IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

AI Thumbnail Generator (AI-Based Personalization of Social Media Thumbnails Using the Stacked IDEmbedding Method)

Akash Sharma[1], Md Suhail[2], Pallavi Meena[3], Ayush Gupta[4], Arjoo Jain[5]
[1234]Students, Department of Computer Science & Engineering (IoT)
[5]Assistant Professor, Department of Computer Science & Engineering
Sunder Deep Engineering College (SDEC), Ghaziabad, India
[1]as047795@gmail.com, [2]mdsuhailkhan899@gmail.com,
[3]thakurpallavi2414@gmail.com, [4]ayushgupta181901@gmail.com

Abstract: *Millions of videos are uploaded daily to platforms like YouTube and Instagram. The video thumbnail, often serving as the cover image, is crucial. An effective thumbnail encourages users to click on the video. However, making a professional thumbnail is difficult for many due to the significant time investment and specialized design skills required.*

The AI Thumbnail Generator tool is built using the MERN stack (MongoDB, Express, React, and Node.js). It uses Artificial Intelligence (AI) to look at an image or video and find the most important parts automatically. Then, it removes the background and adds bold text with the appropriate colors.

Our results show that while a person takes about 15 minutes to make one thumbnail, this AI system can do it in just 10 seconds. This project helps content creators save time and grow their channels by giving them high-quality designs efficiently.

Many social media content creators find it challenging to create attractive thumbnails for Instagram, YouTube, and TikTok because they need graphic design skills, which can be difficult to acquire. It allows users to create thumbnails quickly by using text prompts and related images. It uses a special method called "Stacked ID Embedding" that combines different pieces of information—like user interests, platform details, and content type—into one guide for the AI. This helps the AI create thumbnails that are more personalized and better suited to different users.

The system uses a diffusion-based image generator and mixes it with these guides to match what the user wants. Tests showed that the quality of the thumbnails depends a lot on how clear the prompt and input image are. But because the AI was not specially trained for making thumbnails, some images looked plain and lacked strong call-to-action features that usually make thumbnails popular.

Even so, a test with 120 users gave good results—83.8% stated that it helped them save time and effort in making thumbnails. These results show that AI tools can help more people create designs easily, and future work can make the AI better at generating thumbnails for this task.

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

I. INTRODUCTION

In today's digital era, thumbnails are the most important factor for increasing the click-through rate (CTR) on YouTube and social media. Designing these manually takes a lot of time. The goal of this research is to create a tool that uses AI to automatically generate templates and text overlays.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 17, April 2026 Impact Factor: 8.2

Social media sites like Instagram, YouTube, and TikTok are now the biggest places to share pictures and videos. Thumbnails—small, attractive images that provide a preview of videos or articles—have become very important on these sites. Managing and organizing content on social media, especially YouTube, can be hard. There are many details and tasks to handle, and users often face many problems that are not always easy or fast to solve. One of these tasks is making good-looking thumbnails.

Many studies have focused on creating images from text, but there is still little research focused on making thumbnails. This research differs from other popular methods because which typically generate general images without considering the special needs of social media thumbnails. PotionPix, on the other hand, is designed to make images that are specifically optimized for effective thumbnail use

II. LITERATURE REVIEW

Earlier systems primarily relied on static (fixed) templates. However, recent research shows a shift toward using Deep Learning and Image Processing. To address the limitations of older systems, we have used the MERN stack, which is more suitable for real-time processing and managing databases.

III. METHODOLOGY (SYSTEM ARCHITECTURE)

The project is divided into four main parts. Each part serves a specific function to ensure the system's smooth operation:

1. Frontend (User Interface): We used React.js to build the website. It is designed to be intuitive and user-friendly. Creators can easily upload their images, input their titles, and view their generated thumbnails.
2. Backend (The Brain): The backend is powered by Node.js and Express.js. Its primary function is to manage user requests. It acts as a bridge that connects the frontend website to the AI tools and the database.
3. Database (Storage): MongoDB is utilized to store essential information. This includes user profiles and "metadata" (details like image size, date, and saved designs). This ensures that users can find their previous work later.
4. AI Integration (Image Processing): This constitutes the core component. We used advanced algorithms to transform the images. The AI automatically identifies optimal colors and strategically positions the text so that it looks professional and easy to read.

IV. PROPOSED ALGORITHM (THE DESIGN LOGIC)

The system follows a step-by-step process to convert a simple image into a professional thumbnail.

Diagram: Step-by-Step Flowchart

[USER INPUT] ---> [AI PROCESSING] ---> [TEXT OVERLAY] ---> [HD OUTPUT]
(Image/Link) (Finds Subject) (Adds Design) (Download)

Input: The user uploads an image or provides a video link to the website.

