Armenia's First Virtual Influencer
Spring 2024

Author: Ofelya Stepanyan
American University of Armenia
BS in Data Science

Author: Natela Azoyan
American University of Armenia
BS in Data Science

Supervisor: Elen Vardanyan
American University of Armenia

***Abstract*—In an era where virtual personas are getting popular in social spaces daily, creating virtual influencers and bloggers through AI changes our perspective of how we engage with online platforms and content. This paper presents the development of an Armenian virtual influencer - Sahmi, with the integration of different AI tools, who can automatically share various types of content, such as fashion trends, travel experiences, etc, on Instagram. With the help of the *Realistic_Vision_V2.0* model of Stable Diffusion, deep fake technology, automatic captioning, and automatic Instagram posting, we have created a virtual influencer that looks realistic and can preserve its identity in different scenarios. It was possible to achieve this with the *DreamBooth-fine-tuning* method. During our research, we explored different models and techniques to achieve the best possible outcome with limited resources integrated. Our research also discusses how AI can be applied in the future in social media and digital communication.**

I. Introduction

Can you imagine a virtual influencer from Armenia who shares interesting places inside and outside of Armenia, the latest fashion trends, interacts with her followers using videos, writes captions, and posts all by herself? This scenario, which seems impossible to do, is now a reality with the help of AI-driven technologies. We use the *Realistic_Vision_V2.0* model of Stable Diffusion to determine the influencer's authenticity [1], [2]. Moreover, we use a deepfaking model to integrate the face of the influencer in different videos. Additionally, we are also using a model that does automatic captioning based on the image.

By harnessing the latent potential of text-to-image synthesis, our system can produce high-quality images of the virtual influencer with different poses, scenes, clothing, etc. The virtual influencer has a name, personality, and story, all presented on her Instagram account. In order to have minimal human oversight, postings, and captioning are being done automatically.

One of our biggest problems was having the same identity every time we generated a person. This was possible with the help of a model called *DreamBooth* [4], which is a fine-tuning technique that takes several images and a unique identifier as an input, extracts the features of the input, and, based on that, generates images with different positions, different places, etc. by keeping the identity of the subject/human. We have also created videos using deep faking [7]–[9] techniques to make her more interactive with her followers.

To bring the virtual persona to life, the following steps are done:

- Image generation
- Caption generation
- Caption translation to Armenian
- Video deepfaking
- Automated image/video posting with a caption on Instagram.

This project aims to demonstrate how a combination of current multimodal AI tools can allow us to flexibly create an even more complex model, automating every step of the way.

Having multiple AI technologies in one place following each other might be challenging. Our paper explores these challenges and shows how, with the help of AI, we create a new human being influencer who will be able to do anything she wants on social media.

II. Related Work

A. Image composition

Image composition, used mainly in 3D reconstruction techniques [5], [6], merges a given subject into a chosen background to create a whole image. 3D techniques require a larger number of views and work on nonflexible objects. The drawbacks of such techniques are that it is challenging to create integration between objects because of the lighting and shadows, eventually leading to unrealistic images. Another problem is that these kinds of models cannot make totally new scenarios or go beyond the existing data it has. In contrast, our approach allows the generation of images in entirely new contexts, poses, and places.

B. Controllable generative models

During the recent years, controllable generative models have undergone significant advancements in subject driven and prompt guided image generations [11], [12], where control mechanisms were integrated in order to generate images in a more accurate way. For example, diffusion-based models leverage text/reference images to lead the variations of the images. There are methods that provide user-specified masks, in order to control the modifications of the areas that are not selected [13]. Other techniques focus on utilizing latent space of generative models to keep their main identity while changing the other contexts [14]. Last but not least, some

methods also offer a new approach to image editing, which is mainly done by a prompt-to-prompt method. It helps to do not only local but also global changes without defining the input masks [15].

C. Text-to-Image Editing and Synthesis.

Recent advances in text-to-image synthesis have been impressive, especially with the introduction of *Generative Adversarial Networks (GANs)* [16] in concurrence with image-text representations such as *Contrastive Language-Image Pretraining (CLIP)* [17], [19]. These developments have made text input manipulations more realistic. Although these algorithms perform well in well-defined settings, such as human face editing, they may perform poorly in heterogeneous datasets with a wide range of subjects. Researchers have used methods like *Vector Quantized Generative Adversarial Networks (VQ-GAN)* [20] and training across a wider variety of data to increase performance and solve this difficulty.

Diffusion models, which achieve state-of-the-art generation [16] quality over extremely heterogeneous datasets, have also emerged as a viable method. These models allow for the creation of high-quality images and frequently outperform *GANs*. Only some current techniques allow for unique interpretations of a given subject in different situations. At the same time, the majority of systems concentrate on either global editing or localized editing depending on text input.

Large-scale models like *Imagen* [21], *DALL-E2* [14], *Parti* [22], *CogView2* [23], and *Stable Diffusion* [1] have recently made significant progress in text-to-image synthesis. These models provide previously unheard-of capabilities for semantic generation. However, they only use text advice and have no fine-grained control over the generated images [18]. One specific difficulty is keeping a subject's identity intact among synthesized photos, which is still a major area needing development.

