International Journal of Computational Mathematics and Computer Science
Volume 1, Issue 1, June (2024)

ISSN: xxxx - xxxx

VISUALISE OUR CUSTOMISED TEXT-TO-IMAGE DIFFUSION MODELS WITHOUT USING ANY SPECIAL TECHNOLOGY

Elakkiya Elango[1], Balasubramanian Shanmuganathan[2]

[1]Guest Lecturer, Department of Computer Science,
Government Arts college for Women,
Sivaganga,Tamilnadu,India,

[2]Assistant Professor , Department of Computer Science,
DDE, Alagappa University,
Karaikudi, Tamilnadu, India,

ABSTRACT

Nowadays considering the advancement of text-to-image frameworks (for example, Stable Diffusion [22]) and associated customizing approaches like DreamBooth [24] and LoRA [13], anybody may express their ideas into images of excellent quality at affordable rates. As a result, picture animations approaches that blend created static images with motion dynamics are in high demand. On this paper, we provide a realistic framework for animating most current personalized text-to-image models for good after everything, therefore saving valuable time on model-specific tweaking. The suggested framework's central idea is to inject a dynamically initialized motion modeling modules into an existing frozen text-to-image model and training it using video clips to extract realistic motion priorities. Once trained, all personalized versions that utilize the same basic T2I quickly transform into text-driven models which generate different and personalized animated visuals by just injecting the aforementioned motion modeling module. We test numerous public indicative personalized text-to-image models on animated images and realistic images, and show how our proposed framework helps these frameworks generate temporally smooth animation clips whereas conserving the domain as well as different perspectives of their outputs. The code and pre-trained weights are going to be made public on our project website.

Keywords:

Personalized Animation, AnimateDiff, Network Inflation, 5D video tensor

1. INTRODUCTION

Text-to-image (T2I) generative models [17, 21, 22, 25] have received immense interest within as well as outside the field of research in recent years, owing to their high resolution and text-driven controllability, i.e., a low-barrier point of entry for non-researcher users which allows artists and beginners to conduct AI-assisted content creation. To encourage the imaginative use of existing T2I generative models, multiple light-weighted individualization techniques, such as DreamBooth [24] and LoRA [13], are proposed that allow tuned fine-tuning of those models on small datasets using a consumer-grade device like an a laptop featuring the RTX3080, following which both of these models may generate tuned content with substantially improved excellence. Users may integrate novel ideas or approaches to a pre-trained T2I template in this way at a relatively minimal cost, leading in the countless personalised models submitted by artists and amateurs on models-sharing sites like CivitAI [4] and Huggingface [8]

*Corresponding author

Although personalized text-to-image models trained using DreamBooth or LoRA has garnered notice for their exceptional visual quality, the results they produce are static pictures. Additionally, there is, specifically, a lack of temporal level of flexibility. Given the wide range of uses for animation, we'd like to find out how we can convert most of the existing personalized T2I models into representations that generate animated visuals whilst retaining their original visual appeal. Modern approaches to universal text-to-video generation [7, 12, and 33] advocate embedding temporal modeling into the initial T2I models and refining the frameworks using video datasets. However, personalized T2I models face difficulties since consumers often cannot afford sensitive hyper parameter tweaking, personalized video collecting, and intensive computing resources.

In this paper, we introduce AnimateDiff, a universal approach for generating animating visuals for any personalized T2I models while needing no models-specific adjustment and ensuring visually appealing stability across time. considering the fact that many personalized T2I models are created from the same base one (e.g. Stable Diffusion [22]) and the fact that acquiring the appropriate videos for each personalized domain is impractical, we decided to create a motion modeling section which could animate the most personalized T2I models at once and for all. Using specific terms, a motion modeling module is added to an initial T2I models before being fine-tuned on massive video clips [1], developing the appropriate motion priors. It should be noted that the elements of the fundamental model stay unchanged. We demonstrated that the derived personalized T2I could benefit through the experienced motion priors, creating smooth and attractive animations during fine-tuning. That means, the motion modeling modules is capable of animating all of the related personalized T2I models without the need for extra data collection or customized training.

We test our AnimateDiff on a variety of typical DreamBooth [24] and LoRA [13] models, including anime images and realistic photos. Most personalized T2I models might be directly animated by introducing the well-trained motion modeling module without any special adjustment. In practice, we discovered that simple attention across the temporal dimension is sufficient for the motion modeling module to acquire the correct motion priors. We also show how the motion priors may be generalized to domains like 3D cartoons and 2D anime. To that aim, our An- imateDiff might lead to a simple yet effective baseline for personalized animation, allowing consumers to receive personalized animations rapidly while just incurring the cost of personalizing the picture models.

2. LITERATURE SURVEY

2.1. Models of text-to-image diffusion: Text-to-image (T2I) diffusion models have increased in popularity in and outside of the scientific field in recent decades, owing to large-scale text-image data pairs [26] and the robustness of diffusion models [5, 11]. GLIDE [17] was one of them, as it added text conditions to a diffusion framework and proved that classifiers assistance delivers visually appealing results. DALLE-2 [21] enhances text-image alignment by utilizing the CLIP [19] joint feature space. To accomplish photorealistic picture production, Imagen [25] combines a big language model [20] pre-trained using text corpora with a cascade of diffusion models. The latent diffusion model [22], also known as Stable Diffusion, suggested executing the denoising method within the latent space of an auto-encoder, effectively decreasing the necessary computer resources while keeping the level of quality and flexibility of the produced pictures. In contrast to the previous research, which shared parameters throughout the generation process, eDiff-I [2] trained an ensemble of diffusion models specialized for distinct stages of synthesis. Our approach is based on a pre-trained text-to-image model and may be tuned to any personalized version.

