COMS4995 Applied CV Project Report
Identity-Preserving Comic Story Generator

Xuezhen Wang, Yucheng Wang, Mingyu Deng

UNI: ww2604, yw3720, md4053

Abstract

Personalized text-to-image generation has emerged as a promising solution for creating customized digital content while preserving the identity of input individuals. This paper presents an efficient personalized text-to-image generation system that extends the StableIdentity method, which allows identity-consistent recontextualization with only one face image, to work with the Stable Diffusion XL model. Specifically, the method employs a face encoder with an identity prior, an editability prior, a two-phase diffusion loss, and Adaptive Instance Normalization (AdaIN) to generate high-quality, identity-preserved images. We also embed the system in the creative use case of personalized comic stories using Gradio. With uploaded face images, the user-input storyline is processed by Large Language Model (LLM) and input into the text-to-image model, which generates a personalized comic story versatile in style. Experimental results demonstrate the effectiveness of our approach, which has the potential to be applied in various domains where personalized digital content is in high demand.

1 Introduction

Personalized text-to-image generation systems have gained significant attention due to their ability to create customized images while preserving the identity of input individuals. These systems have a wide range of applications, including generating personalized portraits, creating realistic characters for movies or video games, and enhancing user experience in immersive virtual environments.

This paper presents an efficient personalized text-to-image generation system that leverages deep learning techniques to produce high-quality images while maintaining a strong resemblance to the input person's identity. Our approach extends the StableIdentity method proposed by Wang, et al.(15) to work with the more advanced Stable Diffusion XL model (10). We also develop an intuitive user interface using the Gradio library to facilitate easy creation of personalized comic stories. While Wang et al.'s work (15) only allows for image generation based on one identity, we extended the application to allow the generation of images with more than one face identity for the smooth creation of comic images.

The main contributions of this work are as follows:

- Implementing, improving, and extending the StableIdentity method (15) for identity-preserving image generation
- Integrating StableIdentity with the Stable Diffusion XL model (10) for improved image quality and diversity
- Creating an interactive user interface using the Gradio library to facilitate easy creation of personalized comic stories

The remainder of this paper is organized as follows: Section 2 provides an overview of related works in personalized image generation; Section 3 describes our methodology, including the StableIdentity model, its extension to Stable Diffusion XL, and the system pipeline; Section 4 presents the user interface design; Section 5 details our experiments and analysis; and Section 6 discusses limitations and future work.

2 Related Works

2.1 Text-to-Image Diffusion Models

Text-to-image diffusion models learn to map the distribution of text embeddings to the distribution of image embeddings, allowing for the generation of diverse and coherent images.

Stable Diffusion employs a latent diffusion model (14) to generate images from Gaussian noise and text prompts. The use of Contrastive Language-Image Pre-training (CLIP) (11) further enhances the alignment between text and image embeddings, and other methods aim to improve the controllability and flexibility of text-to-image generation (6) (2). To increase the resolution and quality of generated images, the Stable Diffusion XL model (10) (12) extends the capabilities of the basic Stable Diffusion model by utilizing a second text encoder and expanding the cross-attention context.

But since the training dataset usually contains lots of photos of celebrities and their corresponding names, these models can generate high quality images based on names of celebrity. To democratize image generation to the general public, advancements in the field have increasingly focused on the customized generation task (4) (16) (3).

2.2 Personalized Generation

For the reason above, personalized image generation, particularly in preserving facial identity, has garnered significant attention lately. Existing approaches can be categorized into two main categories: two-phase methods that involve optimization and fine-tuning, and single-pass methods that bypass the fine-tuning process (15).

Two-phase methods, such as DreamBooth (13) and Textual Inversion (5), involve an initial training phase followed by a fine-tuning phase during testing. While the results are promising, these approaches can be computationally expensive and time-consuming. On the other hand, single-pass methods aim to eliminate the fine-tuning process by either creating personalized training datasets or encoding images into a semantic space for customization (9). These approaches offer more efficient

inference but may sacrifice some level of identity preservation.

Our work builds upon StableIdentity (15), which introduces several key innovations to address the challenges of personalized image generation. StableIdentity employs a face recognition encoder (FR-ViT) to extract identity-rich representations from input face images, an editability prior, a two-phase diffusion loss, and Adaptive Instance Normalization (AdaIN) to align the identity embeddings with celebrity name embeddings.

We extend StableIdentity to work with the more advanced Stable Diffusion XL model (10), which generates higher-quality and more diverse images compared to its predecessor. This extension involves modifying the model architecture and recalculating the editability prior. Our work also draws inspiration from Li et al. (8), who propose a method for identity-preserved image generation by reorganizing the pipeline and overriding certain functions. While we encountered challenges in replicating their results, their approach influenced our decision to create a more modular and reusable pipeline.

