Inversion based Style Transfer with Diffusion Models

Tammireddy Sri Vallabh *210101126* Posa Mokshith *210101077* Nitish Kumar Pinneti *210101125* Mukka Koushik *210101069* Rangu Rishvanja Simha *210101084*

***Abstract*—Style transfer is a deep learning technique that synthesizes an image by blending the content of one image with the style of another. Recent advancements, particularly in diffusion-based methods, have enhanced the ability to capture intricate details in style and structure. The Inversion-based Style Transfer (InST) approach uses diffusion models and textual inversion to transfer detailed stylistic elements, such as brushstrokes and color schemes, from a reference painting to a natural image while preserving content integrity. However, InST faces challenges in style representation quality, especially when using textual encodings. In this work, we introduce an improvement to InST by incorporating ShareGPT-4V LLaVA(Large Language and Vision Assistant) model to generate refined style descriptions directly from images, enabling enhanced prompt optimization. This modification allows the model to better encode the unique characteristics of style images, producing more stylistically accurate and visually compelling results. Our experiments show that this approach reduces content deviation and improves the fidelity of style transfer, yielding higher-quality images.**

***Index Terms*—style transfer, diffusion models, Inversion-based Style Transfer (InST), ShareGPT-4V, prompt optimization, artistic representation, content preservation**

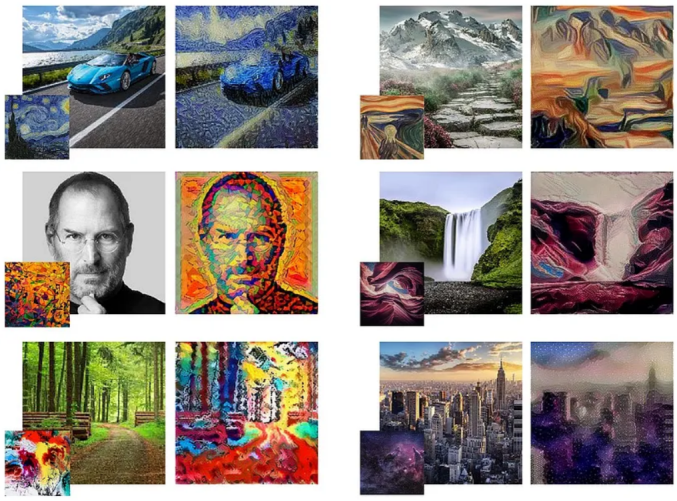

Fig. 1. An example of style transfer

I. Introduction and Overview

Style Transfer is a deep learning technique that combines the content of one image with the artistic style of another. Traditional approaches use Convolutional Neural Networks (CNNs) to separate content (such as shapes and objects) from style (such as colors, textures, and patterns). This allows for the creation of an image that retains the structure of the content image while exhibiting the visual style of the style image. This technique has numerous applications: it enables artistic image generation by transforming photos into artworks inspired by famous paintings, enhances photo and video editing with unique filters, and enables augmented reality (AR) to transform live video feeds with artistic effects in real time. Additionally, style transfer aids fashion and design by generating innovative patterns for textiles and interior decor, and it enhances advertising, marketing, and film production by creating visually engaging and stylized content.

The core challenge in style transfer is to produce an image that successfully blends the semantic structure of the content image with the aesthetic attributes of the style image. Given a content image and a style image, the task is to generate a new image that maintains the structural and semantic content of the original while adopting the visual style—such as color palette, textures, and patterns—of the style image.

Our method builds upon the classical optimization-based approach, where an image is iteratively refined to minimize content and style losses derived from feature representations in a pre-trained CNN. By incorporating adjustments to improve texture consistency and color accuracy, our approach dynamically adapts to variations in content and style complexity. This allows it to achieve high-quality, nuanced results, preserving fine details without compromising the structural integrity of the original content. Consequently, our method is versatile and can handle a wide range of style transfer applications, from detailed artistic rendering to more simplified stylization tasks.