2.2. Modify the text-to-image model: Although there currently have been numerous robust T2I generative algorithms developed, individual users still remain unable to train their own models owing to the need for large-scale information and computing resources, which are merely available to major enterprises as well as research organizations. As a result, numerous strategies for introducing new domains (new ideas or patterns defined mostly by a limited number of photos gathered by users) into pre-trained T2I models are being developed [6, 9, 10, 14, 16, 24, 27]. Textual Inversion [9] advocated that each notion, the word encoding be optimized and the initial networks be frozen during training. An additional method that finetunes the entire networks using preservation losses as regulation is Dream- Booth [24]. Custom Diffusion [16] enhances fine-tuning performance by upgrading only a selected group of parameters and enables idea merge via closed-form optimization. Dream Artist [6] also compresses each input into a single photograph. Recently, LoRA [13], a language-based model adaption approach, was used enabling text-to-image models fine-tuning and obtained high visual quality. Although these approaches are mostly centered on parameter adjustment, other research [10, 14, 27] have attempted to train a more generic encoder for idea personalization. Among all of the personalization techniques in the field of research, our work is primarily focused on tuning-based methods, such as DreamBooth [24] and LoRA [13], because they preserve the basic model's feature space.

2.3. Personalized T2I animation: Since the context in this paper is novel, there's presently minimal work on it. Despite the fact that it is usual practice to augment an existing T2I model with temporal features for video creation [7, 12, 15, 28, 31, 33], entire Method should be updated**.**

3. METHODOLOGY

In this dataset taken from Github. This part Sec. 3.1 presents preliminary knowledge regarding the generic text-to-image paradigm and its personalized versions. Sec. 3.2 then gives the formulation of personalized animation as well as the reason for our technique. Finally, Section 3.3 presents the actual implementation of AnimateDiff's motion modeling module, which animates distinct personalized models to provide appealing synthesis.

3.1. PRELIMINARIES

Wide Text-to-image generator: In this work, we employed Stable Diffusion (SD), a popular text-to-image model, as a broad T2I generator. SD relies upon the Latent Diffusion Model (LDM) [22] Fig 1, which performs the denoising process in the latent domain of an auto-encoder, notably E() and D(), pre-trained on big image datasets like VQ-GAN [14] or VQ-VAE [29]. This approach has the advantage of lowering computational costs while maintaining good visual quality. During latent diffusion network training, the input picture x0 is initially assigned to the latent space by the frozen encoder, generating z0 = E(x0), and then disturbed using a predetermined Markov process:

$$\mathbf{q(zt|zt-1) = N\ (zt;\ p\ 1 - \beta tzt-1,\ \beta tI)}\text{--}\ \mathbf{(1)}$$

Where t = 1,..., T, where T is the total amount of stages in the advance diffusion process. The noise intensity at every stage is determined by the order of hyper-parameters t. The iterative procedure described above may be restated in a closed-form as follows:

$$\mathbf{zt = \surd\ \alpha^{-} tz0 + \surd\ 1 - \alpha^{-} t\epsilon,\ \epsilon \sim N\ (0,I)}\text{--}\ \mathbf{(2)}$$

Where $\alpha^{-} t = Qt\ i=1\ \alpha t$, $\alpha t = 1 - \beta t$. The vanilla training target proposed in DDPM [5] is used by Stable Diffusion, which may be stated as:

$$\mathbf{L = EE(x0),y,\epsilon{\sim}N(0,I\ ),t\ \| \epsilon - \epsilon\theta(zt,\ t,\ \tau\theta(y))}\ \|\ [2/\ 2]\text{--------------------------------}\ \mathbf{(3)}$$

Where y is the equivalent textual description, $\tau\theta(\cdot)$ is a text encoder mapping the string to a sequence of vectors.

In SD, $\epsilon\theta(\cdot)$is executed using an enhanced UNet [23] with four down-sample / up-sample blocks and one middle block, which produces four resolution levels inside the network's latent space. Each resolution level includes 2D convolution layers along with methods for self- and cross-attention. The CLIP [19] ViT-L/14 text encoder is used for implementing the text model ().

3.1.1. Personalized Image Generation:

As generic image generation advances, more focus is being dedicated to personalized image generation. DreamBooth [24] and LoRA [13] are two well-known and commonly utilized customization methods. A simple way to add a new domain (new concepts, styles, etc.) to a pre-trained T2I model is to fine-tune it on photos from that domain. However, tweaking the predictive algorithm directly without regularization frequently results in over fitting or catastrophic forgetting, particularly if the dataset is small. To address this issue, DreamBooth [24] use an uncommon text as the indication for the target domain and augments the collection of data with photos created by the original T2I model. These regularization pictures are created without the indication, enabling the model to be used to acquire the knowledge to link the rare string with the rare string. LoRA [13], on the other hand, takes a different approach by attempting to fine-tune the model weights' residual, which is, training ΔW instead of W.

