

AI-based Personalization of Social Media Thumbnails Using the Stacked ID Embedding Method

Meivi Kartikasari^{1*}, Hashfi Andira Putra², Mukhlis Amien³

¹Department of Information System, Universitas Bhinneka Nusantara, Malang, East Java, Indonesia

^{2,3}Department of Informatics, Universitas Bhinneka Nusantara, Malang, East Java, Indonesia

E-mail: ^{1*}meivi.k@ubhinus.ac.id, ²hashfi.ap@gmail.com, ³amien@ubhinus.ac.id

(Received: 17 Feb 2025, revised: 10 Apr 2025, accepted: 15 Apr 2025)

Abstract

Social media content creators find it hard to make thumbnails shine on Instagram, YouTube, and TikTok, where graphic design skills happen to be a significant bottleneck. To address this issue, scientists developed a text-based image generation model, PotionPix, that allows users to generate thumbnails on the fly based on text prompts and relevant images via a "Stacked ID Embedding" method. This method combines multiple identity embeddings—e.g., user interests, platform context, and content genre—into one vector representation to guide the AI to create more personalized and contextually appealing thumbnails. The system integrates a diffusion-based image generator with the stacked embedding vectors to enable dynamic adaptation to different user intents. In tests, it was observed that how relevant and good the generated thumbnails were very much a function of how specific the input image was and how clear the prompt was. However, since the AI model used was not fine-tuned on the task of thumbnail generation specifically, the visual outputs sometimes were generic and lacked the strong call-to-action elements usually found in high-performing thumbnails. Despite this constraint, the usability test conducted with 120 respondents showed promising results—83.8% of the participants confirmed that PotionPix was indeed assistive in the thumbnail design process, particularly in terms of time and effort savings. The findings show the promise of AI-driven tools in enabling the democratization of design tasks for social media content creators, as well as suggesting future work in model fine-tuning for more domain-specific outcome.

Keywords: Thumbnail, Stacked ID Embedding, Artificial Intelligence, Content Creator, Text-to-image Generation.

I. INTRODUCTION

Social media platforms such as Instagram, YouTube, and TikTok have become the largest platform for sharing visual content. The thumbnail—small, eye-catching visuals that give a hint of larger content such as videos or articles—has become one of the most crucial elements of such websites. Organizing and managing content on social media platforms, especially YouTube, can be a challenging task [1]. It addresses numerous specifics and operations, and the users oftentimes encounter tremendous amounts of troubles which can't always be done so quickly or so easily. It encompasses, inter alia, generation of good-looking and correct thumbnails that typically requires rather a lot of time and skills in computer graphics. Numerous authors of content often fail to generate interesting thumbnails. Therefore, the development of a software tool named PotionPix is intended to streamline and assist content creators in creating attractive and pertinent thumbnails. PotionPix employs the SDXL AI model and the "Stacked ID Embedding" method to achieve maximum accuracy and efficiency in image

generation [2]. While many studies have focused on text-to-image synthesis, there is still limited research specifically targeted at thumbnail generation [3].

We contrast against other dominant methods in other pieces of work in order to emphasize the novelty of this research since all the former methods utilize general-purpose image generation without attempting to optimize the distinctive visual and contextual requirements for social media thumbnails. Therefore, PotionPix seeks to focus its output on generating images that are specially tailored to function as thumbnails.

II. RESEARCH METHODS

There was the collection of data through questionnaires. There were two questionnaires: the first one was to serve as preliminary research to build upon the problem formulation and the second one served as evaluation data to monitor whether or not the software works for content creators. Also,

input data was collected from available datasets and images or photos submitted by users for experimentation [4].

Throughout the image creation process, the Stacked ID Embedding method was used through the combination of a number of user-specific identity vectors such as visual style preference, content type, and platform context into an aggregate composite embedding. The composite vector served as a reference for the SDXL-based AI model at inference time to construct more personalized, relevant, and contextual thumbnails. This approach was preferred to the traditional single-vector embedding methods as it allows the processing of multiple layers of user data simultaneously and the provision of a more advanced and dynamic output.