Text-to-image synthesis has advanced remarkably, with several methods proving they can produce lifelike images from written descriptions. More research is necessary to improve control and fidelity in creating photos that accurately depict the desired individuals and surroundings.

D. Image to Text Generation

Significant progress has been made in the image-to-text generation field, especially by combining *GANs* with models such as *CLIP* [10], [24], [26]. By combining the creative power of *GANs* with the multimodal qualities of *CLIP*, this hybrid technique improves the semantic relevance of the generated visuals to the input text. The capacity to produce diverse and artistically rich images is demonstrated by innovations like *BigSleep* and *VQ-GAN+CLIP*, which maximize the *CLIP* score in the latent space of the *GAN* [25], [26]. Moreover, the FuseDream framework improves upon this paradigm by offering methods for effectively navigating the nonconvex optimization landscapes characteristic of these models. This helps to reduce frequent data biases and improves image quality by using methods such as AugCLIP [25]. These developments highlight the field's quick evolution, pushing the envelope regarding image synthesis from textual descriptions while tackling the inherent difficulties associated with model training and image variety [26].

E. Deep Faking Video Generation

One of the ways to create interactive videos was lip-syncing. There were only a few open-source models, and the comparably best one among them was the *Wav2Lip*.

The primary purpose of *Wav2Lip* is to sync lip movements with audio in video content. It generates realistic lip movements by processing audio inputs and using a pre-trained model with video frames. The pre-trained model that *Wav2Lip* utilizes is *SyncNet* [33], which mainly detects whether the video and audio streams are in sync. To ensure the lip matches the audio, the model combines a lip-sync discriminator with a visual quality discriminator. While the lip-sync discriminator ensures the timing and dynamics of the lip movements match the spoken words, the visual quality discriminator concentrates on the appearance. This technique makes the video more realistic.

Wav2Lip architecture consists of a generator and two different discriminators. The generator comprises an Audio Encoder, Face Encoder, and Face Decoder, mainly developed by several blocks of convolution layers.

As the *Wav2Lip* did not give good results, we tried the video deep faking method. With advanced artificial intelligence techniques, video deepfaking creates or manipulates video content that accurately mimics the appearance and movements of actual people [7].*GANs* and *autoencoder-based* techniques, which effectively alter and synthesize video frames to achieve incredibly realistic results, are essential to deepfaking.

A generator and a discriminator [8] are the two primary parts of *GANs* used for video deepfaking. The discriminator assesses the validity of the video frames produced by the generator, which imitate the appearance of a target person [7], [8]. The generated content's realism is increased due to the generator's iterative training [9], which increases its capacity to make frames that the discriminator cannot discern from actual videos.

In video deepfaking, autoencoders are specifically used for face-swapping tasks [7]. They employ a dual structure in which separate autoencoders are trained to compress and decompress facial data belonging to distinct people. To create videos where the original person's face is replaced with another while maintaining their natural expressions and movements, the encoded aspects of one person's face are fed into the decoder of another during the generation process [7], [8]. This technique is frequently applied in fields ranging from entertainment to producing dangerous content, needing continuous improvements in detection techniques to reduce the potential of misuse of this advantages [9].

III. Methods

Our objective is to generate a person with a given prompt that didn't exist before and automate its functioning as an

influencer guided by caption generation and video generation. There were no restrictions for the prompts, and images can vary with different poses, backgrounds, colors, shapes, locations, etc.

We have used text-to-image diffusion models fine-tuned by the technique called *DreamBooth* which allowed us to personalize text-to-image models (Sec. III-A), then we have done image-to-text generation using *Large Language-and-Vision Assistant(LLaVA)* in order to do captioning for our influencer's post (Sec. III-B). Additionally we have done deepfaking in order to solve the problem of good-quality videos with a technique called Roop (Sec. III-C). In the end in order to automate instagram content posting we have used Selenium (Sec. III-D).

A. Text-to-Image Diffusion Models

Diffusion models are a type of generative model that generate images by gradually adding a noisy image into a clear one. The process involves two main parts: the forward and reverse processes [1], [2]. During the forward process, in the clear image a noise is gradually added step-by-step (from Gaussian Distribution), transforming it into a total noise by the end of the process. The forward process can be described with the following formula:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \tag{1}$$

$$\epsilon_t \sim N(0, I) \tag{2}$$

where β_t is a noise schedule from (0,1).

In the reverse process, where the actual image is generated, the model attempts to recover the original image from the noise, step-by-step, guided by a neural network trained to predict the noise that was added at each step. This process can be described the following way:

$$x_0^{\text{pred}} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \sqrt{1-\alpha_t}E_0(x_t)\right) \tag{3}$$

$$x_{t-1} = \sqrt{\alpha_{t-1}}x_0^{\text{pred}} + \sqrt{1-\alpha_{t-1}}\epsilon_{t-1} \tag{4}$$

Where $\alpha_t = \prod_{i \le t}(1-\beta_i)$ and the denoising network $E_0(x_t)$ approximates E_t , improving the clarity with each backward step. The main objective of the training is to minimize the difference between the added noise and the predicted noise.