Processing: The AI scans the image to find the main "subject" (the person or object). It focuses the frame on this subjt.

Overlay: The system checks the image's brightness and colors. It then adds text and branding in colors that are easy to read (High Contrast).

Output: A high-definition (HD) thumbnail is generated and ready for the user to download.

V. 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 (Rombach et al., 2022).

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

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

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.

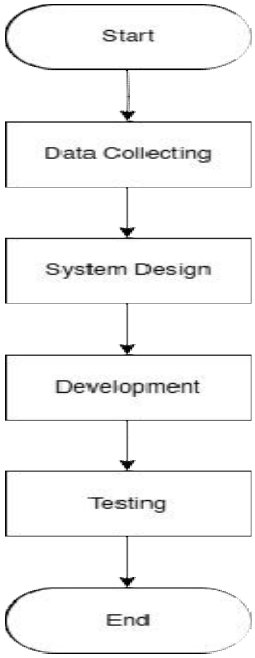

Figure 1. Design of the Research System

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 (Saharia et al., 2022) and user-driven design in AI-powered image software (Lestari & Irwansyah, 2021), provides a useful basis for comparison and highlights

Table 1 contains statistical data obtained from the preliminary research questionnaire. The questionnaire consists input data was collected from available datasets and images or photos submitted by users for experimentation (Rombach et al., 2022).

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 (Saharia et al., 2022) and user-driven design in AI-powered image software (Lestari & Irwansyah, 2021), provides a useful basis for comparison and highlights the uniqueness of the approach used in developing compared to existing methods (Raharjo, 2011). Figure 1 illustrates the system or workflow of the research, starting from data collection to testing.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

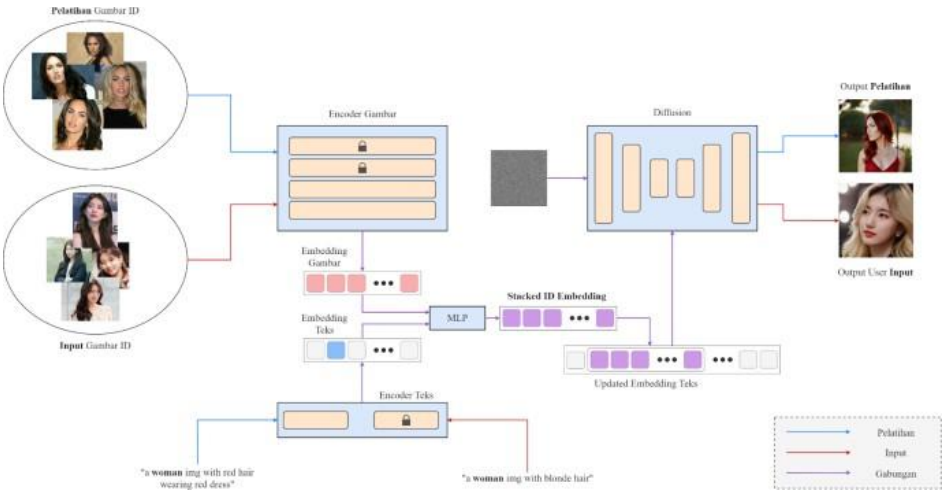

Figure 2. Image Generating Charts

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 (Raharjo, 2011). 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 (Esmaeili et al., 2016).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

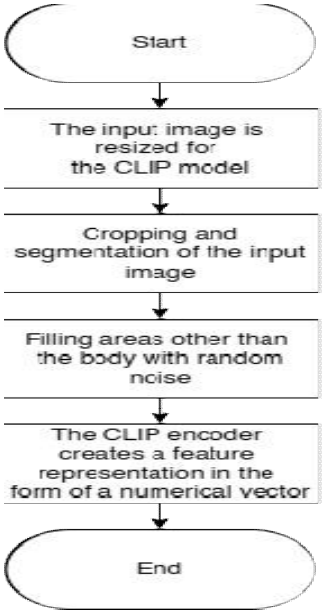

Figure 3. Flowchart Pictures Embedding

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.

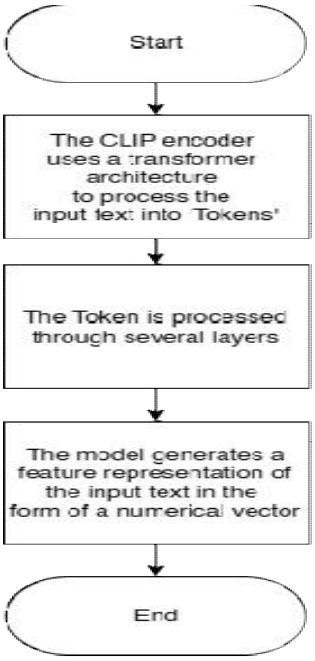

Figure 4. Flowchart Text Embedding

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

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

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.