For further reinforcement of the uniqueness of this research, there is a need to include more references of previous research on AI use in visual content generation. Research on personalized image generation [5] and user-driven design in AI-powered image software [6], provides a useful basis for comparison and highlights the uniqueness of the approach used in developing PotionPix compared to existing methods [7]. Figure 1 illustrates the system or workflow of the research, starting from data collection to testing.

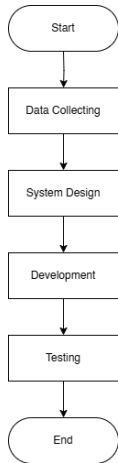


Figure 1. Design of the Research System

Table 1 contains statistical data obtained from the preliminary research questionnaire. The questionnaire consists

of eight questions addressing whether respondents require a tool to assist in creating thumbnails. The analysis began by calculating the mean for each question, which was done by summing the weights or points from all respondents and then dividing by the total number of respondents. Next, the grand mean was calculated by summing the means of all questions and dividing by the total number of questions, which is eight. This resulted in a grand mean of 3.775. An average value ranging from 3.4 to 4.2 is generally interpreted as indicating 'agree' or 'important' in frequency distribution data [8].

Therefore, the grand mean of 3.775 indicates that, on average, respondents agreed or had a positive response to each question in the questionnaire. Additionally, the population standard deviation obtained from this analysis was 0.76. This standard deviation is considered moderate, suggesting that while respondents generally agreed with most questions, there was still a notable variation in their responses. In fact, the majority of respondents tended to respond more neutrally to some questions in the questionnaire. This indicates that the respondents did not exhibit a biased perspective. Thus, the analysis of the questionnaire data demonstrates that creating thumbnails is a significant challenge for content creators without a background in graphic design. This conclusion strongly supports the need for a solution to address the aforementioned problem [9].

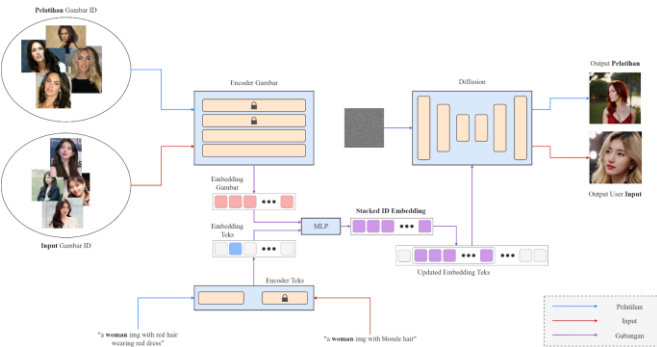


Figure 2. Image Generating Charts

Table 1. Data Preliminary Research

Question	1		2		3		4		5		Total		Mean
	F	%	F	%	F	%	F	%	F	%	F	%	
P.1	1	6.67	0	0	0	0	6	40	8	53.33	15	100	4.33
P.2	0	0	0	0	0	0	5	33.33	10	66.67	15	100	4.67
P.3	0	0	0	0	4	26.67	9	60	2	13.33	15	100	3.87
P.4	2	13.33	6	40	4	26.67	2	13.33	1	6.67	15	100	2.6
P.5	1	6.67	1	6.67	3	20	6	40	4	26.67	15	100	3.73
P.6	0	0	10	66.67	2	13.33	3	20	0	0	15	100	2.53
P.7	0	0	0	0	1	6.67	5	33.33	9	60	15	100	4.53
P.8	2	13.33	0	0	1	6.67	6	40	6	40	15	100	3.93
Grand Mean													3.79

The workflow depicted in Figure 2 begins with the extraction of image and text embeddings from both the input and the pre-trained model. Once the embeddings for each are obtained, they are combined using a Multi-Layer Perceptron (MLP) to generate the “Stacked ID Embedding”. The process of updating the Stacked ID Embedding is a crucial step aimed at improving the accuracy of the generated images and ensuring higher relevance [7]. The updated embedding is then fed into a diffusion model, which starts from random noise. In this process, the diffusion model is tasked with ‘imagining’ the image based on the updated Stacked ID Embedding, thereby producing the desired image that aligns with the embedding.