Hence the training objective of $\mathcal{E}_\theta(x_t)$ is:

$$L = \mathbb{E}_{t,\epsilon}\left\|\mathcal{E}_\theta\left(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon\right) - \epsilon\right\| \to \min_\theta$$

The architecture of the Stable Diffusion consists of Encoder-Decoder, Condition mechanisms and transformer blocks which are shown in 1.

B. Personalization of Text-to-Image Diffusion Models using Dreambooth

We have employed a fine-tuning model called *DreamBooth*, which utilizes a pre-trained generative model on a small number of images (typically around 3-5) of a specific subject, such as a particular person, pet, or object [4]. This process incorporates a unique identifier for each subject. The model operates with few-shot personalization; it receives a few images of a subject, each labeled with a simple text prompt containing a unique identifier and a class descriptor (e.g., "a [unique identifier] girl"). This allows the model to bind the identifier with the subject, leveraging its prior knowledge of the class (like "girl") while focusing on the specific instance. To minimize language drift and help the model concentrate on the unique features of the subject, *DreamBooth* uses rare tokens as identifiers. These tokens are uncommon words or character sequences that do not carry strong pre-existing associations in the model, making them more effective as unique identifiers. The unique identifier/rare token identifier is usually composed of up to three unique English letters (e.g., "zwx") [4].

DreamBooth adds a class-specific prior preservation loss to solve the model's possible loss of language drift and preserve the model's capacity to produce a variety of images from the larger class [4]. This loss function helps the model to keep producing different examples from the class.

The fine-tuning process 2 employs a diffusion loss formula that helps integrate the subject-specific training into the model's broader capabilities without overfitting or forgetting its existing knowledge base.

We generate data $x_{\text{pr}} = \hat{X}(z_{t_1}, c_{\text{pr}})$ by using an inherited sampler on the frozen pre-trained diffusion model with random initial noise $z_{t_1} \sim N(0, I)$ and conditioning vector $c_{\text{pr}} := \Gamma(f(\text{“a [class noun]”}))$. The loss becomes:

$$\mathbb{E}_{x,c,e,e',t}\Big[w_t\left\|\hat{X}_\theta(\alpha_t x + \sigma_t e, c) - x\right\|^2 + \lambda w_t'\left\|\hat{X}_\theta(\alpha_t' x_{\text{pr}} + \sigma_t' e', c_{\text{pr}}) - x_{\text{pr}}\right\|^2\Big]$$

With different experiments and trials, based on our resources, we have achieved the best results with $\lambda = 1$, learning rate 1e-6, 600 iterations for Stable Diffusion, specifically $Realistic_Vision_V2.0$ which is open source and available in Hugging Face [30], and 16 examples of image input. The training process took about 5-10 minutes with the NVIDIA T4 Tensor Core GPU.

C. Image-to-text generation (Captioning)

For image-to-text generation, we have used the *LLaVa* model, which integrates a vision encoder with a language model. The vision component uses a pre-trained *CLIP* visual encoder to extract visual features from input images [29]. These visual features are then transformed into a format that can be processed alongside textual data by the language

