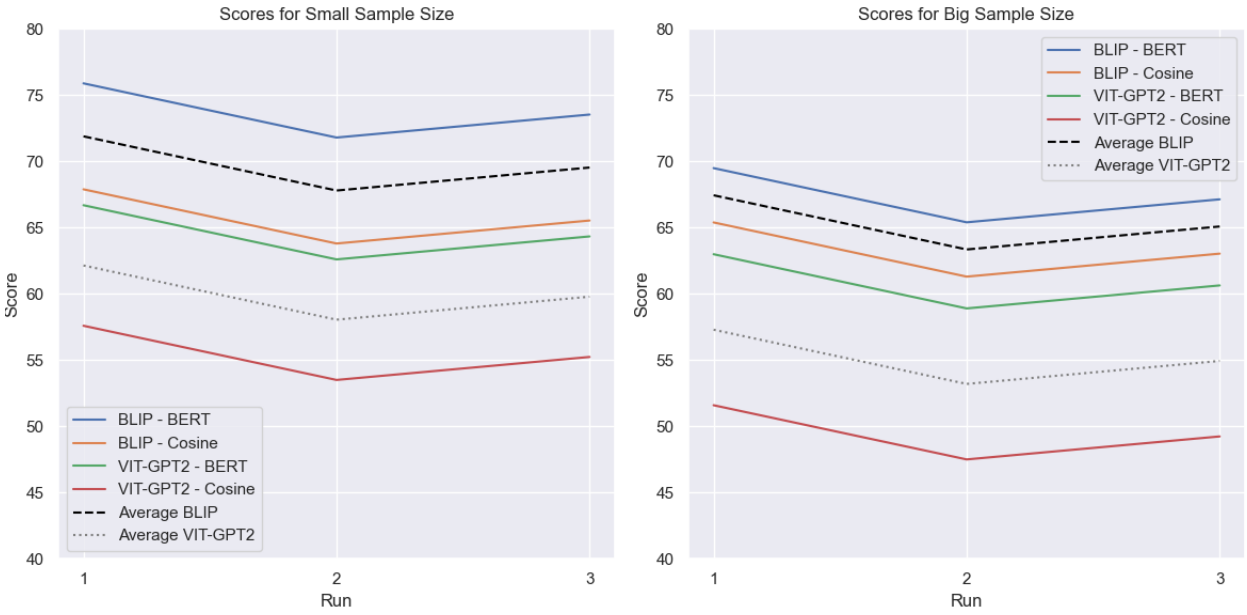

Figure 3. Accuracy scores of ViT-GPT2 and BLIP model across three different runs based on BERT and Cosine similarity.

The comparative analysis of the model evaluation revealed that the BLIP model consistently outperformed the ViT GPT-2 model, demonstrating greater accuracy and relevance in generating image descriptions (see Figure 3). This difference in performance was particularly evident in the BERT-based evaluation, which assesses the model's ability to capture contextual nuances and generate semantically accurate descriptions. The findings of the evaluation clearly indicate the BLIP model as the preferred choice for the design inspiration tool. The BLIP model's ability to provide designers with accurate and contextually relevant image descriptions will significantly enhance the tool's ability to inspire and guide designers in their creative process.

Having examined the functionality and evaluation of both the keyword suggestion generator and the image captioning model, the operational processes of these components can now be visualized in Figure 4. This graphical representation provides a clearer understanding of how these features work. The first diagram depicts the keyword suggestion process. It begins with the system prompting the user to enter a minimum of five keywords. Once the user provides the input, an API request is sent to OpenAI, configured to retrieve

relevant suggestions based on the entered keywords. The response is carefully processed and formatted, and the user is presented with six random suggestions from the total generated keywords. The user has the option to either shuffle through these suggestions or input additional words to generate new ones. The second diagram in Figure 4 focuses on the image captioning workflow. This process is initiated when the user selects images for captioning. The system then retrieves the paths of these images, and a generator function is invoked to produce relevant image descriptions. If any errors occur during this step, the captioning process is skipped; otherwise, the generated captions are stored in the user's database.