II. Related Work

In deep learning, style transfer has evolved significantly over the years. Below are some key methods that have shaped this field:

- **Traditional Style Transfer (CNN-based)** - The foundational work in neural style transfer was proposed by Gatys et al. in 2015. They used Convolutional Neural Networks (CNNs) with VGG architecture to separate and recombine content and style. The approach optimizes two losses—content loss and style loss. However, due to the optimization process, inconsistent style application can occur, leading to poor quality in generated images.

- **Image Style Transfer with Transformers (StyTr^2)** - A transformer-based approach that addresses long-range dependencies, which are challenging in CNN approaches. It introduces Content-Aware Positional Encoding (CAPE) to improve scale invariance and enhance image style transfer.
- **AdaAttN (SOTA)** - Enhances local visual quality by adaptively normalizing content features based on shallow and deep features from both content and style images. AdaAttN learns spatial attention scores and calculates per-point weighted statistics, improving style transfer quality and reducing distortions. However, it may not fully capture interdependencies and relationships between different channels.
- **Contrastive Arbitrary Style Transfer (CAST)** - Uses contrastive learning to learn style representations directly from image features. CAST employs Multi-Layer Style Projector and domain enhancement modules to capture style distributions effectively. However, it requires positive and negative style examples, which may be challenging to curate and could limit style diversity.
- **StyleShot (Current SOTA)** - Focuses on extracting expressive style representations with a style-aware encoder trained for generalized style transfer, without needing test-time tuning.

A. Diffusion Models

Traditional style transfer methods, such as convolutional neural network (CNN)-based approaches, focus on the texture and color of reference images, often neglecting complex structural elements like object shapes and high-level semantics. To address these limitations, recent advancements have turned to diffusion models, a class of generative models that gradually transform random noise into structured images. Diffusion models operate by introducing noise into an image over several steps and then learning to denoise it, thereby generating new images with a high degree of detail and structure. This iterative process allows them to capture finer details, such as brushstrokes and object contours, which makes them particularly effective for complex tasks like style transfer.

Fig. 2. Diffusion Process: Adding and removing noise during stochastic inversion.

Diffusion models are composed of two primary phases:

- **Forward Process:** During this phase, noise is progressively added to an input image, transforming it into a fully noisy representation. This phase is designed to introduce variability and allows the model to learn how to reconstruct images from a noisy base.
- **Reverse Process:** This denoising phase gradually removes noise, guiding the noisy input back to a structured image by learning to predict and remove noise at each step. The reverse process allows diffusion models to achieve more coherent and semantically rich outputs compared to traditional models.

This ability to retain structure while introducing new styles or attributes makes diffusion models suitable for style transfer applications, where maintaining the original content is essential.

B. Inversion-based Style Transfer (InST)

Inversion-based Style Transfer (InST) leverages the advantages of diffusion models and combines them with textual inversion techniques to capture the stylistic nuances of an image directly from a reference. Unlike text-based methods that rely on broad or ambiguous style descriptions, InST learns the specific attributes of a painting, including semantic elements, object shapes, brushstrokes, and color distributions, by using a reference image alone. This process allows the model to apply a style transfer that is deeply aligned with the reference's visual characteristics, without requiring extensive textual input.

The InST model operates in three distinct spaces: textual, latent, and pixel, allowing for a comprehensive approach to style transfer. This involves both textual inversion to learn style embeddings and stochastic inversion to guide the generative process while preserving content.

*1) **Textual Inversion**:* In the textual inversion phase, the style image is encoded into a text-like embedding using CLIP (Contrastive Language-Image Pretraining) architecture:

- The style image y is encoded using the CLIP image encoder to obtain a style embedding.
- This embedding is optimized through cross-attention layers, which refine and focus on the style-relevant aspects of the reference image.
- The final output is a learnable vector v, which acts as a textual embedding for the style, effectively capturing the painting's aesthetic attributes.

This embedding enables the generative model to incorporate style features into the output by conditioning on this representation during image synthesis.

*2) **Stochastic Inversion**:* To preserve the structure of the content image, InST uses stochastic inversion, which involves a diffusion process:

- The content image x is first modified by adding noise, creating a noisy version of the original.
- This noisy image then undergoes a denoising process that predicts and removes the added noise in a way that retains the semantic structure of the content image, allowing the style to be applied without losing core content details.