The meaning of the key icon in Figure 2 indicates that the corresponding input remains unchanged throughout the process. For example, the text input does not change during the process because any alteration to the text input would result in generated outputs that do not meet user expectations. Similarly, the key icon on the image training indicates that certain parameters or layers within the image encoder remain fixed during the training process. This is because the dataset is used for stable feature extraction without updates or fine-tuning. The goal is to maintain the consistency of the learned features and prevent them from being affected by noise or irrelevant changes, thereby ensuring that the ID image characteristics are extracted accurately and stably. In other words, the locked parameters help preserve the integrity of the learned features and support the model’s ability to generate high-quality image representations [10].

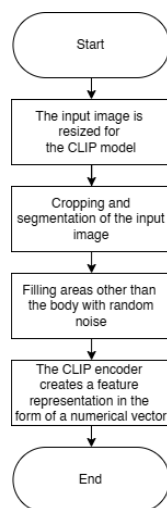


Figure 3. Flowchart Pictures Embedding

The CLIP encoder in Figure 3, functioning as the image encoder, begins the process by feeding the input image into the model. The image undergoes pre-processing, which includes resizing to match the specifications of the CLIP model, thereby enhancing the model's efficiency. Next, CLIP performs cropping and segmentation to focus on the main object, reducing interference from other elements such as the background and irrelevant objects. CLIP also adds random noise to areas outside the main object to minimize disturbances in feature detection and extraction. After this pre-processing

stage, CLIP utilizes the Vision Transformer (ViT) architecture to process the image in patches, generating feature representations in the form of numerical vectors known as “Image Embedding”, which encapsulate essential visual information from the image.

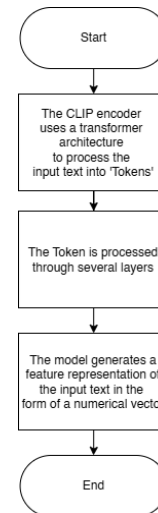


Figure 4. Flowchart Text Embedding

In Figure 4, the text encoder also utilizes the CLIP encoder. The process of obtaining “Text Embedding” begins with CLIP using its transformer architecture to convert the input text into token units. Once the tokens are formed, they are processed through multiple transformer layers to capture the context and relationships between words, which are essential for understanding the overall meaning of the text. This process ensures that the model can better comprehend the semantics of the text, making the generated results more relevant. Finally, CLIP converts the processed text into a unique feature representation in the form of numerical vectors, which can then be used in subsequent processes [11].

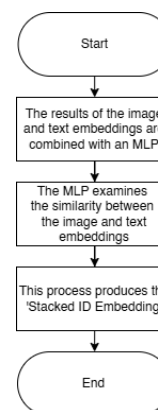


Figure 5. Flowchart MLP

The image and text embeddings obtained from the previous processes are combined using a Multi-Layer Perceptron (MLP). The MLP performs a fusion operation by stacking

layers of image and text embeddings to produce a unified representation, as shown in Figure 5. This process allows the MLP to evaluate the similarity between the text and image embeddings, thereby minimizing mismatches and improving accuracy. The result of this process is the “Stacked ID Embedding”, a more complex and feature-rich representation that aligns with the input (Figure 6).