9.5 Conclusion and discussion

In concluding this section of the broader thesis, it's important to reflect on the key findings and their implications within the context of AI's role in design. The successful implementation of the keyword suggestion generator marks a significant stride in enhancing user interaction with AI tools, offering dynamic and contextually relevant suggestions based on minimal user input. This achievement demonstrates the potential of AI to augment human creativity, providing a broader spectrum of options that might not be immediately apparent to the user.

The evaluation of pre-trained models for image captioning, particularly the ViT GPT-2 and BLIP models, revealed insightful outcomes. The BLIP model's superior performance, especially in BERT-based evaluations, underscores its effectiveness in accurately understanding and describing images. This capability is crucial in design tools, where the relevance and accuracy of content can significantly impact the creative process. The findings from this section of the research highlight the advanced capabilities of AI models in assisting and enhancing the creative design process.

Reflecting on the methodology, the use of pre-trained models and the Flickr 8K dataset for evaluation was instrumental in achieving the objectives set out in this research. However, it's also crucial to acknowledge the inherent limitations. The dependency on pre-existing models means inheriting their biases and training limitations. Moreover, the evaluation metrics, while robust, may not fully encompass the subjective aspects of image captioning and keyword relevance, which are important in design contexts.

The implications of these findings are significant for the field of AI in design. The enhanced performance of the BLIP model in image captioning and the dynamic nature of the keyword suggestion generator collectively represents a substantial advancement in integrating AI into the creative design process. This integration makes the process more efficient and more intuitive, opening new avenues for designers to explore and create.

As this research concludes, it becomes evident that the integration of AI in design is an evolving journey. The potential of AI to enhance the creative process is immense, offering innovative and practical tools for designers. The future of AI in creative fields is promising, with the potential to transform how designers interact with technology, enabling them to push the boundaries of creativity and innovation. This section of the thesis contributes to the broader understanding of AI's role in design, highlighting its capabilities and the exciting possibilities it holds for the future of design innovation.

10 References and appendix

References

3DWays. (2023a). *LinkedIn*. Linkedin.com. https://www.linkedin.com/company/3dways/about/

3DWays. (2023b). *Website*. 3dways.pt. https://3dways.pt/

Aditya, J. N. (2020). *Flickr8K Dataset*. Kaggle. https://www.kaggle.com/datasets/adityajn105/flickr8k/code

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'reilly.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Bostrom, N., & Eliezer Yudkowsky. (2018). The Ethics of Artificial Intelligence. *Chapman and Hall/CRC EBooks*, 57–69. https://doi.org/10.1201/9781351251389-4

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Hesse, C. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html?utm_medium=email&utm_source=transaction

Chi, L., Jiang, B., & Mu, Y. (2020). Fast Fourier Convolution. *Advances in Neural Information Processing Systems*, *33*, 4479–4488. https://proceedings.neurips.cc/paper/2020/hash/2fd5d41ec6cfab47e32164d5624269b1-Abstract.html

Chipambwa, W., & Chikwanya, E. (2022). *Design communication: Fashion design student's perspectives on digital vs physical mood boards*. https://www.grid.uns.ac.rs/symposium/download/2022/76.pdf

Choi, W.-H., & Choi, Y.-S. (2022). Effective Pre-Training Method and Its Compositional Intelligence for Image Captioning. *Sensors*, *22*(9), 3433–3433. https://doi.org/10.3390/s22093433

Davenport, T. H. (2018). *The AI advantage : how to put the artificial intelligence revolution to work*. The Mit Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. ArXiv.org. https://arxiv.org/abs/1810.04805

EPSU. (2018). *The General Data Protection Regulation (GDPR) AN EPSU BRIEFING*.

https://www.epsu.org/sites/default/files/article/files/GDPR_FINAL_EPSU.pdf

Fenza, G., Gallo, M., Loia, V., Orciuoli, F., & Herrera-Viedma, E. (2021). Data set quality in Machine Learning: Consistency measure based on Group Decision Making. *Applied Soft Computing*, *106*, 107366–107366. https://doi.org/10.1016/j.asoc.2021.107366

Filippi, S. (2023). Measuring the Impact of ChatGPT on Fostering Concept Generation in Innovative Product Design. *Electronics*, *12*(16), 3535–3535. https://doi.org/10.3390/electronics12163535