Our contribution lies in the successful integration of StableIdentity with Stable Diffusion XL and the development of an intuitive user interface for creating personalized comic stories. By combining state-of-the-art techniques with an accessible interface, we aim to make this technology more democratized and applicable to various domains.

3 Method Overview

3.1 Model

Our application is based on the method proposed in the StableIdentity paper by Wang, etc. (15). The method allows identity-consistent recontextualization with one-time training of just one face image and the learned features can be reused in various scenarios. In our application, we implemented the pipeline for StableIdentity and adopted the method to a larger pretrained stable diffusion model.

3.1.1 StableIdentity

At the core of the StableIdentity framework is the FR-ViT, a specialized face recognition encoder that processes the input face image to extract a representation rich in identity features. This component is fine-tuned specifically for face recognition tasks, enabling it to capture nuanced aspects of individual identity, such as facial structure, expression, and possibly even subtle features that distinguish one person from another.

After the identity representation is obtained via FR-ViT, the model employs MLPs to project this identity into a new space that accounts for the flexibility of editing. This is referred to as the editability prior, which derives from the semantic understanding that celebrity names, used during the training of large text-to-image models, come with a rich context that can be adapted to various scenarios. By anchoring the identity representation in this editability space, the model gains the ability to customize images while maintaining the core identity features consistent across different edits.

The two-phase diffusion loss is a novel contribution to the training regime. The masked diffusion loss targets only the salient parts of the image, such as the face and hair, ensuring that these identity-carrying regions are preserved across different generations. This masking helps the model to focus on these areas and avoid irrelevant background information that does not contribute to identity preservation.

The training process itself is divided into two temporal phases, each with its own objective:

- In the early phase of denoising, the model's objective is to ensure that the basic layout and context of the image are conducive to diverse contextual embeddings. Here, the model learns to adapt the core identity features to various scenarios, essentially teaching the identity to be flexible.
- In the later phase, as the image becomes clearer and noise is reduced, the emphasis shifts to pixel-level detail. This is where the model fine-tunes the identity features to achieve a high-fidelity reproduction of the original input face.

The implementation of AdaIN is critical in bridging the gap between the extracted identity features and the celeb embedding space. AdaIN is employed to adjust the distribution of the identity embeddings so that they align with the distribution of the celebrity name embeddings. By doing this, the identity features are not only maintained but are also made adaptable to the kinds of transformations and contexts that celebrities' images would typically undergo in the text-to-image model. This ensures that the personalized images generated are not only identity-consistent but also versatile in terms of the edits they can undergo.

Ultimately, the integration of the identity prior (knowledge from face recognition) with the editability prior (the versatility of celebrity contexts) allows StableIdentity to maintain a subject's recognizable features while manipulating the image in various ways, whether it be changing the attire, setting, or even artistic style of the image, without losing the essence of the individual's identity.

3.1.2 Extending to Stable Diffusion XL and multiple IDs

The StableIdentity paper implements their methods on the Stable Diffusion 2.1-base. In our application, we retrain the model and extend the method to a larger pretrained stable diffusion model, Stable Diffusion XL-base-1.0 (SDXL) (10). During testing, although the original model of SD 2.1-base (SD2.1) captures human face identity well, the quality of the generated image is highly unstable. It struggles to illustrate complex scenes and a wide range of styles, with the style of Stable Diffusion 2.1 also being unsatisfactory. As the application aims to generate a comic story, which is a series of images around the same topic, it is challenging for the model to meet a reasonable standard. According to the evaluation conducted by the Stable Diffusion XL model's paper, SD 2.1 accounts for only 3.42% of user preference in the evaluation, while SDXL 1.0 (base and refiner) accounts for a 26.2% share among the comparison of six models.