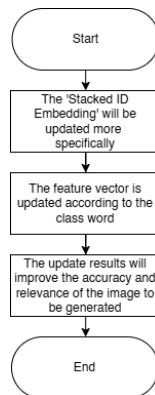


Figure 6. Flowchart Update of Stacked ID Embedding

The Stacked ID Embedding obtained from the MLP process is then used to update the initial text embedding. Specifically, the feature vectors at positions corresponding to class words (e.g., “man” or “woman”) in the original text embedding are replaced with the Stacked ID Embedding. This update aims to enhance the alignment of the generated image with user expectations, as it allows the integration of specific identity features from the Stacked ID Embedding into the text embedding. In this way, the generated image becomes more relevant and accurate, matching both the text description and the desired identity characteristics. The illustration in Figure 7 below depicts the process of this step [12].

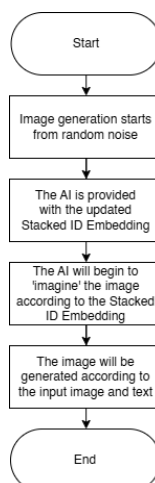


Figure 7. Flowchart Diffusion

The diffusion model requires training on a specialized dataset before it can generate images. During training, the model gradually transforms the input image into simple noise through a series of steps. After this transformation process, the model is tasked with reconstructing the image from the noise,

ensuring that the integrated image information is preserved. This process continues until the image is completely transformed into noise. The model then uses the updated Stacked ID Embedding vector to ‘imagine’ and generate a new image that aligns with the user’s input image and text. This process involves a cross-attention mechanism to combine the ID information from the Stacked ID Embedding with the text embedding, ensuring that the generated image matches the specified identity and the provided text description.

III. RESULT AND DISSCUSSION

The implementation of the PotionPix application was mainly done by using Visual Studio Code as the primary software, along with Python as the used programming language. The SDXL AI model, which is open source, is the dominant component of this application. The second dominant component is the approach that was followed by PhotoMaker, which is also open source. Data used in this research involves the use of photos or images that have been extended to JPG, PNG, and JPEG. The data includes existing data and images or photos that are user-tested. With a combination of data types, the testing becomes more extensive and significant [13].

This research is more focused on generating thumbnails for social media that are visually attractive and significant to the content that is being searched for. It should thus aid in generating more interactive thumbnails, which will generate more user engagement.

To further boost the contribution of this study, the strengths and limitations of the method employed should be examined more deeply. One strength of the Stacked ID Embedding approach is that it can consider multi-dimensional user preferences and produce more personalized and context-specific visual content. Unlike other single-embedding or generic diffusion model approaches, the method adaptively adjusts the generation process to serve different user purposes and platform demands. However, one of its limitations is that the model has not been specifically fine-tuned for thumbnail generation, which at times results in outputs lacking the strong visual signals typical of highly performing thumbnails—e.g., face emphasis, bold text, or high contrast [7].

On the other hand, other studies using fine-tuned models or template-based techniques can have more consistent visual results but without the customization and flexibility of the Stacked ID Embedding method. Therefore, while PotionPix introduces a novel and changing method, future work should include hybrid techniques combining customization with features specifically optimized for thumbnail attractiveness.

The login page will be displayed after the splash screen. Users are required to log in to access the features of the application. Figure 8 illustrates the layout of the login page. This page includes a login feature, which requires a database for verifying registered accounts. If the user successfully logs in, a new folder will be created for the user, corresponding to their username [14].

Users can utilize two features to input the ID image that will be used by the AI in the image generation process. By

using this ID image, the AI model can generate images with characteristics derived from the input image. One of the features provided is the selection of an existing image owned by the user. In this feature, the file path of the selected image is saved and then passed into the image generation function to be processed into an ID image. Figure 9 displays the interface for one of the ID image input features, which involves selecting an image.