Fox, F. (2023, September 19). *5 Challenges Businesses Face When Integrating AI Solutions*. Ratherlabs.com. https://www.ratherlabs.com/blog/5-challenges-businesses-face-when-integrating-ai-solutions

Gunawan, D., Cinta Sembiring, & Mohammad Andri Budiman. (2018). The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *Journal of Physics*, *978*, 012120–012120. https://doi.org/10.1088/1742-6596/978/1/012120

Han, T., Adams, L. C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., & Bressem, Keno K. (2023). *MedAlpaca -- An Open-Source Collection of Medical Conversational AI Models and Training Data*. ArXiv.org. https://arxiv.org/abs/2304.08247

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., Wen, J.-R., & Yuan, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225–250. https://doi.org/10.1016/j.aiopen.2021.08.002

Hulth, A. (n.d.). *Improved Automatic Keyword Extraction Given More Linguistic Knowledge*. https://aclanthology.org/W03-1028.pdf

Karabiber, F. (2023). Cosine Similarity. LearnDataSci. https://www.learndatasci.com/glossary/cosine-similarity/?fbclid=IwAR2sTQUDdMFbxA4AQFdDZDOtz7nBRDn7-nNzHxQVkeVvle8eQSS1m206-iw

Karpathy, A., & Fei-Fei, L. (2015). Deep Visual-Semantic Alignments for Generating Image Descriptions. *Cv-Foundation.org*, 3128–3137. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.html

Kotter, J. P. (1996). *Leadership change. Harvard Business School Press: Boston, MA, USA.*

Krasin I., Duerig T., Alldrin N., Ferrari V., Abu-El-Haija S., Kuznetsova A., Rom H., Uijlings J., Popov S., Veit A., Belongie S., Gomes V., Gupta A., Sun C., Chechik G., Cai D., Feng Z., Narayanan D., Murphy K. OpenImages: A public dataset for large-scale multi-label and multi-class image classification, 2017. Available from https://github.com/openimages

Kulkarni, G., Visruth Premraj, Ordóñez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2013). BabyTalk: Understanding and Generating Simple Image Descriptions. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, *35*(12), 2891–2903. https://doi.org/10.1109/tpami.2012.162

Lewin, K. (1951). *Field theory in social science: selected theoretical papers* . https://psycnet.apa.org/record/1951-06769-000

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. ArXiv.org. https://arxiv.org/abs/2201.12086

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & Mcclosky, D. (2014). *The Stanford CoreNLP Natural Language Processing Toolkit* (pp. 55–60). Association for Computational Linguistics. https://aclanthology.org/P14-5010.pdf

McKinsey. (2020). *Global survey: The state of AI in 2020*. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Global%20survey%20The%20state%20of%20AI%20in%202020/Global-survey-The-state-of-AI-in-2020.pdf

Miller, R. (2023). Holding Large Language Models to Account - PhilSci-Archive. *Pitt.edu*. http://philsci-archive.pitt.edu/22103/1/llmaccountability-paper.pdf

Nagalapuram G.D., Roopashree S., Varshashree D , Dheeraj D , & Nazareth D.J. (2021). Controlling Media Player with Hand Gestures using Convolutional Neural Network. *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*. https://doi.org/10.1109/mysurucon52639.2021.9641567

NLP Connect. (2022). *vit-gpt2-image-captioning (Revision 0e334c7)*. https://doi.org/10.57967/hf/0222

OpenAI Platform. (2023). Openai.com. https://platform.openai.com/docs/guides/text-generation

Paullada, Inioluwa Deborah Raji, Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336–100336. https://doi.org/10.1016/j.patter.2021.100336

PyFPDF. (2023). Readthedocs.io. https://pyfpdf.readthedocs.io/en/latest/ReferenceManual/index.html

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. ArXiv.org. https://arxiv.org/abs/1908.10084

Rieuf, V., Bouchard, C., & Aoussat, A. (2015). Immersive moodboards, a comparative study of industrial design inspiration material. *J. Of Design Research*, *13*(1), 78. https://doi.org/10.1504/jdr.2015.067233

Rosenthal, P., & Niggemann, O. (2022). *Problem examination for AI methods in product design*. ArXiv.org. https://arxiv.org/abs/2201.07642

Russel, S., & Norvig, P. (2020). *Artificial intelligence: a Modern approach.* (4th ed.). Prentice Hall.