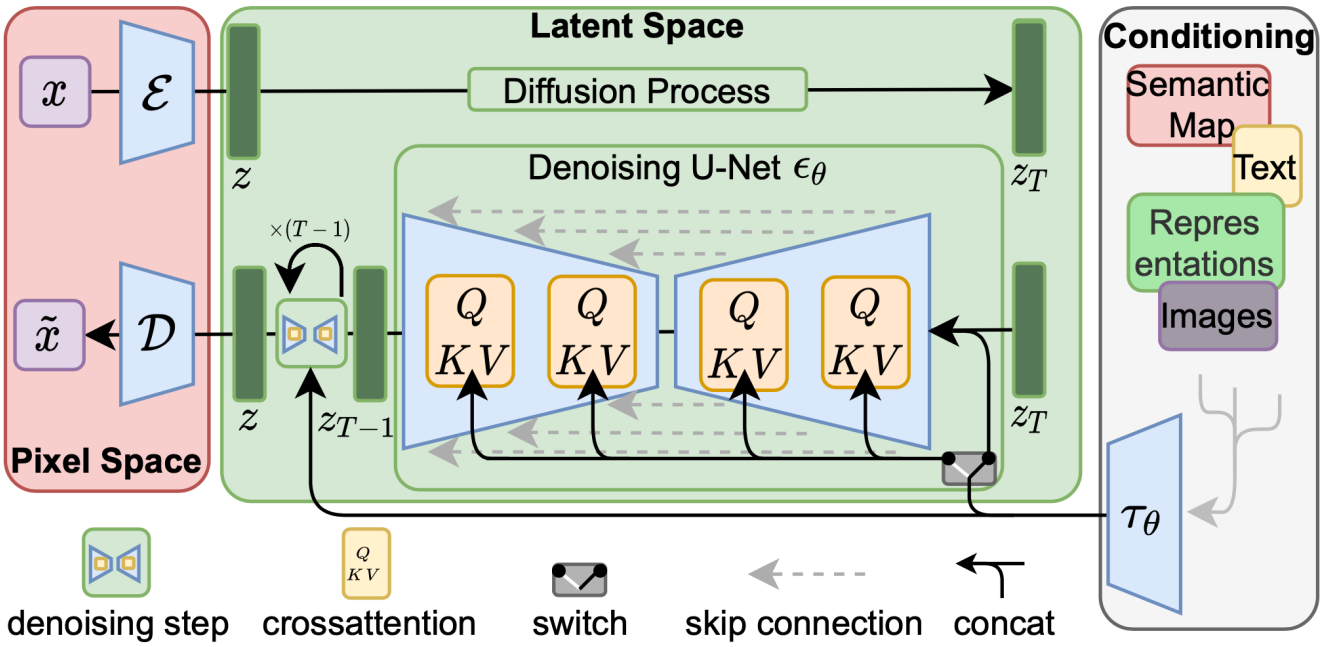

Fig. 1: The architecture of the latent diffusion model (LDM)

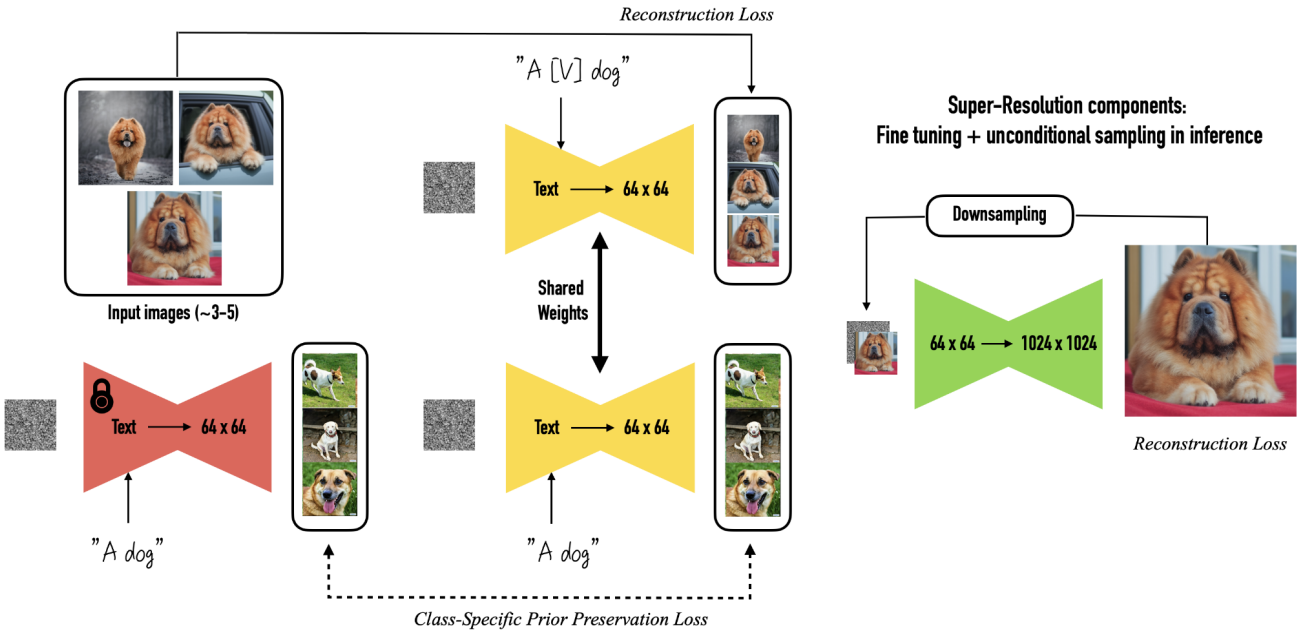

Fig. 2: DreamBooth Fine-Tuning

model. To generate training data, *LLaVa* leverages language-only models like GPT-4 to create multimodal instruction-following data from image-text pairs. This involves generating questions and answers that form a conversation around the visual content, which is crucial for training the model in a way that aligns with human-like interaction and understanding.

LLaVa is fine-tuned using specially created multimodal instruction data [29]. This process involves adjusting the model to improve its ability to follow detailed instructions linked to visual content, enhancing its performance on tasks that require understanding and interacting with both text and images [29]. It has so many models, and the best one that fits our caption

generation was with “llava-hf/llava-1.5-7b-hf" which is open source and available in Hugging Face [31].

One of the advantages of *LLaVa* is that it is data efficient and smart. Haotian Liu et al. did an experiment showing that *LLaVa* have impressive multimodal chat abilities that sometimes overcomes the behaviors of multimodal GPT-4 and has 85.1% relative score compared with GPT-4 on a synthetic multimodal instruction-following dataset [29].

As our influencer is Armenian, we used *googletrans* library of Python to translate those captions generated by LLaVa into Armenian [32].

The caption given with the model *LLaVa* is the following.3

Fig. 3: A generated image using Realistic Vision v2.0

Caption:

I recently visited a charming cafe in Paris called Collective. The cafe is known for its unique and cozy atmosphere, which is perfect for a relaxing afternoon or a casual meet-up with friends. The interior design is a mix of vintage and modern elements, creating a warm and inviting space. The cafe offers a variety of delicious food and beverages, including a selection of pastries, sandwiches, and coffee. The friendly staff and the opportunity to sit by the window and watch the bustling city life outside make Collective an unforgettable experience.