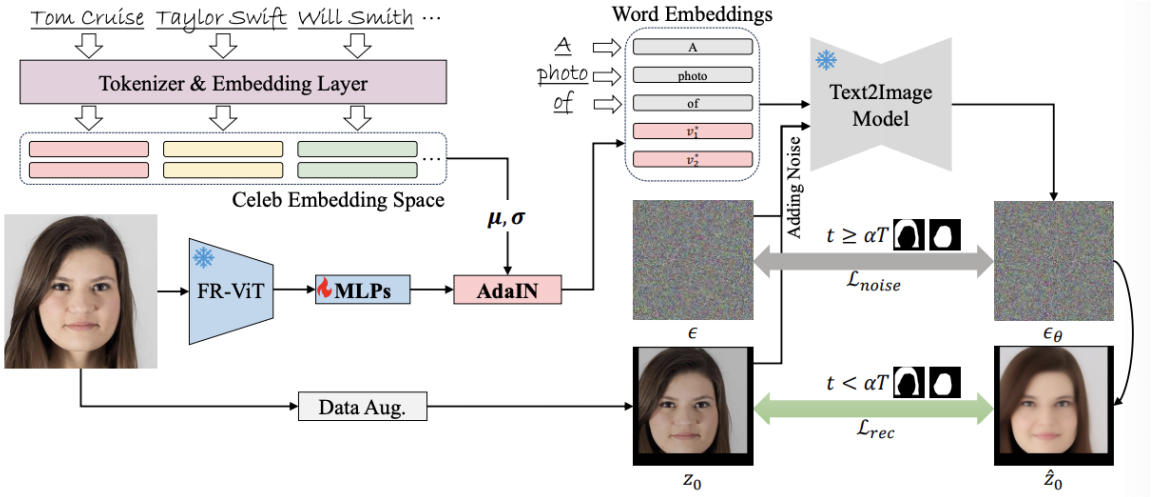

Figure 1: StableIdentity Model Pipeline

Stable Diffusion 2.1-base employs the CLIPTokenizer and CLIPTextModel (clip-vit-large-patch14) to process text (11), with a token dimension of 1024. The structure of Stable Diffusion XL differs from that of the basic Stable Diffusion model. SDXL utilizes a second text encoder to allow for more attention blocks and an expanded cross-attention context, leading to an increase in model parameters (10). In the specific model checkpoint Stable Diffusion XL-base-1.0, it uses two separate text tokenizers and text encoders, CLIPTextModel and CLIPTextModelWithProjection, with corresponding token dimensions of 768 and 1280. To accommodate this modification, we implement an additional set of multilayer perceptrons (MLPs) to insert the image embedding into the text embedding from the second text encoders. Concurrently, the editability prior is recalculated for each encoder respectively.

Building on the original capabilities of the StableIdentity model, which was primarily demonstrated with single identity image generation, we have expanded the model's functionality to support multiple identities within a single image while preserving each individual's distinct identity. The extension of this method is straightforward yet effective. Initially, the model utilized a single pair of signal words (e.g., `v1*`, `v2*`) to replace with trained personal identity embeddings. In our enhanced implementation for inference, we introduce multiple signal words (e.g., `v3*`, `v4*`) into the prompts, corresponding to the number of individuals depicted in the image. Each pair of signal words is directly associated with a unique identity embedding. This approach has enabled us to successfully generate images that maintain multiple distinct identities simultaneously, marking a significant advancement in the model's capability.

3.1.3 Failed Attempts

The original method delineates a clear distinction between the training and inference stages. During training, the original paper describes a unique forward pass pipeline that incorporates all the necessary models from the StableDiffusionPipeline, yet remains distinct from the pipeline's standard forward pass. This approach results in a training script that is isolated, hindering reusability. Additionally, the trained model is tailored to a specific identity and, as such, does not lend itself to reuse; only the identity embedding, which is generated via a single forward pass based on the input image, proves to be reusable. In the inference stage, this identity embedding can be loaded to replace specific token embeddings within a freshly loaded pre-trained model, a process again distinct from the training phase. In our endeavor, we aimed to reorganize the overall pipeline, taking cues from the original StableDiffusionPipeline or StableDiffusionXLPipeline, and overriding certain functions within the pipeline. Our code structure draws inspiration from another paper by Li et al. (8), which also focuses on identity-preserved image generation.

Despite these efforts, our modified model did not manage to replicate the desired outcomes achieved by the former implementation. This shortfall could be attributed to some overlooked predefined behaviors within the original stable diffusion pipeline. Currently, this issue remains unresolved.

3.2 System Pipeline

We systematically demonstrate the proposed system, which is designed to generate personalized images using a combination of machine learning models and processing of the user inputs, in the following Figure 2. The process encompasses several phases:

Phase 1: Training the Personalization Model. Initially, the system ingests a set of face images paired with corresponding full names (the reason that they must and only contain the first and the last name is to be made clear later). This data is utilized to individually train the model for each pair (see Section 3.1). The training yields a unique embedding for each image-name pair that will be stored in a "database," capturing facial feature representations without the reliance on extensive datasets—this is a single-instance training approach per individual.