Salierno, G., Leonardi, L., & Giacomo Cabri. (2021). The Future of Factories: Different Trends.

Applied Sciences, *11*(21), 9980–9980. https://doi.org/10.3390/app11219980

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, *35*, 25278–25294. https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

Specker, A. A. S. (2022). *Accelerating Digitalization within 3dways: Implementation of a Machine Learning Tool for Design Inspiration and Process Optimization*.

Tasq.ai. 2023. "What Is Bounding Box | Tasq.ai." Tasq.ai. February 22, 2023. https://www.tasq.ai/glossary/bounding-box/#:~:text=What%20is%20the%20bounding%20box,of%20interest%20inside%20each%20image.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator. *Cv-Foundation.org*, 3156–3164. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html

Wang, S. (2022). The Influence of the Sustainable Evolution of Artificial Intelligence Technology(AIT) on the Blossom of the Intelligent Technology Industry(TI). *2022 International Conference on Artificial Intelligence in Everything (AIE)*. https://doi.org/10.1109/aie57029.2022.00058

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Ruslan Salakhudinov, Zemel, R., & Yoshua Bengio. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *PMLR*, 2048–2057. https://proceedings.mlr.press/v37/xuc15.html

You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image Captioning With Semantic Attention. *Thecvf.com*, 4651–4659. https://openaccess.thecvf.com/content_cvpr_2016/html/You_Image_Captioning_With_CVPR_2016_paper.html

Yu, T., Feng, R., Feng, R., Liu, J., Jin, X., Zeng, W., & Chen, Z. (2023). *Inpaint Anything: Segment Anything Meets Image Inpainting*. ArXiv.org. https://arxiv.org/abs/2304.06790

Appendix

Databases	Variables	Type	Description	Example
user_data	username	TEXT	Represents the username of the user, and it is a primary key.	Lisa
	password	TEXT	Stores the individual keywords entered by the user.	12345
data	username	TEXT	Represents the username of the user.	Lisa
	keywords	TEXT	Stores the individual keywords entered by the user	Red_car
	sug_keywords	TEXT	Contains the generated keyword suggestions.	Canine Hound Puppy Mutt Feline Kitten Pussycat Tomcat Human Individual Being Entity Rodent Cursor Pointing device Trackball Scarlet auto Cherry motor Crimson vehicle Rubicund car
	current_datetime	TEXT	Records the timestamp when the data is stored.	06/12/2023 20:31
	images_paths	TEXT	Stores the paths of images associated with each keyword.	C:/Users/lmrod/fiftyone/open-images-v6/validation/data/4bbd710d9e8483b0.jpg
	selected_images	TEXT	Indicates which images are selected by the user.	C:/Users/lmrod/fiftyone/open-images-v6/validation/data/4bbd710d9e8483b0.jpg
	recomended_words	TEXT	Stores the recommended words for each keyword.	horse,birds
	recomended_images	TEXT	Contains the paths of recommended images for each keyword.	C:/Users/lmrod/fiftyone/open-images-v6/validation/data/1700351849_a1a11e24-3511-4730-98ce-4d886a7ed64c_remove_inpaint.png,C:/Users/lmrod/fiftyone/open-images-v6/validation/data/1700352753_aa972123-fd5e-4b9d-880b-fc35bd3f8376_remove_inpaint.png,C:/Users/lmrod/fiftyone/open-images-v6/validation/data/1700351849_a1a11e24-3511-4730-98ce-4d886a7ed64c_remove_inpaint.png
	images_descripition	TEXT	Holds the descriptions associated with each image.	a red car parked in front of a store
	inpaint_save	TEXT	Stores the paths of images after modification using inpainting.	C:/Users/lmrod/fiftyone/open-images-v6/validation/data/1701895364_380300c3-101f-4b81-bfb3-621ad598099a_sam.png
keywords	username	TEXT	Represents the username of the user.	Lisa
	keywords	TEXT	Stores the individual keywords entered by the user.	dog,cat,person,mouse,red_car
	current_datetime	TEXT	Records the timestamp when the data is stored.	06/12/2023 20:31
	recomended_words	TEXT	Stores the recommended words for each set of keywords.	horse,birds

Table3. Results of the Databases

```
Final Suggestions based on the input keywords of the user ['dog', 'blue', 'teddybear', 'car', 'blackwidow']
1. widowmaker
2. sapphire
3. navy
4. pup
5. doll
6. spider
```

Figure 2. Showcasing final keyword suggestions based on user input.