This ensures that the output image maintains the structure of the content while seamlessly blending the stylistic attributes of the reference image.

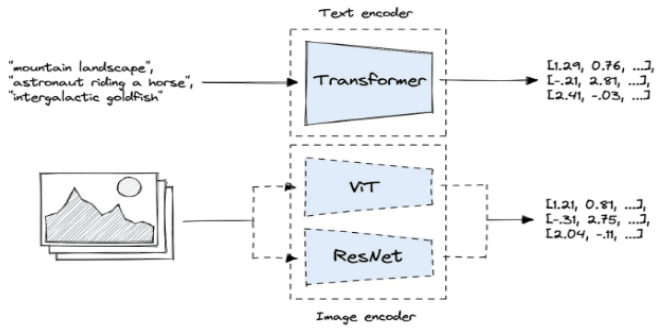

Fig. 3. CLIP Architecture used for encoding style images into textual embeddings.

*3) **Conditioned Synthesis Process in Stable Diffusion Models**:* For efficient style transfer, InST employs the Latent Diffusion Model (LDM), which performs the diffusion process in the latent space:

- During inference, noise is added to the content image according to the predicted noise from the stochastic inversion phase.
- The model combines the style embedding v with this noisy latent representation, guiding the denoising process in latent space.
- The decoder then transforms the final stylized latent representation into pixel space, producing a synthesized image that combines both style and content effectively.

This approach balances high-quality style transfer with computational efficiency, as processing in latent space reduces the computational load compared to pixel-space operations.

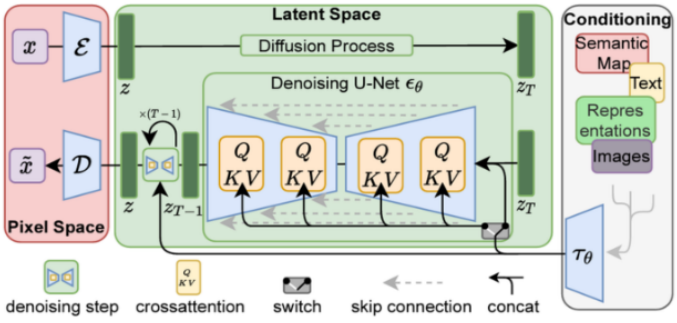

Fig. 4. Latent Diffusion Model (LDM) for efficient style transfer in latent space.

The combination of textual inversion to capture style information and stochastic inversion to maintain content allows InST to produce images that retain the essence of both style and content in a single output.

III. Method

In our work, we mainly utilized ShareGPT4V-7B chatbot which utilized llava-v1.5-7b model pretrained on ShareGPT-4V dataset and inversion as the basis of the InST framework and Stable Diffusion Models (SDMs) as the generative

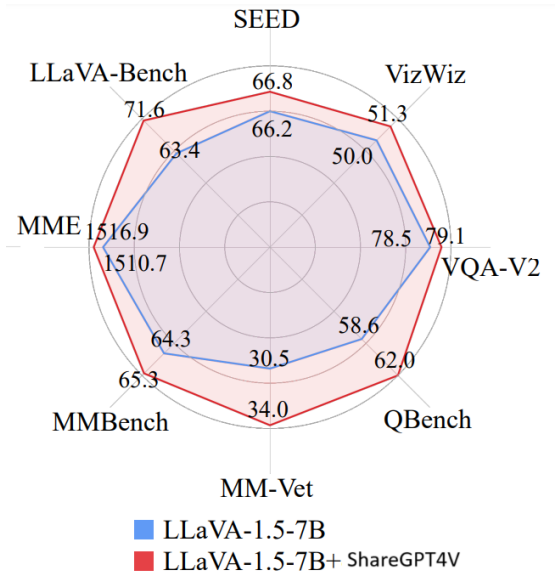

Fig. 5. Performance comparision of LLaVA model with and without ShareGPT4V

backbone. ShareGPT-4V data is a pioneering large-scale resource that features 1.2 million highly descriptive captions. LLaVA-1.5v-7b (Large Language and Vision Assistant) is a multimodal LLM for general-purpose image and language understanding. It can process an input image, and a task or question relevant to the image, and generate an appropriate response.