Phase 2: Story Guideline Interpretation. Users provide a narrative guideline detailing the characters and rough ideas, which must correspond to the names used in the initial training phase. For example:

> *Hugh Jackman was selecting an apple at a local grocery store when Taylor Swift, who*

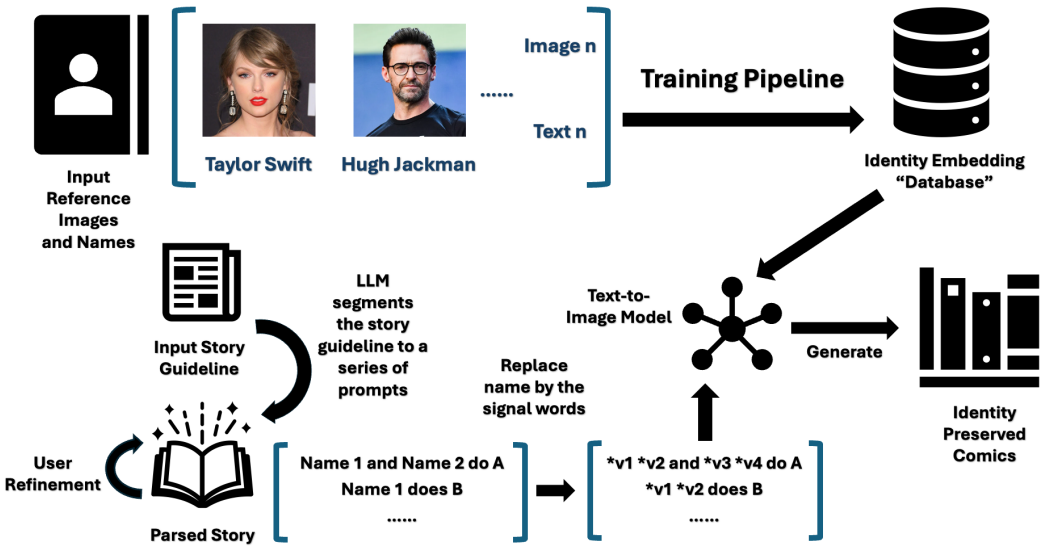

Figure 2: System Pipeline

dislikes apples, reached for a banana nearby. They struck up a conversation and discovered they were both headed to the same school. After paying for their purchases, they walked to school together, chatting amiably about their day.

An LLM interface processes the narrative by breaking it down into individual sentences and ensuring that each instance includes the full name (e.g., "Taylor Swift" instead of "Taylor" or any pronouns). We will submit the narrative guidelines alongside a prompt that specifies the sentences should be simple and adhere closely to the development outlined in the guidelines, specifically avoiding the use of pronouns. We have evaluated several free LLM APIs, such as Mistral 7B, Cohere, and Eden AI, and empirically found that Gemini offers the most effective parsing. A typical resulting narrative processed by the LLM API will be similar to the one as follows:

- *Hugh Jackman was selecting an apple at a local grocery store when Taylor Swift reached for a banana nearby.*
- *Hugh Jackman and Taylor Swift struck up a conversation and discovered they were both headed to the same school.*
- *Hugh Jackman and Taylor Swift walked to school together, chatting amiably about their day.*

Notice that all the pronouns are replaced by the full names of the people and are grouped into a list of sentences. The processed narrative is presented back to the users via a web interface, allowing for modifications and corrections to align with the user's intentions.

Phase 3: Embedding Integration. In this phase, the system substitutes actual names within the story with designated signal words (e.g., "v1* v2*"). As an example, the previous sentences become:

- *v1* v2* was selecting an apple at a local grocery store when v3* v4* reached for a banana nearby.*
- *v1* v2* and v3* v4* struck up a conversation and discovered they were both headed to the same school.*
- *v1* v2* and v3* v4* walked to school together, chatting amiably about their day.*

where $v1*$, $v2*$, $v3*$ and $v4*$ stand for Hugh, Jackman, Taylor and Swift, respectively. We change the original encoder's embedding for the signal words in the pre-trained text-to-image model into the trained embeddings we obtain in Phase 1. During the inference stage, these placeholders are then mapped to their respective embeddings within the text-to-image model, enabling the integration of personalized elements into the generated images.

Phase 4: Image Synthesis. Finally, a series of sentences in the parsed and annotated story are input into the modified pre-trained text-to-image model, specifically a diffusion model in this instance, which synthesizes a sequence of images which can be stacked as a comic story for the original story guideline. Inherently, The system offers versatility in output styles, ranging from comic to cinematic or painting-like visuals, which can be selected by appending stylistic keywords to the input prompt.

Our approach delineates a novel method for creating identity-preserved visual narratives, offering high customization and ensuring that the generated images retain the personalized features of the characters as specified by the users.

4 User Interface

Gradio is an open-source Python library that simplifies the creation of user interfaces for machine learning models (1). It is especially useful for developers and researchers looking to swiftly prototype and share their models with others. In the context of our system, Gradio serves as the interface layer that allows users to interact with the personalized image generation pipeline. Through the Gradio interface, users can upload headshot

images, enter and refine their story guidelines, and select desired image styles, thus making the complex process of generating personalized images both accessible and user-friendly.