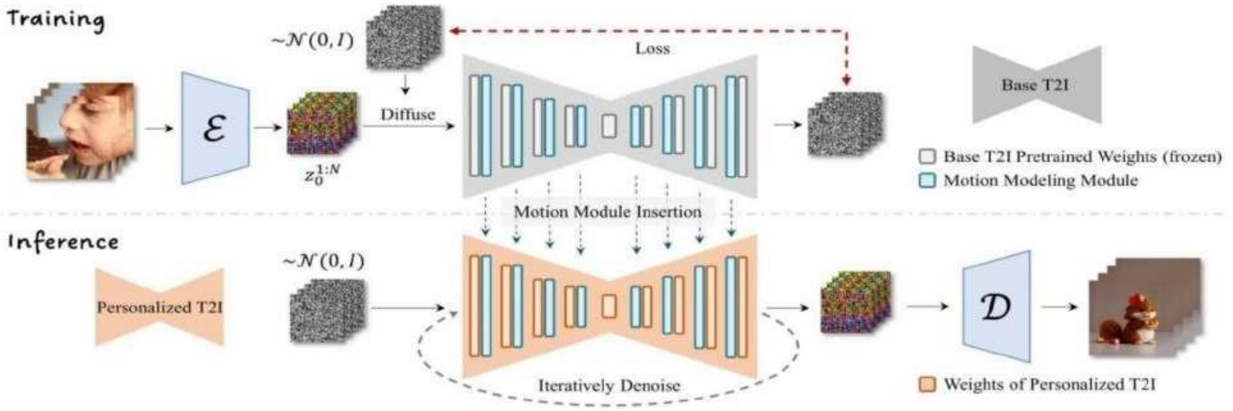


Fig1. Shows the AnimateDiff pipeline. Given a foundation T2I model (for example, Stable Diffusion [22]), we train a motion modeling module on video datasets to encourage it to distil motion priors. Only the parameters of the motion module are modified at this step, keeping the feature space of the underlying T2I model. Once trained, the motion module can transform any personalized model based on the basic T2I model into an animation generator, then generate various and personalized animated visuals via an iterative denoise process.

3.2. Personalized Animation

Animating a personalized picture model frequently necessitates extra tweaking with a related video collection, making the task considerably more difficult. Fig 2. explains in this part, we focus on personalized animation, which is precisely defined as: given a personalized T2I moded device, such as a DreamBooth [24], LoRA [13] checkpoint taught by users or taken from CivitAI [4] or Huggingface [8]), the objective is to convert it into an animation generator with minimal or no training cost while retaining its original domain knowledge and quality. Assume a T2I model is customized for a certain 2D anime style. In such situation, the related animation generator ought to be trained in producing animation clips with appropriate motions, including foreground/background segmentation, character body movements, and so on.

One naive technique would be to inflate a T2I model [7, 12, 33] through including temporal-aware structures and learning suitable motion priors from large-scale video datasets. However, gathering enough personalized movies for personalized domains is expensive. In the meantime, a lack of data would result in a loss of understanding in the source domain. As a result, we decide to train a generalizable motion modeling module independently and then insert it into the personalized T2I at inference time. By doing so, we avoid having to tune each personalized model individually and maintain their expertise by leaving the pre-trained weights alone. A further important benefit of this technique is that, once trained, the module can be introduced into any personalized T2I based on the same basic model with no requirement for particular tweaking, as demonstrated in the subsequent trials. This is due to the fact that the personalizing process little affects the feature space of the original T2I model, as proved by ControlNet[32].

3.3. Motion Modeling Module

3.3.1. Network Inflation: While the initial SD can only analyze picture data in batches, model inflation is required in order to make it suitable with our motion modeling module that accepts as inputs a 5D video tensor in the style of batchchannels × frames × height × width. In order to accomplish the above, we use a method comparable to the Video Diffusion Model [12]. In particular, we reshape the frame axis into the entire batch axis and enable the network to go through each frame individually to convert each 2D convolution and attention layer in the original picture model into spatial only pseudo-3D layers. In contrast to the previous example, our newly included motion module acts across frames in every batch to create motion smoothness and content uniformity in the animation clips. Figure 3 shows further information.

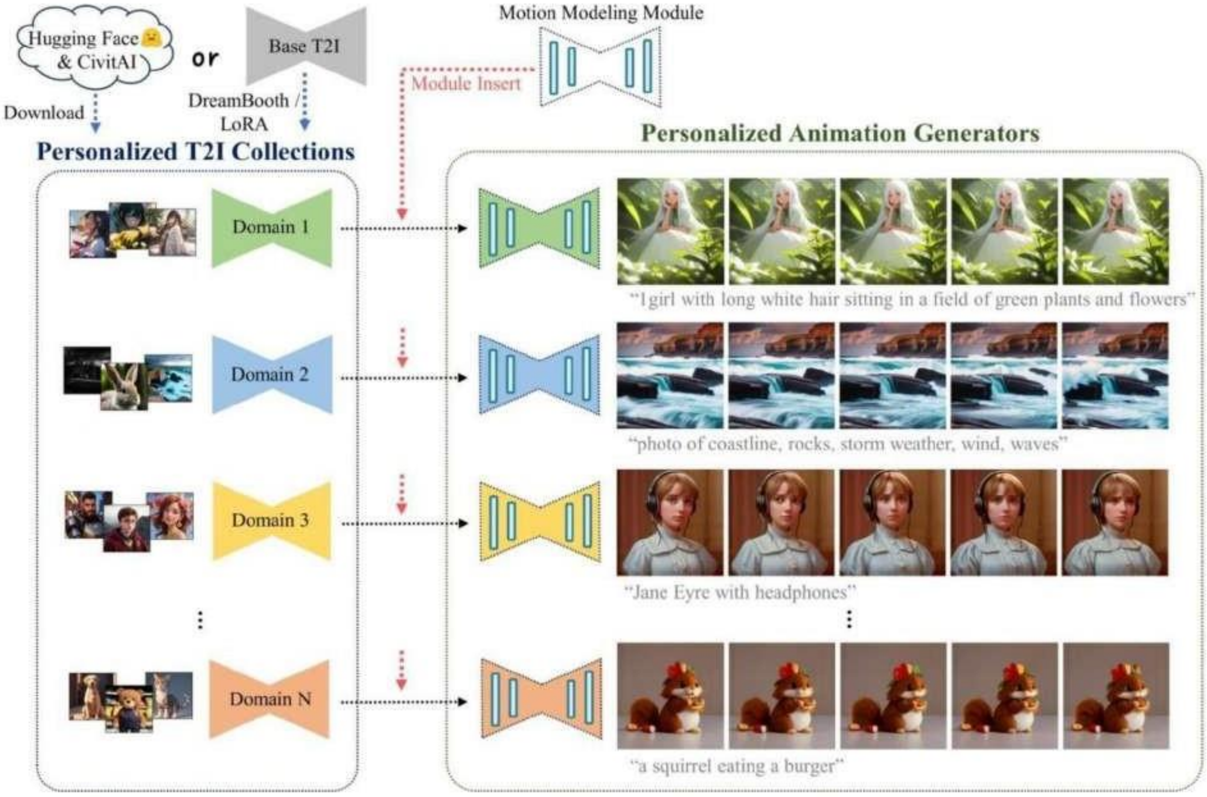

Hugging Face
& CivitAI
or
Base T2I
Download
DreamBooth / LoRA
Personalized T2I Collections
Motion Modeling Module
Module Insert
Personalized Animation Generators
Domain 1
Domain 2
Domain 3
Domain N
"1girl with long white hair sitting in a field of green plants and flowers"
"photo of coastline, rocks, storm weather, wind, waves"
"Jane Eyre with headphones"
"a squirrel eating a burger"