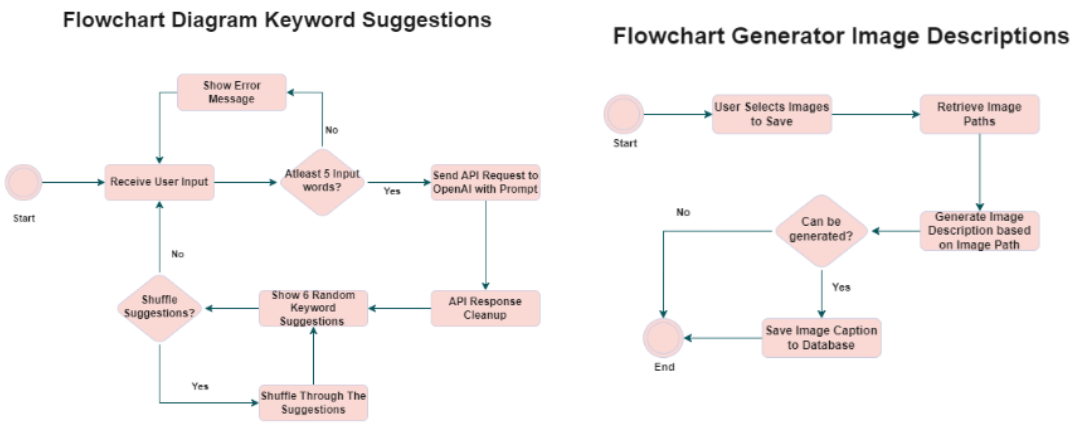

Figure 4. Flowchart diagrams of functionalities implemented.