4.1 Identity Training and Embedding Storage

We provide a user interface that enables users to upload headshots for identity preservation in the generated comics. Users may upload multiple image-name pairs into the system. It is imperative that the sequence of uploaded images corresponds exactly with the sequence of full names entered into the designated text box. Due to system constraints, only headshots with a resolution of 512×512 pixels are supported. Additionally, the list of names must follow the format: *[FirstName1 LastName1, FirstName2 LastName2, FirstName3 LastName3, ...]*, as shown in Figure 4. Clicking the "Upload" button will initiate the training of models in the background which takes usually about 5∼10 minutes on a customer GPU and store the embeddings for each identity after training.

After storing the trained embeddings, the interface will display the available identities along with their corresponding names, as illustrated in Figure 6. Additionally, we provide several pre-trained embeddings, shown in Figure 5, complete with their corresponding images and names. Users can directly click on these to add them to the trained embedding display, as shown in Figure 6.

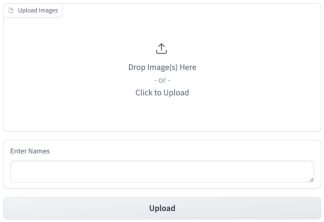

Figure 3: Empty Uploader

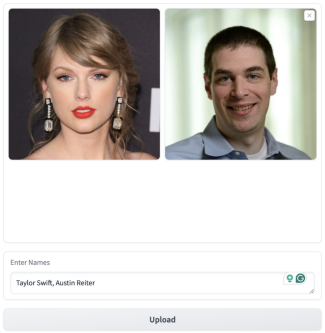

Figure 4: Uploader with Chosen Images

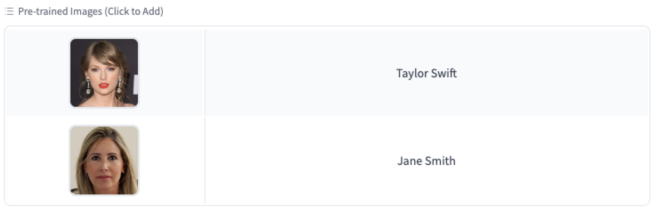

Figure 5: Displayed Pre-Trained Embeddings

4.2 Parsed Story Generation

We provide users with an interface, as depicted in Figure 7, to input their story guidelines, which must include the full name of the trained identities at least once. By clicking "Process Text," the LLM API parses the story guideline and displays a list of sentences as the result. Given the unpredictability of the LLM model, users are allowed to modify the results directly based on their needs. For example, they may replace a pronoun that the LLM ignored with a full name. Note that **Phase 3**, as mentioned in Section 3.2—which involves replacing the full name with signal words—is handled in the backend to simplify the user experience.

As previously mentioned, we provide an interface for users to select the style of the images to be generated. Options include "2D minimalistic," "8K," "cartoon," "chiaroscuro lighting technique," "cinematic," "painting," among others.

4.3 Comic Story Generation and Display

After obtaining the trained embeddings and the parsed story, clicking the "Generate Images" button will activate the modified pre-trained diffusion pipeline to generate comic stories and the corresponding text for each image, as shown in Figure 8.