Fig 2. shows AnimateDiff, a powerful framework for transforming personalized text-to-image (T2I) models into animation generators without requiring model-specific tweaking. Once motion priors have been learnt from huge video datasets, AnimateDiff may be placed into personalized T2I models that have been trained by the user or obtained directly from services like CivitAI [4] or Huggingface [8] to make animation clips with suitable movements.

3.3.2. Module Design: We want our motion modeling module's network design to allow for efficient information transmission between frames. To do this, we designed our motion module using vanilla temporal transformers. It is worth mentioning that we have also tested alternative network topologies for the motion module and discovered that a simple temporal transformer is sufficient for modeling the motion priors. We will look for improved motion modules in future efforts. The vanilla temporal transformer is made up of a number of self attention blocks that operate along the temporal axis (Fig. 3).

When the feature map z passes through our motion module, the spatial dimensions height and width are first reshaped according to the batch dimension, resulting in batch × height × width sequencing at a number of frames. The altered feature map will next be projected and subjected to a series of self-attention blocks, i.e.,

$$\mathbf{z = Attention\ (Q, K, V\ ) = Softmax\ (\ QKT\ /\sqrt{\ d}\ ) \cdot V} \text{ --------------------- } \mathbf{(4)}$$

where, Q = WQz, K = W Kz, and V = WV z are three projections of the reshaped feature map. This procedure allows the component to record the temporal relationships between features that are in the same place throughout the temporal axis. We put our motion module at each resolution level of the U-shaped diffusion network to broaden its receptive field. Furthermore, we add sinusoidal position encoding [30] to the existing self-attention blocks to allow the network to be trained aware of the temporal placement of the current frames in the animation clip. To introduce our module without causing any problems during training, we zero initialise the temporal transformer's output projection layer, this is an effective practice proven by ControlNet [32].

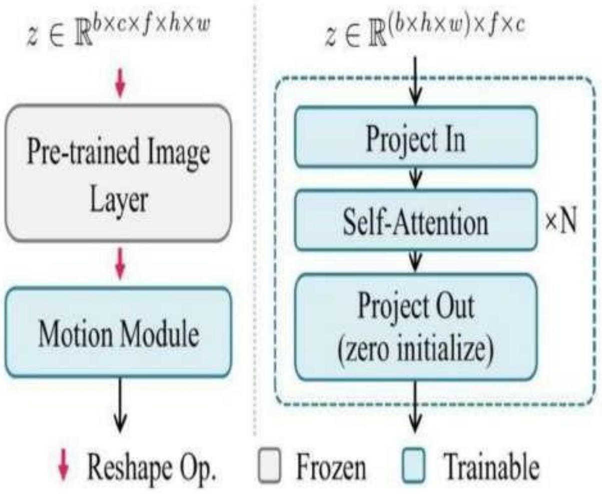

$z \in \mathbb{R}^{b\times c\times f\times h\times w}$
Pre-trained Image Layer
Motion Module
$z \in \mathbb{R}^{(b\times h\times w)\times f\times c}$
Project In
Self-Attention
×N
Project Out (zero initialize)
Reshape Op.
Frozen
Trainable

Fig 3. Motion Module Specifications. Insertion of a module (left): Our motion modules are placed between the picture layers that have already been taught. When a data batch is processed by the picture layers and our motion module, its temporal and spatial axes transform individually into the batch axis. Module layout (right): Our module is a simple temporal transformer with a project layer with a zero initialization.

3.3.3. Training Objective: Our motion modeling module's training procedure is comparable to the Latent Diffusion Model [22]. Sampled video data x 1: N 0 are first encoded into the latent code z 1:N 0 frame by frame via the pre-trained autoencoder. Then, the latent codes are noised using the defined forward diffusion schedule: z 1:N t = √ α¯tz 1:N 0 + √ 1 − α¯tϵ. The diffusion network inflated using our motion modules receives as inputs learning noised latent codes and related text prompts and forecasts the amount of noise intensity added to the latent code, which is encouraged by the L2 loss term. Our motion modeling module's ultimate training goal is:

$$\mathbf{L = EE(x\ 1{:}N\ 0\ ),y,\epsilon \sim N(0,I\ ),t\ \Vert \epsilon - \epsilon\theta(z\ 1{:}N\ t\ ,\ t,\ \tau\theta(y))\Vert\ [2/\ 2]} \text{ ------------------------------ } \mathbf{(5)}$$

It should be noted that the pre-trained weights of the base T2I models are frozen throughout optimization to maintain its feature space unaltered.

4. EXPERIMENT OUTCOMES

4.1. Implementation Details

4.1.1. Training: Since most public personalized models depend on this version of the software, we picked Stable Diffusion v1 as our basis model to train the motion modeling module. The WebVid-10M [1], a text-video pair dataset, was used to train the motion module. The video clips within the dataset are initially sampled at stride 4, then shrunk and center-cropped to 256 X 256 resolutions. Our tests indicate that the 256-resolution module may be generalized to higher resolutions. As a result, we picked 256 as our training level since it strikes a good mix between training efficiency and visual quality. The final duration of the training video segments was set to 16 frames. Throughout the studies, we noticed that employing a diffusion schedule that is somewhat distinct from the original schedule on which the basic T2I model was trained aids in achieving higher visual quality and avoiding artifacts such as weak saturability and flickering. We hypothesize that slightly altering the initial schedule will aid the model's adaptation to new tasks (animation) and data distribution. Thus, we used a linear beta schedule, where βstart = 0.00085 and βend = 0.012, which is slightly different from that used to train the original SD.