Translated Caption:

Վերջերս այցելեցի Փարիզի մի հմայիչ սրճարան, որը կոչվում էր Collective: Սրճարանը հայտնի է իր յուրահատուկ և հարմարավետ մթնոլորտով, որը կատարյալ է հանգստանալու կեսօրին կամ ընկերների հետ պատահական հանդիպման համար: Ինտերիերի դիզայնը վինտաժային և ժամանակակից տարրերի խառնուրդ է՝ ստեղծելով ջերմ և գրավիչ տարածք: Սրճարանը առաջարկում է մի շարք համեղ ուտեստներ և խմիչքներ, ներառյալ խմորեղենի, սենդվիչների և սուրճի ընտրանի: Բարյացակամ անձնակազմը և պատուհանի մոտ նստելու և դրսում աշխույժ քաղաքային կյանքը դիտելու հնարավորությունը Collective-ին դարձնում են անմոռանալի փորձ:

D. Video DeepFaking

To make our influencer more interactive, we explored video generation tools. However, with different experiments, we have concluded that we would use deep faking [7], [8] in our videos instead of video creation because of the sparsity of the good open-source models and the existing models' limitations for this task. Thus, we have the initial video of a real person and we swap the face of the video with our influencer's face. We have found a model named *ROOP* that works best for this task. *ROOP* combines OpenCV for frame extraction with *InsightFace* for face identification to alter faces across video frames. Adjusting frame rates, preserving audio synchronization, and utilizing various processing tools for multiple effects, like face swapping and augmentation, enable the production of high-quality, deep-faked videos. *ROOP* is an excellent example of the developments in deepfake technology, which uses *GANs* to produce incredibly real fake movies.

E. Automated Posting with Selenium

We did automated posting, for which we used the *Selenium* package. *Selenium* package automates web browser interaction using Python. With the help of the *Selenium* package, we first open the Chrome driver, log in to Instagram, click on the post button, select the file from the computer, take the caption from the .txt file, and post it in just a few minutes. Our only problem that could arise was that Instagram would sometimes block our account because of many requests in specific time intervals because of the automated logins. To solve this problem, we made random time sleeps for each of the actions so that Instagram would not understand that the login and posting process was automated.

All the posts and captioning of Sahmi's Instagram page were done automatically.4

IV. Dataset and Evaluation metrics

To create an AI-generated influencer, we utilized the Generated Photos website [27], which allows for the generation of people using various poses, characteristics, and styles. We specified our vision of what the influencer should look like and generated multiple photos of her in different styles and poses. These images were used during the fine-tuning process to ensure that the model comprehensively understood her appearance. For our final model, we compiled a dataset of 16 images of our generated influencer. Additionally, we employed experimental datasets to assess which types of images the model performs best with and to proof-test the model.

We have collected 20 prompts: 10 descriptive, 5 creative and contextual, and 5 attribute-specific prompts. We have generated 4 images per prompt for evaluation purposes, totaling

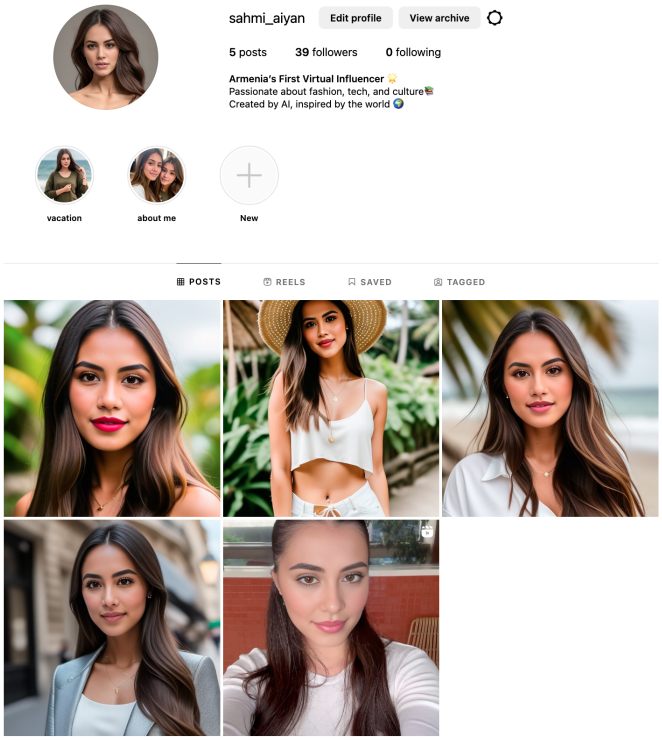

Fig. 4: Sahmi's Instagram account

Fig. 5: Input images for fine-tuning Realistic Vision v2.0 with DreamBooth

80. This allows us to measure a method's performance and generalization capabilities robustly. The full list of prompts can be found in the APPENDIX section.