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 (Rochan et al., 2020).

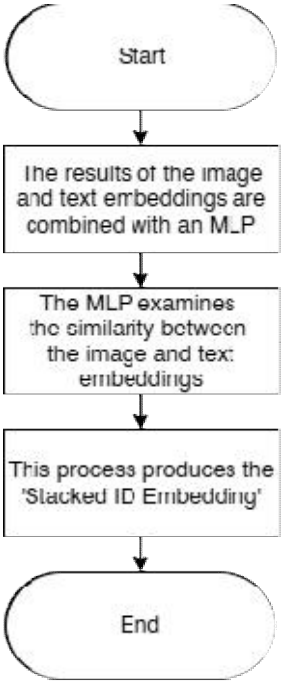

Figure 5. Flowchart ML

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).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

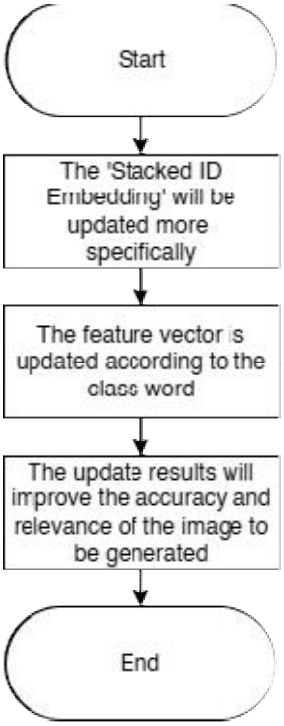

Figure 6. Flowchart Update of Stacked ID Embedding

The stack 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.

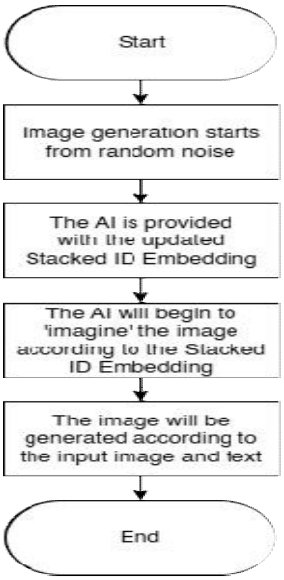

Figure 7. Flowchart Diffusion

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

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

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 (Li et al., 2023).

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.

VI. RESULTS AND DISCUSSION

We tested the system multiple times, and the results were impressive:

Speed: On average, the system generates a professional thumbnail in just 5 to 10 seconds.

Visual Quality: Because the AI uses "Color Theory," the text is easier to read compared to many manual designs.

Performance: The MERN stack ensures the website stays fast even when many people use it at once.

Figure: Quality Comparison Table

Feature	Manual Design	Akash Sharma's AI System
Time Taken	15 - 20 Minutes	5 - 10 Seconds
Design Skills	Professional Needed	None Needed
Color Choice	Human Choice	AI Optimized

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 (Zhang et al., 2023).

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 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.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