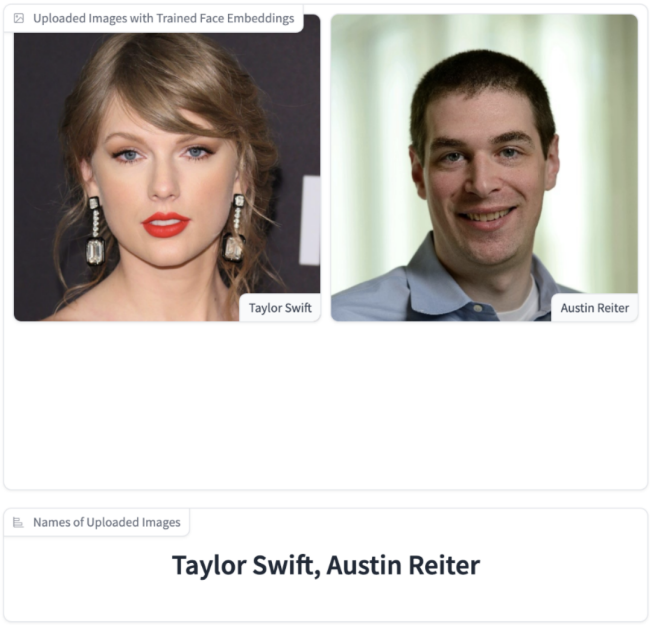

Figure 6: Displayed Trained Embeddings and Names

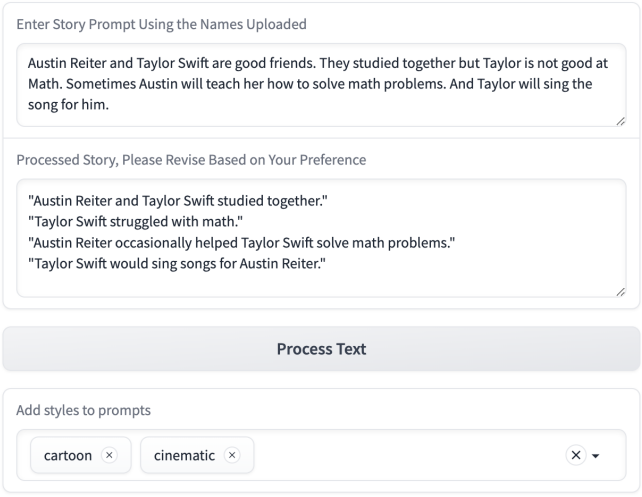

Figure 7: LLM Story Parser

Figure 8: Comic Story Generation and Display

5 Experiments and Analysis

5.1 Setup

We conduct our experiments on one single NVIDIA L4 Tensor Core GPU. For the detailed setup, we basically followed the setup in the original paper. The batch size remains to be 1 but we choose the learning rate to be $1e-5$ instead of $5e-5$. The MLPs are trained for 500 steps while training the image embedding. The scale of classifier-free guidance (7)is set to 8.5 by default.

5.2 Evaluation

Evaluation of the Training. Though the pipeline based on StableDiffusion XL entails training more parameters, the training process can still converge swiftly, within 500 steps. As previously mentioned, the training utilizes a two-phase diffusion loss, which divides the loss calculation into two parts at timestep αT. The larger the α is, the greater the portion of data evaluated by $\mathcal{L}_{rec}$. Since approximating the true distribution of the denoised prediction becomes more challenging at larger timesteps, the original paper suggests a range for the division parameter $\alpha \in [0.4, 0.6]$ to yield good results that balance identity preservation and diversity. In the original implementation, $\alpha = 0.6$ is set as the default. However, in our implementation with StableDiffusion XL, we found it is even more challenging to approximate the distribution of the denoised prediction, prompting us to choose a slightly smaller α. In the code provided, we select $\alpha = 0.5$ as the default.

Qualitative Analysis of the Generated Images As shown in Figure 9, we conduct a comparative analysis between the generated images of two models using seven different prompts based on Taylor Swift's face. All prompts specify close-up portraits to focus the analysis on facial features, facilitating a direct comparison. Our evaluation reveals significant limitations with the original model, StableDiffusion 2.1, which often fails to interpret prompts accurately. For instance, the prompts "wearing a space suit" and "wearing a doctoral cap" consistently result in incorrect scenarios, despite multiple trials. Similarly, other prompts also yield incorrect images on initial attempts. In contrast, our new model based on the pre-trained StableDiffusion XL consistently delivers higher quality and more stable results. The facial identity is largely preserved, and the prompts are clearly expressed. Moreover, images generated by StableDiffusion XL demonstrate greater diversity, whereas those from StableDiffusion 2.1 often exhibit repetitive expressions.

In addition to the close-up portrait analysis, we also compare the two models across various scenarios in Figure 10. These prompts extend beyond facial focus to include more dynamic body gestures and complex background environments. The results indicate that StableDiffusion 2.1 struggles with these tasks, often overlooking significant elements of the prompts such as "teaching in the class" and "winning a Nobel prize." Furthermore, even when actions are correctly represented, like in "playing basketball," the facial features tend to distort. On the other hand, StableDiffusion XL again outperforms by producing more natural body gestures and maintaining consistent facial expressions. The backgrounds also better complement the subjects, enhancing the overall realism of the images.

Evaluation of the Application. We distribute simple questionnaires to our friends at Columbia, which contain the following statements:

1. The Gradio interface was intuitive and easy to navigate.
2. I was able to upload images and enter story guidelines without any difficulties.
3. The system accurately reflected the personalization of the characters in the generated images.
4. The image generation process met my expectations in terms of style and quality.
5. I am satisfied with the speed of the image generation process.

Participants were asked to respond by choosing one of the following options: Strongly Agree, Agree, Neutral, Disagree, and Strongly Disagree. Due to time constraints, we were only able to collect 78 responses from them. The collected results are shown below in Figure 11.