Subject fidelity is another important aspect to evaluate: the preservation of subject details in generated images. For this, we computed *CLIP-I* [17], the average pairwise cosine similarity, a metric of the cosine of the angle between two vectors in a multidimensional space, between *CLIP* embeddings of generated and real images. The second important aspect to evaluate is prompt fidelity, measured as the average cosine similarity between prompt and image *CLIP* embeddings. Let's denote this as *CLIP-T*. The TABLE I shows the average of the results of the output of the metrics:

Methods	CLIP-I	CLIP-T
Realistic_Vision_V2.0	0.64	0.32

TABLE I: Performance Metrics

Fig. 6: Samples from generated images using different types of prompts

V. Experiments

We have used the world-famous model Kendall Jenner for proof testing to demonstrate that *DreamBooth* works well on personalized subject generation, specifically for people. We used 12-14 images(7) photos of her from different angles and fine-tuned the *Stable Diffusion v1.5* model. The results were promising as they closely resembled the input data. However, we wanted to challenge the model further since we assumed that the model had already learned some information about Kendall.

Fig. 7: Input images of Kendall Jenner

We aimed to check how the model would perform with a person it had never seen before since our influencer had

Fig. 8: Generated images of Kendall Jenner using DreamBooth with Stable Diffusion v1.5

never existed. Therefore, we tested the model with a different dataset, providing 15 images of AI-generated women's faces, which were mostly similar. The results were unsatisfactory despite various prompt engineering techniques that emphasized certain words and included effective negative prompts. We also experimented with different parameters to identify those that could yield promising results.

Finally, we decided to use another stable diffusion model that we assumed would provide more realistic outputs than we had seen before. Thus, we tested $Realistic_Vision_V2.0$, which is open-source and available on Hugging Face [30]. By experimenting with different parameters and different types of inputs (as very often overfitting happens), we achieved controlled generation from these images, resulting in realistic outcomes. Additionally, we have explored prompt-engineering techniques to give better prompts. For example, suppose we wanted to give more attention to one word. In that case, we specify parentheses **"()"** and for less attention square brackets **"[]"**. Additionally, with numbers, we can specify how much attention we want to give to a specific word, for example, **"(red lipstick: 1.5)"**.

A clear cut examples can be seen in the following images. 9b 9a

Prompt: photo of zwx girl, near Eifel Tower, at night, high detailed skin, glossy eyes, at 8K UHD.9b

Negative prompt: rz-neg-15-foranalog, CGI, 3d, lowres, text, error, cropped, worst quality, low quality, jpeg artifacts, ugly, duplicate, morbid, mutilated, out of frame, extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, blurry, dehydrated, bad anatomy, bad proportions, extra limbs, cloned face, disfigured, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, fused fingers, too many fingers, long neck, username, watermark, signature

As can be seen in the image 9b "at night" is not well understood.

Prompt: photo of zwx girl, near Eifel Tower, (at night: 1.5), high detailed skin, glossy eyes, at 8K UHD. 9a

Negative prompt: rz-neg-15-foranalog, CGI, 3d, lowres, text, error, cropped, worst quality, low quality, jpeg artifacts, ugly, duplicate, morbid, mutilated, out of frame, extra fingers, mutated hands, poorly drawn hands, poorly drawn face, mutation, deformed, blurry, dehydrated, bad anatomy, bad proportions, extra limbs, cloned face, disfigured, gross proportions, malformed limbs, missing arms, missing legs, extra arms, extra legs, fused fingers, too many fingers, long neck, username, watermark, signature.

Given "at night" with "()" and a number, the night is well understood 9a.

(a) Image generated with prompt engineering attention techniques

(b) Image generated without prompt engineering attention techniques

Fig. 9: Differences of image generation based on prompt engineering techniques

We have tried OpenAI API for caption translation. However, the translation was not as good as the one with *googletrans*. Additionally, to make our influencer more interactive, we used the OpenAI Whisper model to generate speech from the captions, which we then integrated with video using the *Wav2Lip* lip-syncing method [33]. Additionally, to make our influencer more interactive, we used the OpenAI Whisper model to generate speech from the captions, which we then integrated with video using the *Wav2Lip* lip-syncing method [33]. Unfortunately, we could not reach high-quality lip syncing results because of the quality of the video.

VI. Limitations

In this project, we encountered a few limitations that prevented us from getting better results. First, generating full-body images was challenging 11 even though the model was sometimes generating very good results. We tried to fine-tune the model with more full-body pictures so that the model could understand how the girl should look from afar. However, the overfitting 12 happened even when we changed the parameters. Hence, when the model generated a photo on the far side, it became inaccurately or strangely depicted.