4.1.2. Evaluations: To test the efficacy and generalization of our strategy, we gathered numerous typical personalized Stable Diffusion models (Tab. 1) from CivitAI [4,] a public platform where artists may post their personalized models. These models' domains span from anime and 2D cartoon pictures to realistic photos, offering a broad standard for evaluating our method's competence. We integrate our trained module into the target personalized models and produce animations with pre-programmed text cues. We do not utilize standard text prompts since personalized models only create expected material with specific text distribution, which means the prompts must take special forms or contain "trigger words."

Model Name	Domain	Type
Counterfeit	Anime	DreamBooth
ToonYou	2D Cartoon	DreamBooth
RCNZ Cartoon	3D Cartoon	DreamBooth
Lyriel	Stylistic	DreamBooth
InkStyle	Stylistic	LoRA
GHIBLI Background	Stylistic	LoRA
majicMIX Realistic	Realistic	DreamBooth
Realistic Vision	Realistic	DreamBooth
FilmVelvia	Realistic	LoRA
TUSUN	Concept	LoRA

Table 1 shows the personalised models that were utilised for assessment. For our evaluation, we selected many typical personalised models supplied by artists from CivitAI [4], spanning a wide range of domains from 2D animation to realistic photography.

4.2. Qualitative Results

Fig 4 shows some qualitative outcomes from several models. We only show four frames of each animation clip due to space constraints. We strongly advise readers to visit our site for improved visual quality. The image demonstrates how our technology successfully animates personalized T2I models in a variety of domains, ranging from highly styled anime (1st row) to realistic photos (4th row), without sacrificing their domain expertise.

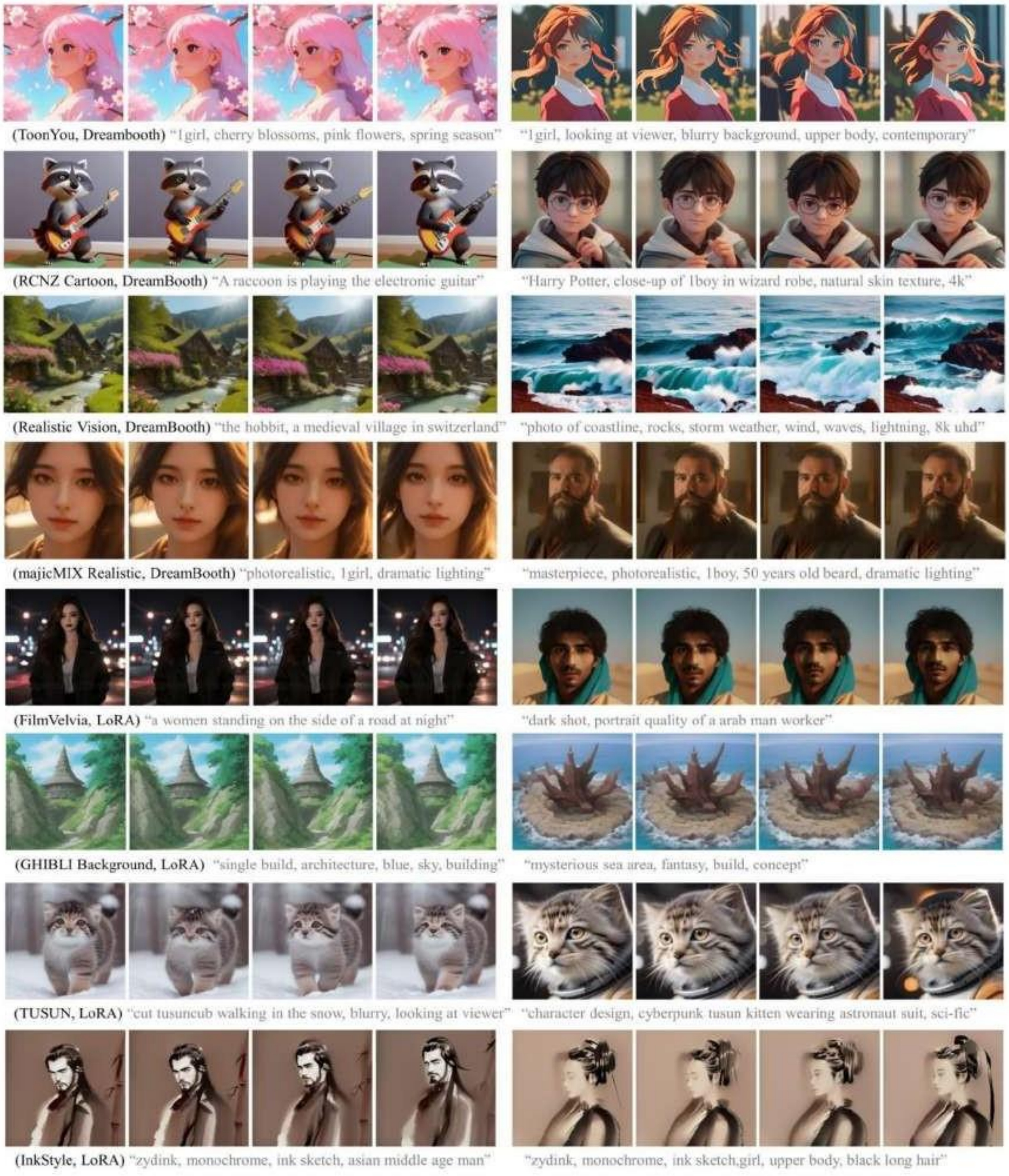
The motion modeling module can grasp the verbal prompt and assign relevant movements to each pixel, such as the motion of sea waves (3rd row) and the leg motion of the Pallas's cat (7th row), owing to the movement priors learnt from the video datasets. We also discovered that our approach can separate important themes in the picture form the foreground and background, generating a sense of brightness and realism. In the first animation, for example, the character and backdrop blooms move individually, at various speeds along with varying blurring intensities. Our qualitative findings show that our motion module is generalizable for animating personalized T2I models across domains. AnimateDiff can produce high-quality animations that are loyal to the personalized domain while still being diverse and aesthetically appealing by including our motion module into the personalized model.