We downloaded pretrained stable diffusion model [11] that works with two images, context and style images respectively, along with an additional input paremeter prompt, which also helps in conditional learning. We completely removed the style image as direct input and instead, passed the style description generated by LLaVA model trained on sharegpt4v.

Our aim was to generate a high level description of style aspects of the style image through this model using the prompt :

Prompt for Style Description Generation:

```
Task: Give high level description
English of the style aspects of the
image, like color scheme used,style
description of objects,etc.
```

By this method, we obtained high level description of the style image about the style attributes which we used to pass as prompt to the InST model(inversion based style transfer diffusion model) while denoising content image completely excluding the style image. The main reason to do so is because InST model uses clip embeddings of both content and style images and use conditional denoising of content image guided with style image. But clip embeddings of style image mostly contain the content aspects of the image which is contrary to what we needed, ie we need only the style aspects, not the contextual description embeddings.

So, In the inference process, x is the content image, and y

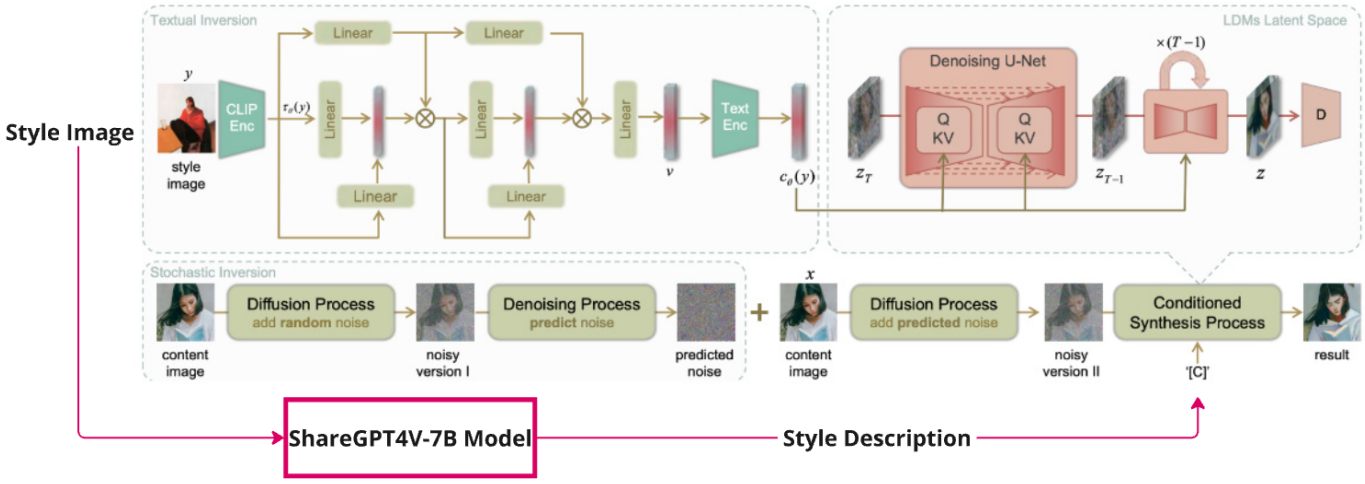

Fig. 6. InST Architecture: The style image is input into the ShareGPT4V-7B model to generate a style description, which is incorporated into the prompt to apply the desired style in the synthesized output. Red path indicates our modifications. The style image is processed through the ShareGPT4V-7B model (indicated by the red path) to generate a style description, which is incorporated into the prompt

TABLE I
AVERAGE PERFORMANCE SCORES FOR EACH MODEL ACROSS DIFFERENT IMAGE CATEGORIES: OVERALL, FACE IMAGES, AND NON-FACE IMAGES. THE HIGHEST SCORE IN EACH CATEGORY IS HIGHLIGHTED FOR BETTER COMPARISON.