From the feedback collected, it is evident that the majority of responses are favorable. Nevertheless, it is significant that most participants opted for "Agree" rather than "Strongly Agree." This distinction suggests that while users found the interface and functionality generally satisfactory, there is ample room for improvement to elevate user satisfaction to a higher level of enthusiasm. Further analysis reveals several potential areas for enhancement, such as simplifying the user interface for even smoother navigation and boosting the speed and quality of image generation. Additionally, despite upgrading our model from Stable Diffusion 2.1-base to

Figure 9: Comparative visualization of generated images using SD2.1 and SDXL models based on varying prompts focusing on closeup portrait.

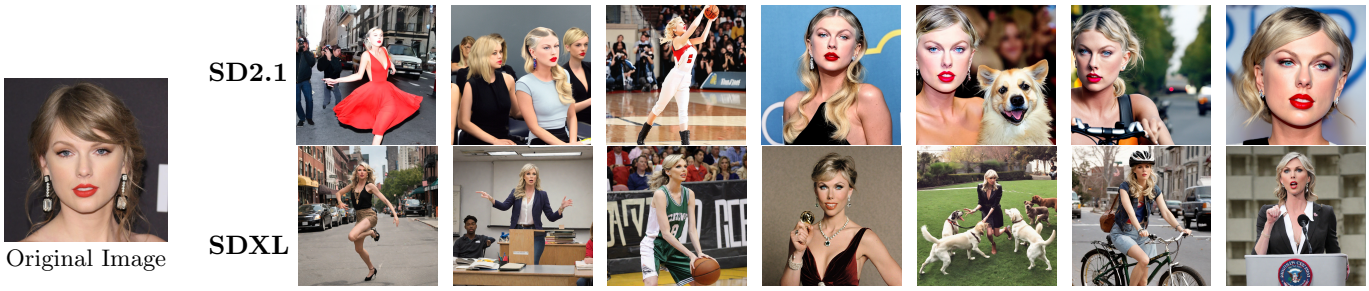

Figure 10: Comparative visualization of generated images using SD2.1 and SDXL models based on varying prompts.

Stable Diffusion XL, the variability in image quality remains an issue, with users frequently encountering distorted images. This indicates that our system might benefit from more refined prompt engineering or a more advanced text-to-image model. We also observed that most users were more likely to vote for "Agree" instead of "Strongly Agree" regarding the preservation of identity. This outcome is partly due to limited time spent on tuning hyperparameters within our project, resulting in occasional dissatisfaction with how identities were maintained. We anticipate that fine-tuning these parameters would enhance the system's ability to preserve identity more accurately. Implementing these improvements could potentially shift user perceptions from mere agreement to strong agreement with the system's performance. Addressing these aspects could lead to more robust and enthusiastic user approval in future evaluations.

Comparison between SD2.1 and SDXL. We also include results generated by using the original Stable Diffusion 2.1 (SD2.1), as described in the original paper, alongside our modified Stable Diffusion XL (SDXL) in our system. We find that all 78 participants favored the comics generated by the SDXL model.

6 Discussions

6.1 Limitation

Cultural Representation Bias. Our model exhibits varying performance levels across different ethnicities, showing a particularly limited capability in accurately generating Asian faces compared to Caucasian faces. In some instances, features typically associated with Asian individuals are not preserved, resulting in outputs that disproportionately resemble Caucasian features. This is partially due to the hyperparameter α, which regulates the balance between identity preservation and diversity, as discussed in Section 5.2. A higher α value enhances diversity but tends to dilute ethnic-specific features, leading to unsatisfactory representations of Asian identities.

Content Safety Concerns. The model also presents potential content safety concerns. Specifically, it lacks robust controls against generating sensitive or potentially harmful content, such as explicit imagery. This susceptibility could be exploited to generate inappropriate images involving specific individuals' likenesses without their consent, posing significant ethical and privacy issues. Specifically, we have tested the generation of images that preserve identity using prompts that include the word "naked." This testing resulted in images depicting the person in a state of undress, which poses significant ethical and privacy concerns. Addressing these concerns is crucial to prevent misuse and ensure the responsible deployment of our technology.

7 Conclusions

In conclusion, this project aims to contribute to the field of personalized digital content creation by developing a robust identity-preserved comic generator. Utilizing the StableIdentity method, extended to operate with the more powerful Stable Diffusion XL model, our system has set new benchmarks in generating high-quality, identity-consistent comic stories. Furthermore, while the original work is limited to generating single identities, we have discovered that the method can be naturally extended to generate images with multiple preserved identities. This advancement allows us to apply the technique to the task of creating comic stories with consistent character representations. The intuitive Gradio user interface significantly enhances accessibility, allowing users to

effortlessly engage with advanced text-to-image generation technology which is trainable using customer GPU in a short time. Our user-study experiments confirm the system's effectiveness, demonstrating its potential across various applications where identity preservation is critical. Despite encountering challenges related to cultural representation biases and content safety, the foundation established here provides a strong basis for future improvements. The ongoing development of this project will focus on enhancing the accuracy of identity preservation, expanding its applicative reach and addressing the ethnic issues, ensuring it can serve a more diverse and inclusive range of users.