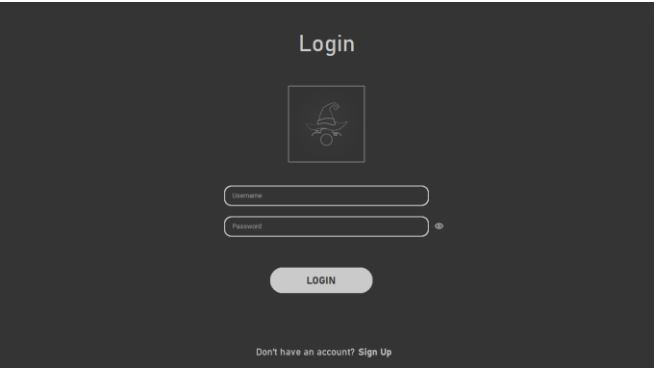


Figure 8. Login Page



Figure 9. Select Image Page

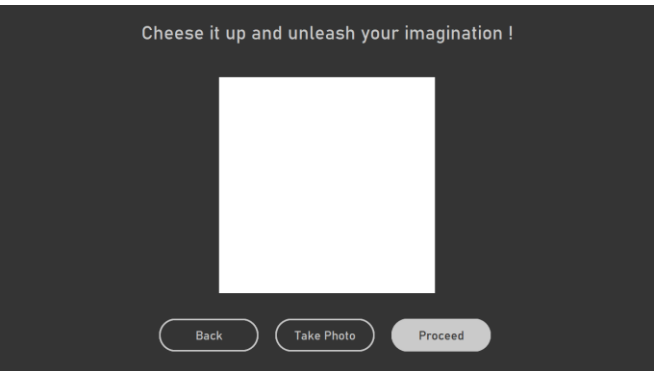


Figure 10. Select Camera Page

Another feature available for inputting the ID image is the use of the user’s camera device. On this page, users can access their camera and capture an image directly, which will then be processed into an ID image. With these two features, users have greater flexibility in inputting ID images according to their needs. Figure 10 illustrates the interface for the camera feature [15].

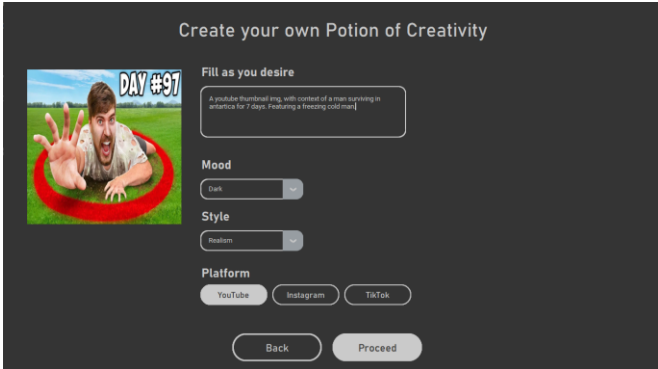


Figure 11. Prompt Page

One of the critical aspects of the image generation process is the input prompt. The quality of the generated image heavily depends on the quality of the prompt provided. To address this issue, the author implemented a hybrid approach in the application’s user interface (UI), which includes the use of a text box and a combo box. This approach aims to minimize the likelihood of poor-quality input prompts while still providing users with high flexibility to type their prompts in the text box. The UI can be seen in Figure 11.



Figure 12. Image Generation Result

The generation result from the input image with the prompt “A YouTube thumbnail image, with the context of a man surviving in Antarctica for 7 days. Featuring a freezing cold man.” is shown in Figure 12. This generation process produces a thumbnail designed for the YouTube platform with a resolution of 1280 x 720 pixels. The image depicts a man striving to survive in Antarctica, highlighting the extreme cold conditions. This thumbnail is designed to capture the audience's attention and provide a clear representation of the challenges faced by the main character in the content [8].

After the design and development of PotionPix were completed, it was necessary to conduct testing with respondents to validate whether PotionPix could indeed assist them. Therefore, the second questionnaire was distributed after the software was finalized. The questionnaire questions focused on whether the respondents felt the software helped them in creating thumbnails.