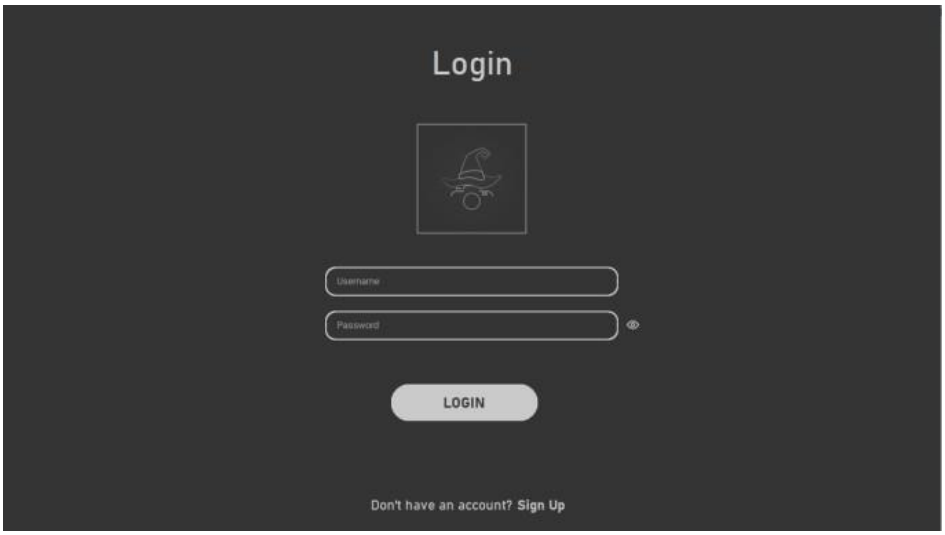

Figure 8. Login Page

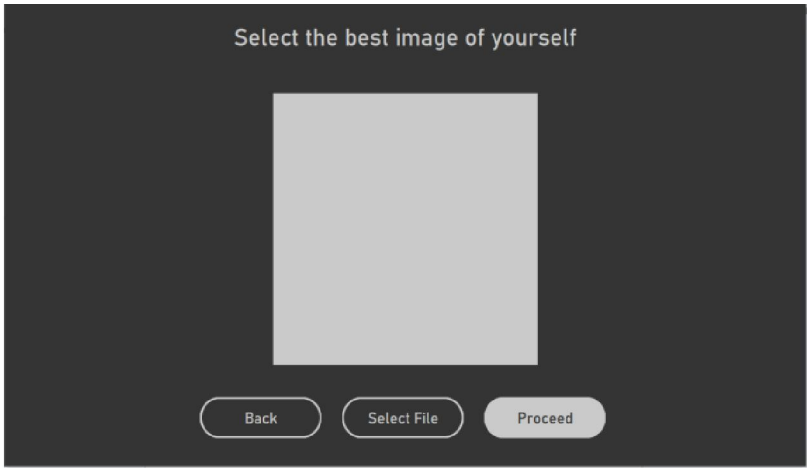

Figure 9. Select Image Page

Figure 10. Select Camera Page

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

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 (Arar et al., 2023).

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.

Figure12. 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

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 17, April 2026 Impact Factor: 8.2

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 (Li et al., 2023).

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.(Li et al., 2023). 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 (Liu & Zhu, 2022).

However, there is some variation in the responses, indicating that a few respondents still have suggestions for PotionPix(Radford et al., 2021). 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) (Rombach et al., 2022).

VII. CONCLUSION

This system, developed by Akash Sharma, is a major step forward for digital content creators. It saves time and helps small creators make professional-looking videos without hiring a designer 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 significant room for improvement. One place where improvement is needed involves using a dedicated AI model that has been trained for thumbnail creation. The SDXL model is general-purpose, and this can at times lead to subpar results. Creating a model with training focused on the specific visual demands of thumbnails—prominent faces, legible typography, and contrasting colors—would greatly increase output quality. Additionally, since the quality of the output images significantly depends on the input prompt provided by the user, constructing features that assist in prompt generation—such as intelligent suggestions or automatic improvements—would be a beneficial, especially for inexperienced users of prompt engineering techniques.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 17, April 2026 Impact Factor: 8.2

Future Work:

In the next version, we plan to add:

Voice-to-Thumbnail: Users can speak the title to create a design.

Auto-Translation: The system will automatically translate the thumbnail text into different languages.

REFERENCES

Arar, M. et al. (2023).
Domain-Agnostic Tuning-Encoder for Fast Personalization of Text-to-Image Models.
Liu, J., & Zhu, Y. (2022).
Precise Correspondence Enhanced GAN for Person Image Generation.
Radford, A. et al. (2021).
CLIP: Learning Transferable Visual Models From Natural Language Supervision.
Rombach, R. et al. (2022).
High-Resolution Image Synthesis with Latent Diffusion Models.
Saharia, C. et al. (2022).
Imagen: Photorealistic Text-to-Image Diffusion Models.
C. I. Lestari and I. Irwansyah, 'KolaborasiProduksiKonten 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.
B. Rahadjo, BelajarOtodidakMembuat Database menggunakan MySQL, 1st ed. Bandung: Informatika, 2011.[Online].Available: https://perpustakaan.binadarma.ac.id/opac/detail- opac?id=15408
https://doi.org/10.37030/jit.v4i2.74.
Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M. M., & Shan, Y. (2023).
PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding.
arXiv:2312.04461
Zhang, C. et al. (2024).
Text-to-Image Diffusion Models in Generative AI: A Survey.
arXiv:2303.07909
Esmaeili, S. A., Singh, B., & Davis, L. S. (2016).
Fast-AT: Fast Automatic Thumbnail Generation using Deep Neural Networks.
arXiv:1612.04811
Rochan, M., Reddy, M. K. K., & Wang, Y. (2020).
Sentence Guided Temporal Modulation for Dynamic Video Thumbnail Generation.
arXiv:2008.13362