Voice generation was another significant limitation that we encountered. Our young girl had no open-source, appropriate, and accent-free voice. However, as we had videos with deep fakes, we found a website where we generated a voice, not the one that we wished to be, but still, at this point, it was good enough to use it until either we had resources to train

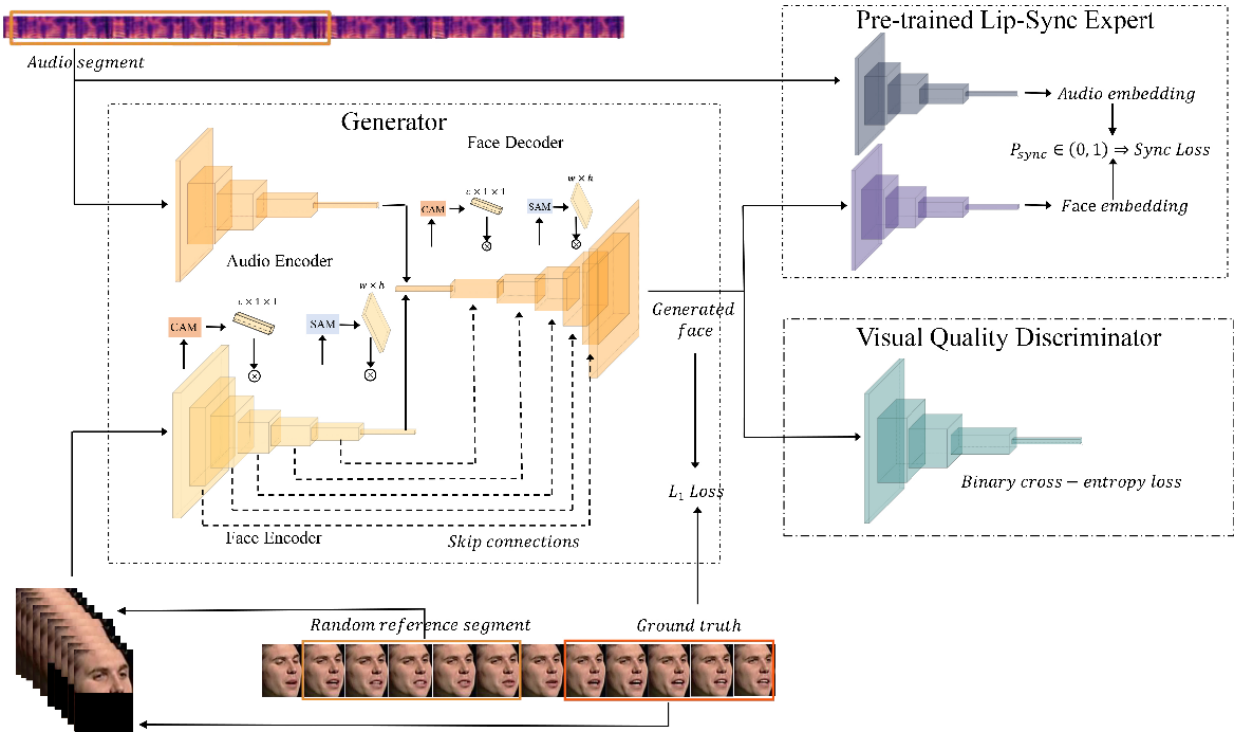

Fig. 10: Attention-Based Lip Audio-Visual Synthesis for Talking Face Generation in the Wild

Fig. 11: Bad generation of images at greater distances

Fig. 12: Overfitting of the model because of many full-body input images

the model with Armenian voice, or a good model appears in the industry.

Additionally, the quality of 2D lip-syncing could have been better; the generated lip movements were often blurry and unnatural. This issue made the video content seem less realistic. Additionally, there were no available 3D open-source models that could be used for this task. Lastly, a lack of resources prevented our efforts in video generation. We have tried the*"stabilityai/stable-video-diffusion-img2vid-xt"* [34] model that was added recently. However, it had several limitations; for example, videos could not be more than 4 seconds, videos might be generated without motion, or the model was not controlled through texts.

These limitations show the current limits in making digital personas and point out where more work and investment are needed to improve the field and our project.

VII. Conclusion and Future Work

This study has shown how a virtual influencer was created and operated using AI techniques. With the $Realistic\_Vision\_v2.0$ model of Stable Diffusion we have developed a virtual persona that can make content on social media platforms, specifically on Instagram. With the deepfake technology and automated captioning and posting tools our influencer is totally independent. Sahmi, with an Armenian identity, proves the power of AI to not only generate realistic

images and videos but also preserve the persona's identity across various interactions and scenarios 13.

Fig. 13: Samples of genereated images of the virtual influencer

In the future, certain aspects of our project will require improvements to enhance the influencer's realism and interactions. Firstly, we aim to generate high-quality video when a suitable open-source model becomes available in the industry. Alternatively, we will seek to enhance lip-syncing methods to achieve better results than we have now with the 2D model. Secondly, we aim to refine the fine-tuning of our model to maintain the persona's identity even in complex prompts where the influencer appears from a distance. We will also utilize models like *Low-Rank Adaption(LoRA)* to define our influencer's style better. Thirdly, we will work on integrating a model that preserves the influencer's Armenian-speaking voice without any accent.

Another possible step further could be using our virtual influencer to promote certain brands and places. The market research about the need and potential effectiveness of such a development of the ideas presented in this paper will be investigated.

In summary, our AI-powered virtual influencer not only stands as a milestone in the aspect of AI utilization in social media platforms but also offers a backbone for further research on the integration of AI in day-to-day human interactions. The venture advances our understanding of AI potential and triggers a deeper discussion and understanding of the consequences of such technologies in social and ethical issues.