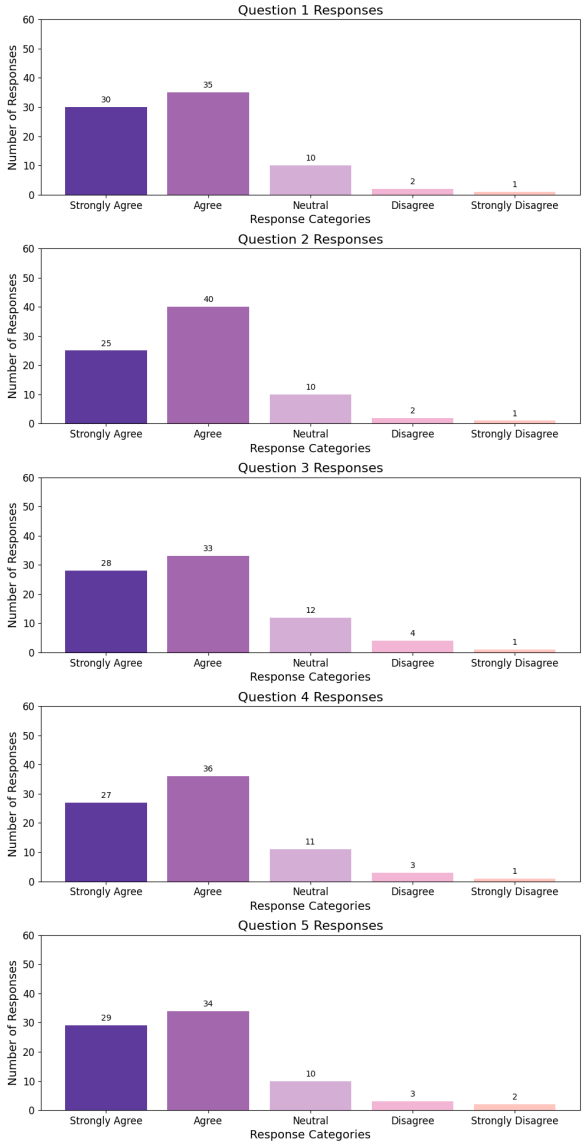

Figure 11: User Feedback Results

8 Github Repo

We provide the link to our Github repo (`https://github.com/DELTA-DoubleWise/Identity_Preserved_Images_Generation`) which contains the main implementation, commit history and code reviews (see PRs) for this project.

References

[1] Abid, A., Abdalla, A., Abid, A., Khan, D., Alfozan, A., and Zou, J. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569* (2019).

[2] Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., and Lischinski, D. Break-a-scene: Extracting multiple concepts from a single image. *arXiv preprint arXiv:2305.16311* (2023).

[3] Chen, L., Zhao, M., Liu, Y., Ding, M., Song, Y., Wang, S., Wang, X., Yang, H., Liu, J., Du, K., et al. Photoverse: Tuning-free image customization with text-to-image diffusion models. *arXiv preprint arXiv:2309.05793* (2023).

[4] Chen, Z., Fang, S., Liu, W., He, Q., Huang, M., Zhang, Y., and Mao, Z. Dreamidentity: Improved editability for efficient faceidentity preserved image generation. *arXiv preprint arXiv:2307.00300* (2023).

[5] Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

[6] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).

[7] Ho, J., and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[8] Li, Z., Cao, M., Wang, X., Qi, Z., Cheng, M.-M., and Shan, Y. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461* (2023).

[9] Nitzan, Y., Aberman, K., He, Q., Liba, O., Yarom, M., Gandelsman, Y., Mosseri, I., Pritch, Y., and Cohen-Or, D. Mystyle: A personalized generative prior. *arXiv preprint arXiv:2301.01319* (2023).

[10] Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[11] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision, 2021.

[12] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695.

[13] Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 22500–22510.

[14] Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[15] Wang, Q., Jia, X., Li, X., Li, T., Ma, L., Zhuge, Y., and Lu, H. Stableidentity: Inserting anybody into anywhere at first sight. *arXiv preprint arXiv:2401.15975* (2024).

[16] Yuan, G., Cun, X., Zhang, Y., Li, M., Qi, C., Wang, X., Shan, Y., and Zheng, H. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926* (2023), 1–4, 10.