4.3. Comparison with Baselines

Our technique is compared to Text2Video-Zero [15], a training-free methodology for extending a T2I model for video creation via network inflation and latent warping. Although Tune-a-Video may be used for personalized T2I animation, it needs an additional video input and hence is not included in the comparison. Because T2V-Zero does not need parameter adjustment, it is simple to use it to animate personalized T2I models simply substituting the model weights with personalized ones.

Using the authors' default hyper-parameters, we make animation clips of 16 frames at 512 X 512 resolutions. We evaluate the baseline's and new technique's cross-frame content consistency on the identical personalized model and prompt ("A forbidden castle situated high in the hills, pixel art, complex details2, hdr, intricate details"). To more effectively depict as well as contrast the fine-grained features of our approach and the baseline, we chopped and zoomed in on the same subpart of each result, as seen at the left/right bottom of each frame in Fig. 5.

As seen in the picture, both strategies maintain the personalized model's domain knowledge, and their frame-level features are comparable. When studied attentively, the T2V-Zero result, while visually identical, lacks fine-grained cross-frame consistency. For example, the form of the pebbles in the foreground (1st row) and the cup on the table (3rd row) varies with time. When the animation is shown as a video clip, this discrepancy becomes much more apparent. Our technique, on the other hand, delivers temporally consistent material while maintaining greater smoothness (2nd, 4th row). Furthermore, our solution demonstrates more suitable content modifications that correlate better with the underlying camera motion, demonstrating its usefulness even further. This is a plausible outcome since the baseline technique fails to acquire motion priors and achieves visual consistency through rule-based latent warping, whereas our method inherits information from huge video datasets and maintains temporal smoothness via efficient temporal attention.



(ToonYou, Dreambooth) "1girl, cherry blossoms, pink flowers, spring season" "1girl, looking at viewer, blurry background, upper body, contemporary"

(RCNZ Cartoon, DreamBooth) "A raccoon is playing the electronic guitar" "Harry Potter, close-up of 1boy in wizard robe, natural skin texture, 4k"

(Realistic Vision, DreamBooth) "the hobbit, a medieval village in switzerland" "photo of coastline, rocks, storm weather, wind, waves, lightning, 8k uhd"

(majicMIX Realistic, DreamBooth) "photorealistic, 1girl, dramatic lighting" "masterpiece, photorealistic, 1boy, 50 years old beard, dramatic lighting"

(FilmVelvia, LoRA) "a women standing on the side of a road at night" "dark shot, portrait quality of a arab man worker"

(GHIBLI Background, LoRA) "single build, architecture, blue, sky, building" "mysterious sea area, fantasy, build, concept"

(TUSUN, LoRA) "cut tusuncub walking in the snow, blurry, looking at viewer" "character design, cyberpunk tusun kitten wearing astronaut suit, sci-fic"

(InkStyle, LoRA) "zydink, monochrome, ink sketch, asian middle age man" "zydink, monochrome, ink sketch,girl, upper body, black long hair"

Fig 4.Qualitative outcomes. In this section, we show 16 animation clips made by models injected using our framework's motion modeling module. Each row's two samples are from the same personalized T2I model. Therefore merely sample four frames in each animation clip because of space constraints, and we encourage that readers visit our project website for a better perspective. For understanding, irrelevant tags such as "masterpieces" and "high quality" are removed from each prompt.