Table 2. Evaluation Data

Question	1		2		3		4		5		Total		Mean
	F	%	F	%	F	%	F	%	F	%	F	%	
P.1	0	0	0	0	0	0	9	60	6	40	15	100	4.4
P.2	0	0	1	6.67	1	6.67	8	53.33	5	33.33	15	100	4.13
P.3	0	6.67	0	0	2	13.33	8	53.33	4	26.67	15	100	3.93
P.4	0	0	0	0	1	6.67	9	60	5	33.33	15	100	4.27
P.5	0	0	1	6.67	1	6.67	7	46.67	6	40	15	100	4.2
Grand Mean													4.19

As shown in Table 2, an average value between 3.4 and 4.2 is typically interpreted as representing the category of "agree" or "important" in frequency distribution data.[8]. Meanwhile, the grand mean obtained from the calculations in Table 2 is 4.19, indicating that the majority of respondents provided very positive or agreeing responses to the questions asked. When combined with a standard deviation of 0.15, this data analysis suggests consistency in the highly positive response category from the participants [16].

However, there is some variation in the responses, indicating that a few respondents still have suggestions for PotionPix [17]. For example, in the PotionPix interface design question, one respondent rated "Poor" and another chose "Neutral". Therefore, it can be concluded that while the majority of respondents provided positive feedback, PotionPix still requires further development in various aspects, including interface design, features, and the quality of the generated images (thumbnails) [18].

IV. CONCLUSION

The present study presents a valuable contribution towards the development of AI-powered software for assisting content creators to design pertinent and visually appealing social media thumbnails. By integrating the SDXL model and Stacked ID Embedding method, PotionPix software facilitates the process for something that occasionally demands advanced graphic design skills. User testing outcomes show that 83.8% of the participants stated that PotionPix successfully solved issues they typically experience when generating thumbnails. This illustrates the practical value of PotionPix in helping creators, particularly its ability to produce more personalized and context-aware visual content, as compared to other methods of generating generic results.

Despite these positive results, there is still plenty of room for improvement. One place where improvement is needed is in using a dedicated AI model that has been trained for thumbnail creation. The SDXL model is general-purpose now, and this can at times lead to subpar results. Creating a model with training focused on the specific visual demands of thumbnails—faces being prominent, legible typography, and contrasting color—would greatly increase output quality. Additionally, since the quality of the output images significantly depends on the input prompt offered by the user, constructing features that assist in prompt generation—such as intelligent suggestions or automatic improvements—would

be a beneficial feature, especially for inexperienced users of prompt engineering techniques.

Another obstacle is processing speed, particularly in the case of limited hardware. To fight this, relocating the program to a cloud-based, web-based environment would allow for faster generation times and increased accessibility. Additionally, the model's limited capability to understand Indonesian-language inputs suggests a need for more language-processing capability. Improved ability for the system to comprehend Indonesian contextually would improve the user's experience among a local population and enhance its practical usage.

Considering each of these updates together, a bright future trend is indicated for PotionPix. With ongoing updates, it may become a best-in-class automatic thumbnail generation tool—adaptive, efficient, and well-positioned to meet content creators' demands in Indonesia as well as across the globe.

REFERENCES

- [1] C. I. Lestari and I. Irwansyah, 'Kolaborasi Produksi Konten YouTube melalui Multi-Channel Network: Studi pada Kreator Sandy SS dengan Collab Asia', *J. Ris. Komun.*, vol. 4, no. 1, pp. 143–159, Mar. 2021, doi: 10.38194/jurkom.v4i1.152.
- [2] M. Arar et al., 'Domain-Agnostic Tuning-Encoder for Fast Personalization of Text-To-Image Models', Jul. 13, 2023, *arXiv*: arXiv:2307.06925. doi: 10.48550/arXiv.2307.06925.
- [3] B. Rahadjo, Belajar Otodidak Membuat Database menggunakan MySQL, 1st ed. Bandung: Bandung, Informtika, 2011. [Online]. Available: <https://perpustakaan.binadarma.ac.id/opac/detail-opac?id=15408>
- [4] F. Firdausi, 'Pengembangan Aplikasi Online Public Access Catalog (Opac) Perpustakaan Berbasis Mobile Pada Stai Auliaurasyidin', *J. Intra Tech*, vol. 4, no. 2, pp. 11–24, 2020, doi: <https://doi.org/10.37030/jit.v4i2.74>.
- [5] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim, 'Text-to-image Diffusion Models in Generative AI: A Survey', Nov. 08, 2024, *arXiv*: arXiv:2303.07909. doi: 10.48550/arXiv.2303.07909.
- [6] J. Liu and Y. Zhu, 'Precise Correspondence Enhanced GAN for Person Image Generation', *Neural Process. Lett.*, vol. 54, pp. 5125–5142, 2022.