Category	InST	InST,syle image with ShareGPT4V Prompt	InST with only ShareGPT4V Prompt
Overall Average	1.500	**2.259**	2.241
Face Images Average	1.475	2.069	**2.456**
Non-Face Images Average	1.525	**2.450**	2.025

is the reference image. The textual embedding v of the prompt the of reference image produced by the chatbot y guides the generative model in generating a new artistic image.

IV. RESULTS

For the experiments conducted with InST, an NVIDIA RTX 3050 GPU was used. The average inference time for generating a single stylized image was approximately 6 minutes.

For generating style descriptions, the experiments were conducted on Kaggle using their provided 2xT4 GPUs. The average time to generate a style description was around 30 seconds.

These timings provide a baseline for understanding the computational requirements and performance of the InST model in both image inference and style description generation.

We compare our method with Zhang's, inversion based style transfer model [3], which was inturn compared with several SOTA image style transfer methods by qualitative measure. Determining which one gives better results is subjective, as there is no perfect measure to capture which model performs better. Therefore, we asked participants to compare the outputs and identify which one extracted more artistic features from the style image.

For 20 participants, 16 content–reference pairs were displayed,the images used were directly from the InST paper, so comparision could be done. We generated results using InST+llava model(with and without style image) and only InST. The participants were informed that the artistic consistency be tween the generated and reference images is the main met ric. Then, they were invited to select which result for each content–reference pair is better. Finally, The percentage of votes demonstrated that our method achieves the best visual characteristic transfer results.

Given the limited data used for accuracy measurements, this model cannot yet be considered a new benchmark. Extensive inference on larger datasets is needed to validate it fully, but due to time and resource constraints, we have left this phase for future work.

Our content-reference pairs are mainly contrasted with two types, facial images and non-facial images. Our results showed that for facial images, pure prompt without using style image for denoising process captured better artistic attributes, and for non-face images, considering both style image and prompt captured better features. The preference scoring system used in this study awards points as follows: 3 points for the 1st place, 2 points for the 2nd place, and 1 point for the 3rd place.

$$\text{Average Score} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \text{Points}_{ij}}{n \cdot m}$$

- Points_{ij} is the score assigned to model i for sample j,
- n is the number of participants,
- m is the number of samples shown (in this case, $m = 16$).

Average score is calculated for all the three variants for comparision study.