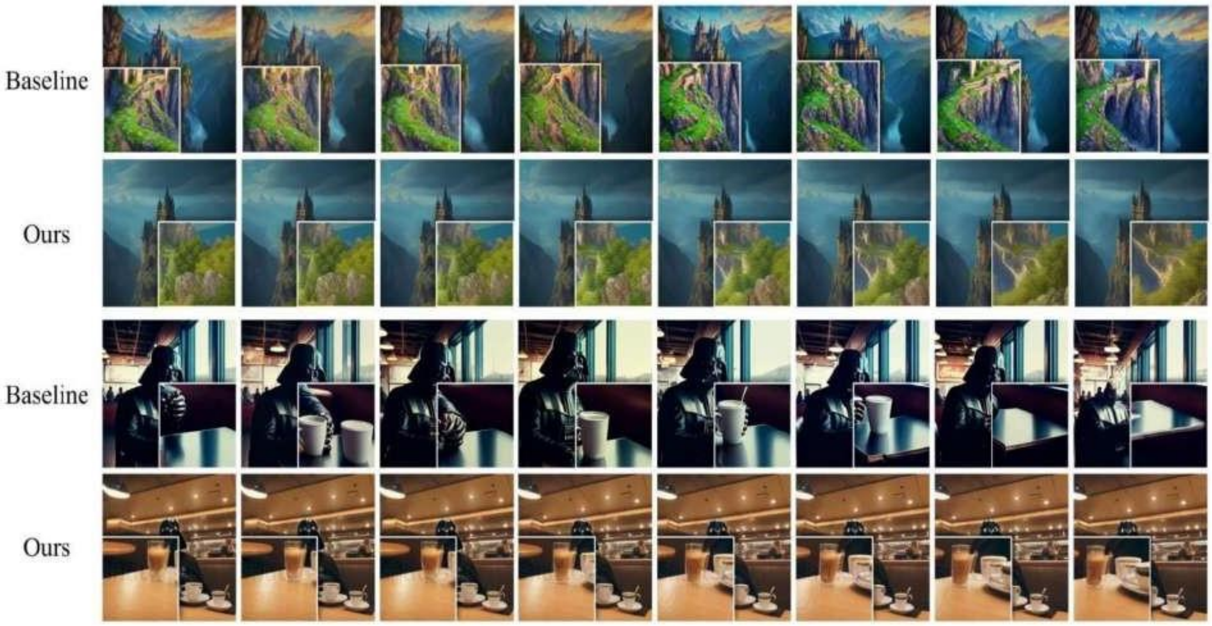


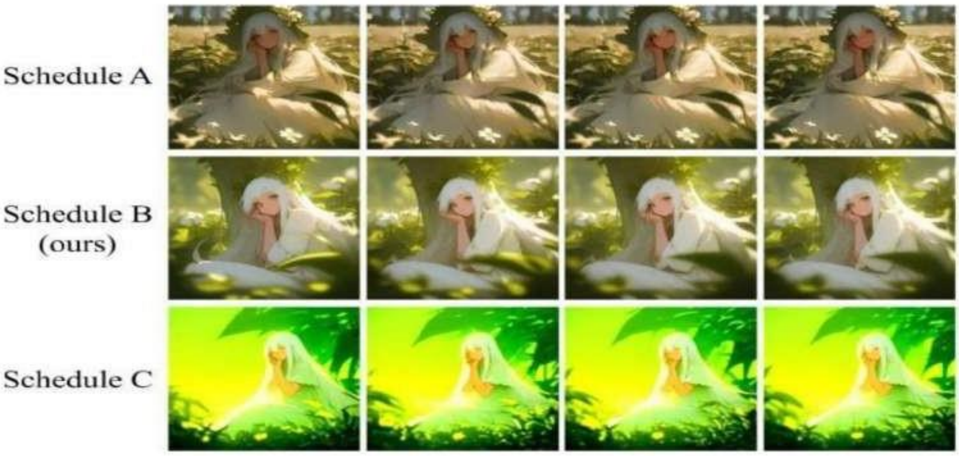
Fig 5. Shows a baseline comparison. We evaluate the cross-frame content consistency of the baseline (1st, 3rd row) with our technique (2nd, 4th row) qualitatively. While the baseline findings lack fine-grain consistency, our technique preserves more temporal smoothness.

4.4. Ablative Study

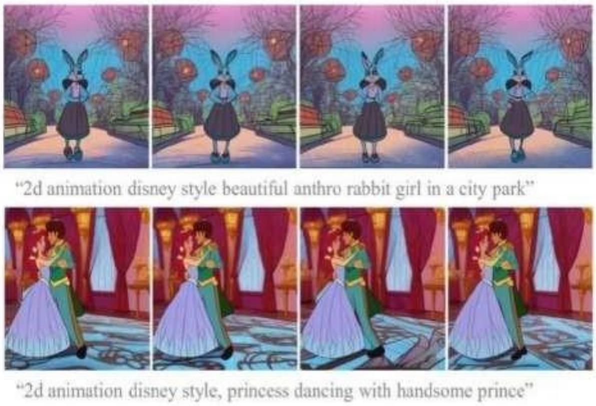
We do an ablative investigation to validate our noise schedule selection in the forward diffusion process during training. We discussed in the last section that employing a slightly adjusted diffusion plan helps create higher visual quality. In this section, we experiment using three example diffusion schedules (Tab. 2) used in prior research and visually contrast the results in Fig. 6.

Configuration	Schedule	β_{start}	β_{end}
Schedule A (SD)	*scaled linear*	0.00085	0.012
Schedule B (ours)	*Linear*	0.00085	0.012
Schedule C	*Linear*	0.0001	0.02

Table 2 shows three different diffusion schedule configurations used in our ablative tests.



Schedule A corresponds to the pre-training Stable Diffusion schedule.
Fig 6. An Ablative analysis. We evaluate the outcomes of three diffusion schedules, all of which have a distinct deviation level from the schedule whereby Stable Diffusion was pre-trained.



"2d animation disney style beautiful anthro rabbit girl in a city park"

"2d animation disney style, princess dancing with handsome prince"

Figure 7. A Ablative analysis. We test having three diffusion schedules, each with a distinct deviation level from the schedule where Stable Diffusion was pre-trained, and then evaluate the outcomes qualitatively.

In Schedule A is the pre-training Stable Diffusion schedule; Schedule B is our choice, which differs from SD's schedule in how the beta sequence is computed; and Schedule C is used in DDPM [5] and DiT [18], and differs even more from SD's pre-training schedule. As seen in Fig. 7, while employing the original SD schedule for training our motion modeling module (Schedule B), the animation results include sallow color artifacts. This is an odd phenomenon because, on the surface, utilizing the diffusion schedule in conjunction with pre-training should benefit the model's ability to retain previously learnt feature space. The color saturation of the produced animations grows as the schedules diverge further from the pre-training plan (from plan A to Schedule C), but the range of motion diminishes. Among these three options, our preference delivers a good mix of visual quality and motion fluidity.

A review of these findings, we hypothesize that a somewhat changed diffusion schedule during the training stage aids the pre-trained model's adaptation to novel tasks and domains. The latest training goal for our architecture is to rebuild noise sequences using a dispersed video stream. This may be performed frame by frame, without taking into account the temporal structure of the video sequence, resulting in the image reconstruction job that the T2I models was pre-trained on. Utilizing the same diffusion schedule may trick the model into thinking it is still optimized for image reconstruction, slowing up the training effectiveness of our motion modeling modules that handles cross-frame motion modeling and leading in more flickering animation and color aliasing.