- [7] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, 'PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding', Dec. 07, 2023, *arXiv*: arXiv:2312.04461. doi: 10.48550/arXiv.2312.04461.
- [8] J. D. Stemple, 'Job Satisfaction Of High School Principals In Virginia', Diss. Submitt. Fac. Va. Polytech. Inst. State Univ., Apr. 2004.
- [9] R. Gal et al., 'An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion', Aug. 02, 2022, *arXiv*: arXiv:2208.01618. doi: 10.48550/arXiv.2208.01618.
- [10] I. M. De La Vega Hernández, A. Serrano Urdaneta, and O. Schiappa-Pietra, 'Scientific mapping of artificial intelligence as an emerging field of knowledge', in Handbook of Research on Artificial Intelligence, Innovation and Entrepreneurship, E. Carayannis and E. Grigoroudis, Eds., Edward Elgar Publishing, 2023, pp. 8–28. doi: 10.4337/9781839106750.00008.
- [11] J. Ho, A. Jain, and P. Abbeel, 'Denoising Diffusion Probabilistic Models', 34th Conf. Neural Inf. Process. Syst. NeurIPS 2020 Vanc. Can., Dec. 2020.
- [12] M. Irfan, H. Siregar, and J. T. Handoko, 'Pengembangan Dan Integrasi Aplikasi Prediksi Jumlah Gagal Produksi PC Menggunakan Metode Triple Exponential Smoothing Pada Sistem Aplikasi Produksi Di PT Tera Data Indonusa, Tbk', Pros. Semin. Nas. Darmajaya, vol. 1, 2023.
- [13] X. Jia et al., 'Taming Encoder for Zero Fine-tuning Image Customization with Text-to-Image Diffusion Models', Apr. 05, 2023, *arXiv*: arXiv:2304.02642. doi: 10.48550/arXiv.2304.02642.
- [14] W. Peebles and S. Xie, 'Scalable Diffusion Models with Transformers', Mar. 02, 2023, *arXiv*: arXiv:2212.09748. doi: 10.48550/arXiv.2212.09748.
- [15] A. Radford et al., 'Learning Transferable Visual Models From Natural Language Supervision', Feb. 26, 2021, *arXiv*: arXiv:2103.00020. doi: 10.48550/arXiv.2103.00020.
- [16] J. Sutrisno and V. Karnadi, 'Aplikasi Pendukung Pembelajaran Bahasa Inggris Menggunakan Media Lagu Berbasis Android', *J. Comasie*, vol. 4, no. 6, 2021.
- [17] I. S. Akbar and T. Haryanti, 'Pengembangan Entity Relationship Diagram Database Toko Online Ira Surabaya', *Comput. Insight J. Comput. Sci.*, vol. 3, no. 2, pp. 28–35, Jul. 2023, doi: 10.30651/comp_insight.v3i2.12002.
- [18] T. H. Trinh, M.-T. Luong, and Q. V. Le, 'Selfie: Self-supervised Pretraining for Image Embedding', Jul. 27, 2019, *arXiv*: arXiv:1906.02940. doi: 10.48550/arXiv.1906.02940.