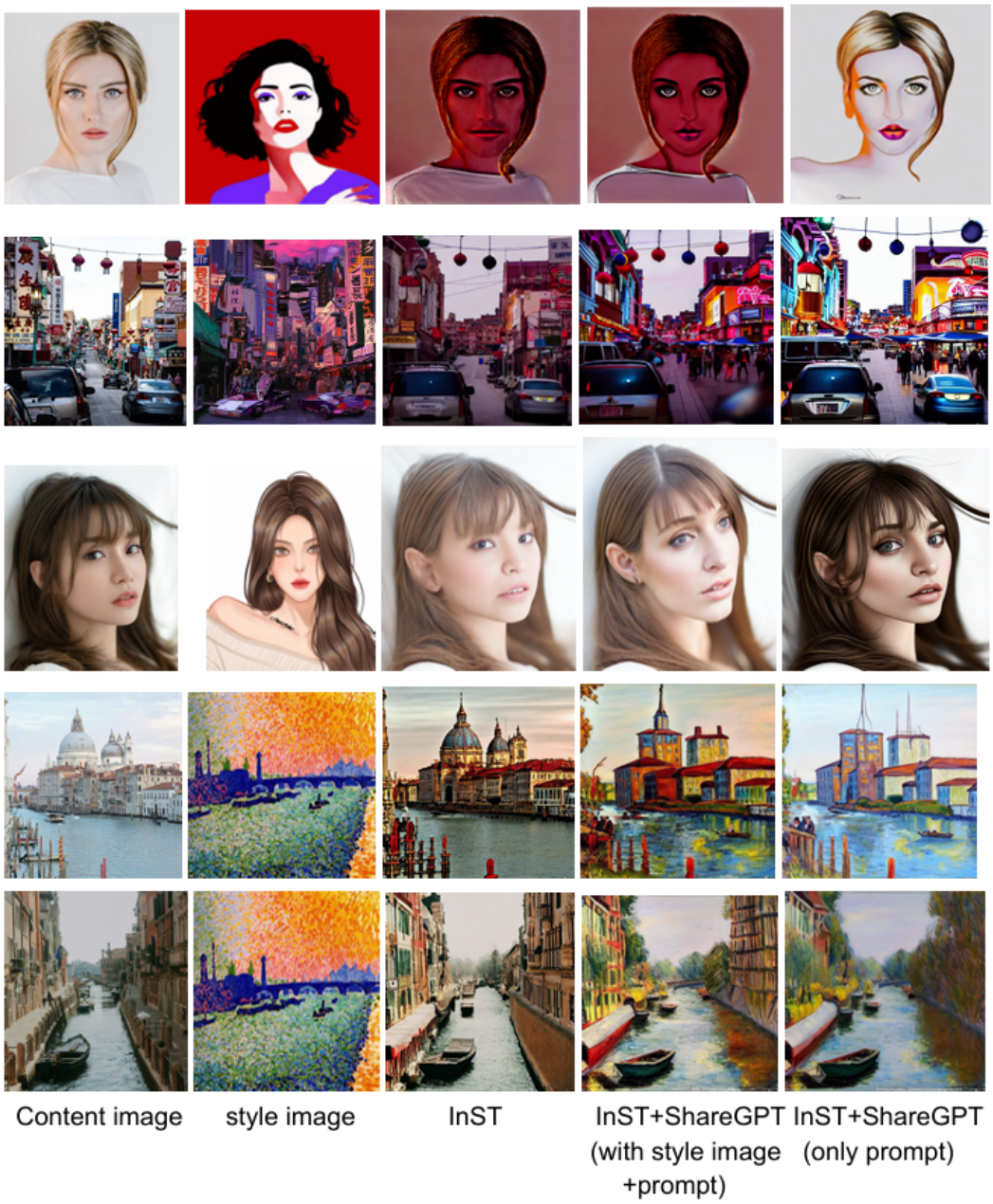

Fig. 7. Qualitative comparison with InST, along with InST conditioned on human captions. Our method can accurately represent the target style image, ie fourth and fifth columns compared to the third column.

REFERENCES

[1] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations (ICLR)*, 2021.

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

[3] Y. Zhang, N. Huang, F. Tang, H. Huang, C. Ma, W. Dong, and C. Xu, "Inversion-based style transfer with diffusion models," *arXiv preprint arXiv:2211.13203*, 2022.

[4] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "ShareGPT4V: Improving large multi-modal models with better captions," *arXiv preprint arXiv:2311.12793*, 2023. [Online]. Available: https://arxiv.org/pdf/2311.12793

[5] J. Gao, Y. Liu, Y. Sun, Y. Tang, Y. Zeng, K. Chen, and C. Zhao, "StyleShot: A snapshot on any style," *arXiv preprint arXiv:2407.01414*, 2024.

[6] Y. Deng, F. Tang, W. Dong, C. Ma, X. Pan, L. Wang, and C. Xu, "StyTr2: Image style transfer with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11326–11336, 2022.

[7] Y. Zhang, F. Tang, W. Dong, H. Huang, C. Ma, Tong-Yee Lee, and C. Xu, "Domain enhanced arbitrary image style transfer via contrastive learning," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022.

[8] L. Gatys, A. Ecker, and M. Bethge, "Image style transfer using convo-

lutional neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2015.

[9] T. Srivallabh, "Fork of ShareGPT4V," [Online]. Available: https://www.kaggle.com/code/tsrivallabh/fork-of-sharegpt4v-226b9e-06785d

[10] T. Srivallabh, "InST code update incorporating ShareGPT4V prompt," [Online]. Available: https://github.com/Sri-Vallabh/InST

[11] Zhang Yuxin, "Pretrained stable diffusion model" [Online]. Available: https://huggingface.co/CompVis/stable-diffusion-v-1-4-original/resolve/main/sd-v1-4.ckpt