5. OUTCOMES

In our research, we found that the majority of failure occurrences occur when the domain of the personalized T2I model is far from realistic, such as a 2D Disney cartoon (Fig. 7). In some circumstances, the animation outputs show visible artifacts and are incapable of producing appropriate motion. We hypothesize that this is owing to the significant distribution gap between the realistic training video and the personalized model. We left it to future efforts to manually gather multiple films in the target domain and somewhat fine-tune the motion modeling module as a possible solution to this problem.

6. CONCLUSION

In this study, we introduce AnimateDiff, an efficient platform for allowing personalized text-to-image model animation, with the goal of permanently converting the vast majority of current personalized T2I models into animation generators. We show how our approach, which incorporates a straightforward motion modeling module trained on base T2I, can extract generalizable motion priors from huge video datasets. Our motion module, once trained, may be put into other personalized models to produce animated visuals with natural and appropriate movements while being loyal to the related domain. Extensive testing on numerous personalized T2I models supports our method's efficiency and generalizability. As a result, AnimateDiff provides a simple yet efficient foundation for personalized animation, which might benefit a wide range of applications.

7. REFERENCES

[1] Max Bain, ARsha Nagrani, etc., Frozen in time: " A joint Video and image Encoder for end-to-end retrievel. In Proceedings of IEEE/CVF International Conference on Computer vision, Pages 1728-1738, 2021,2,5.

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022. 3

[3] Andreas Blattmann, Andreas and Rombach, et.,al., "High-Resolution Video Synthesis with Latent Diffusion Models",IEEE Conference on Computer Vision and Pattern Recognition ,2023, pages 22563–22575, 2023. 3.

[4] Civitai. Civitai. https://civitai.com/, 2022. 1, 2, 4, 7

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34:8780–8794, 2021. 2, 3, 8

[6] Ziyi Dong, Pengxu Wei, and Liang Lin. Drea- martist: Towards controllable one-shot text-to-image gen- eration via contrastive prompt-tuning. arXiv preprint arXiv:2211.11337, 2022. 3

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. arXiv preprint arXiv:2302.03011, 2023. 2, 3, 4

[8] Hugging Face. Hugging face. https://huggingface. co/, 2022. 1, 2, 4

[9] Gal, Rinon and Alaluf, Yuval and Atzmon, Yuval and Patashnik, et.,al., "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion, https://arxiv.org/abs/2208.01618, doi :10.48550/ARXIV.2208.01618,

[10] Rinon Gal and Moab Arar and Yuval Atzmon et.,al.," Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Mod", preprint arXiv: 2302.12228, 2023.3.

[11] Jonathan Ho and Ajay Jain and Pieter Abbeel , "Denoising Diffusion Probabilistic Models", Advances in Neural Information Processing Systems, arXiv: 2006.11239,2020.

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video dif- fusion models. arXiv preprint arXiv:2204.03458, 2022. 2, 3, 4, 5

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen- Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.1,2, 3,4

[14] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642, 2023. 3

[15] Levon Khachatryan, Andranik Movsisyan, Vahram Tade- vosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to- image diffusion models are zero-shot video generators. arXiv preprint arXiv:2303.13439, 2023. 3, 8.

[16] Nupur Kumari and Bingliang Zhang et.al., "Multi-Concept Customization of Text-to-Image Diffusion",. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, arXiv: 2212.04488, pages 1931–1941, 2023. 3

[17] Alex Nichol and Prafulla Dhariwal and Aditya Ramesh and Pranav Shyam .et.al., "Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models", preprint arXiv: 2112.10741, 2021. 2.

[18] William Peebles and Saining Xie. Scalable diffusion models with transformers. arXiv preprint arXiv:2212.09748, 2022,8.

[19] Alec Radford and Jong Wook Kim and Chris Hallacy and Aditya Ramesh and Gabriel Goh ., et al. "Learning Transferable Visual Models From Natural Language Supervision", In International conference on machine learning, pages 8748–8763, preprint arXiv: 2103.00020, 2021. 2, 3.

[20] Colin Raffel and Noam Shazeer and Adam Roberts .et.al.,, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", The Journal of Machine Learning Research, arXiv: 1910.10683, 21(1):5485–5551, 2020. 2.

[21] Aditya Ramesh and Prafulla Dhariwal and Alex Nichol and Casey Chu and Mark Chen ., "Hierarchical Text-Conditional Image Generation with CLIP Latents", arXiv preprint arXiv:2204.06125, 2022. 2.

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjo¨rn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022. 1, 2, 3, 4, 5.

[23] Olaf Ronneberger and Philipp Fischer and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segment", arXiv: 1505.04597 : 2015. 3.

[24] Nataniel Nataniel Ruiz and Yuanzhen Li and Varun Jampani et.al.,. "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, pages 22500–22510, 2023. 1, 2, 3, 4

[25] Chitwan Saharia and William Chan and Saurabh Saxena.,et al., Photorealistic "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding ", Advances in Neural Information Processing Systems, 35:36479–36494, 2022,2.

[26] Christoph Schuhmann and Romain Beaumont and Richard Vencu et al. "LAION-5B: An open large-scale dataset for training next generation image-text models", arXiv preprint arXiv:2210.08402, 2022. 2.

[27] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instant- booth: Personalized text-to-image generation without test- time finetuning. arXiv preprint arXiv:2304.03411, 2023. 3.

[28] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al., “Make-a-video: Text-to-video generation without text-video data”, arXiv preprint arXiv:2209.14792, 2022,3.

[29] van den Oord, Aaron and Vinyals, Oriol and kavukcuoglu, koray.,“Neural Discrete Representation Learning”. , Advances in Neural Information Processing Systems, volume 30. Curran Associates,Inc., 2017. 30.

[30] Ashish Vaswani and Noam Shazeer and Niki et.al.,” Attention Is All You Need”, Advances in neural information processing systems, 30, 2017. 5

[31] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565, 2022. 3.

[32] Lvmin Zhang and Maneesh Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models”, arXiv preprint arXiv:2302.05543, 2023. 5.

[33] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018, 2022