REFERENCES

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis With Latent Diffusion Models," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10689-10699.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising Diffusion Probabilistic Models," arXiv preprint arXiv:2006.11239, 2020

[3] Lilian Weng, "Diffusion Models," 2021.

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," arXiv preprint arXiv:2208.12242, 2022

[5] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8394–8403, 2020.

[6] Shi, T., M. R. Min, R. Lee, K. Kim, and M. Lee. "Stable Diffusion: A New Approach to Training Generative Networks." arXiv preprint arXiv:1803.01837 (2018).

[7] Patel, Y., Tanwar, S., Gupta, R., Vimal, V., et al. "Deepfake Generation and Detection Case Study and Challenges." IEEE Access, vol. 11, pp. 143296-143323, December 2023.

[8] Tolosana, R., et al. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. Information Fusion, 64, 131-148.

[9] Nguyen, T.T., et al. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey. arXiv preprint arXiv:1909.11573.

[10] Xu, L. et al. (2021). "Deep Divergence Learning for Image-to-Text Generation," IEEE Transactions on Neural Networks and Learning Systems.

[11] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castricato, and E. Raff, "VQGAN-CLIP: Open Domain Image Generation and Editing with Natural Language Guidance," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2022.

[12] Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instanceconditioned gan. Advances in Neural Information Processing Systems, 34:27517–27529, 2021.

[13] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938, 2021.

[14] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626, 2022.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in Proc. 38th Int. Conf. Machine Learning (ICML), 2021, pp. 8748-8763.

[18] Ron Mokady, Omer Tov, Michal Yarom, Oran Lang, Inbar Mosseri, Tali Dekel, Daniel Cohen-Or, and Michal Irani. Self-distilled stylegan: Towards generation from internet photos. In Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings, pages 1–9, 2022.

[19] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv preprint arXiv:2108.00946, 2021.

[20] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34:8780–8794, 2021

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo-Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487, 2022

[22] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789, 2022.

[23] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217, 2022.

[24] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. arXiv preprint arXiv:2204.14217, 2022.

[25] Wang,Y. et al. (2023). "Optimizing CLIP Scores in GANs for Enhanced Text-to-Image Synthesis," Neural Information Processing Systems.

[26] Kaushar, S. et al. (2024). "ImageVista: Training-Free Text-to-Image Generation with Multilingual Input," Conference on Innovative Data Communication Technologies and IoT, March 2024, DOI: 10.1109/ID-CIoT59759.2024.10467385.

[27] "Generated Photos." https://generated.photos/

[28] N. A. Rink, A. Paszke, D. Vytiniotis, and G. S. Schmid, "Memory-efficient array redistribution through portable collective communication,"

in Proceedings of the 27th ACM Symposium on Parallelism in Algorithms and Architectures, 2021, pp. 123-136. https://arxiv.org/pdf/2112.01075
[29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," arXiv preprint arXiv:2304.08485v2, Dec. 2023.
[30] SG161222, "$Realistic\_Vision\_V2.0$," Hugging Face Models. https://huggingface.co/SG161222/Realistic_Vision_V2.0.
[31] llava-hf, "llava-1.5-7b-hf," Hugging Face Models, https://huggingface.co/llava-hf/llava-1.5-7b-hf
[32] "googletrans," PyPI. https://pypi.org/project/googletrans/.
[33] G. Wanga, P. Zhang, L. Xiea, W. Huangb, and Y. Zhaa, "Attention-Based Lip Audio-Visual Synthesis for Talking Face Generation in the Wild," arXiv preprint arXiv:2203.03984, 2022
[34] Stability AI, "Stable Video Diffusion - Img2Vid XT," https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt

APPENDIX

Different types of prompts that were used to estimate the average pairwise cosine similarity between CLIP embeddings of generated and real images and the average cosine similarity between prompt and image CLIP embeddings

Descriptive Prompts

1) item Photo of zwx girl sitting at a café terrace, sipping coffee.
2) Photo of zwx girl attending a tech conference as a speaker.
3) Photo of zwx girl riding a bicycle through a quaint village.
4) Photo of zwx girl painting in a sunlit studio.
5) Photo of zwx girl shopping in a luxury boutique.
6) Photo of zwx girl walking a dog in a suburban neighborhood.
7) Photo of zwx girl reading a book in a cozy library.
8) Photo of zwx girl yoga in a peaceful garden.
9) Photo of zwx girl on a ski trip in the mountains.
10) Photo of zwx girl attending a wedding as a guest.

Creative and Contextual Prompts

1) Imagine photo of zwx girl discovering a hidden magical city.
2) Photo of zwx girl as a time traveler visiting ancient Egypt.
3) Photo of zwx girl as a guest at a royal ball in a castle.
4) Photo of zwx girl as a futuristic robot-building competition.
5) Photo of zwx girl as a chef in a magical cooking show.

Attribute-Specific Prompts

1) Photo of zwx girl in hiking attire exploring a national park.
2) Photo of zwx girl wearing traditional Japanese kimono in Kyoto.
3) Photo of zwx girl in a glamorous gown at a film premiere.
4) Photo of zwx girl in desert gear exploring sand dunes.
5) Photo of zwx girl in a chef's uniform in a high-end kitchen