Figure 25. Web application Log in page

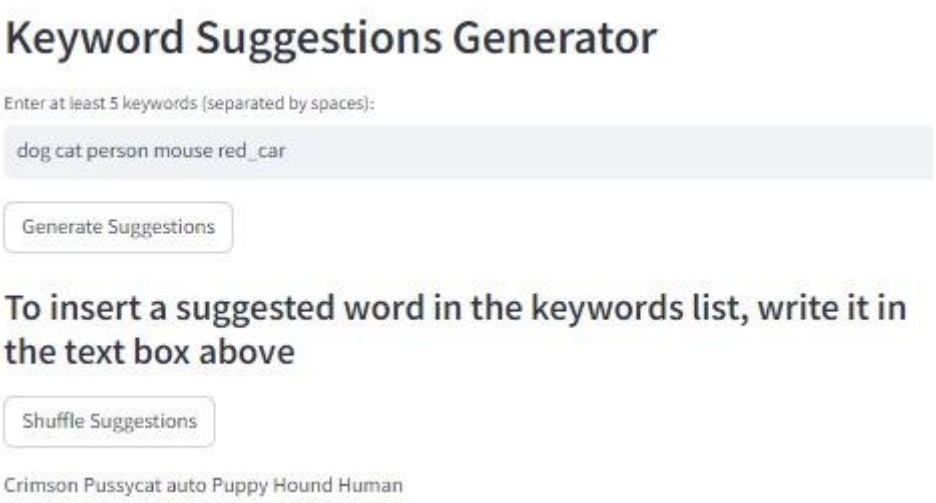

Figure 26. Key word recommender system

Figure 27. Suggested images based on the key words inserted

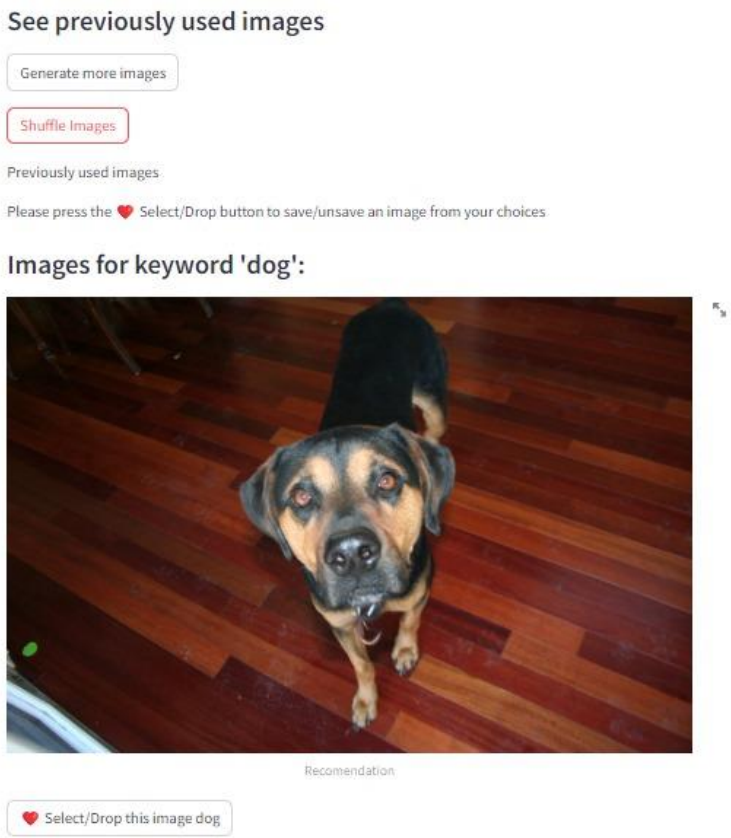

Figure 28. Suggested images based on previous interactions.

Selected Images:

The selected images: red_car_2

The caption of the selected images: a red car parked in front of a store

Modify Images

Only applicable if the user selected images before

Press the button bellow to set up the functionality with the images selected. This may take a while, depending on the number of images. Instructions will follow.

Modify images selected

By clicking in the buttons bellow, a pop up window will appear with the selected image. You will be asked to select points to determine the object to isolate, and optionally set a box limit where the object is contained.

To define object position, please select points with a left mouse click to highlight areas of interest in the image, positive input points, and with a right mouse click to select areas for the model to ignore, negative input points, whatever is there, the model will not include in the final image. To close the pop up window, press the 'q' key.

To define a box where the object should be contained, select 2 points with a left mouse click. The first will be considered the top left box corner and the second the bottom right box corner. The pop up window should close after the second click.

red_car_2

Get Isolated element: Get with element position

Get Isolated element: Get with element + box position

Remove Element: Get with element position

Remove Element: Get with element + box position

Ready to generate the Mood Board?

Create a Mood Board with all images selected and modified. To make any changes, go back and find new images.

Mood Board Generation?

Figure 29. Image modifying section.

Figure 30. Pop up with image selected to gather segmentation coordinates.

Figure 31. Mood board generation section

Figure 32. Mood board generation section

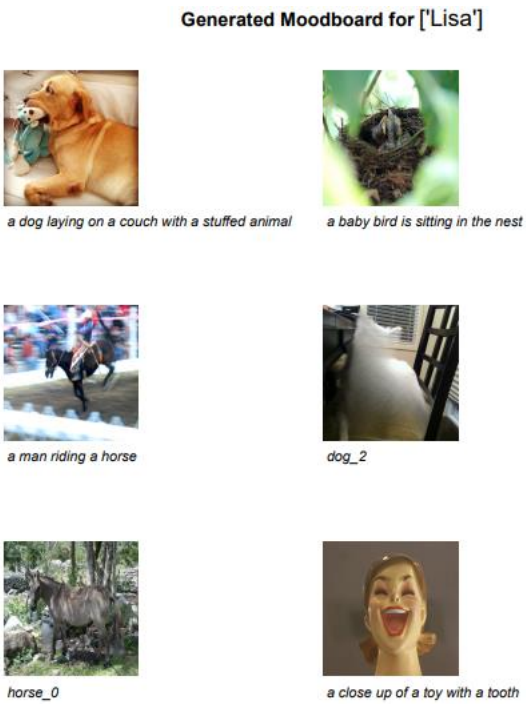

Figure 33. Mood board example