

# DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing

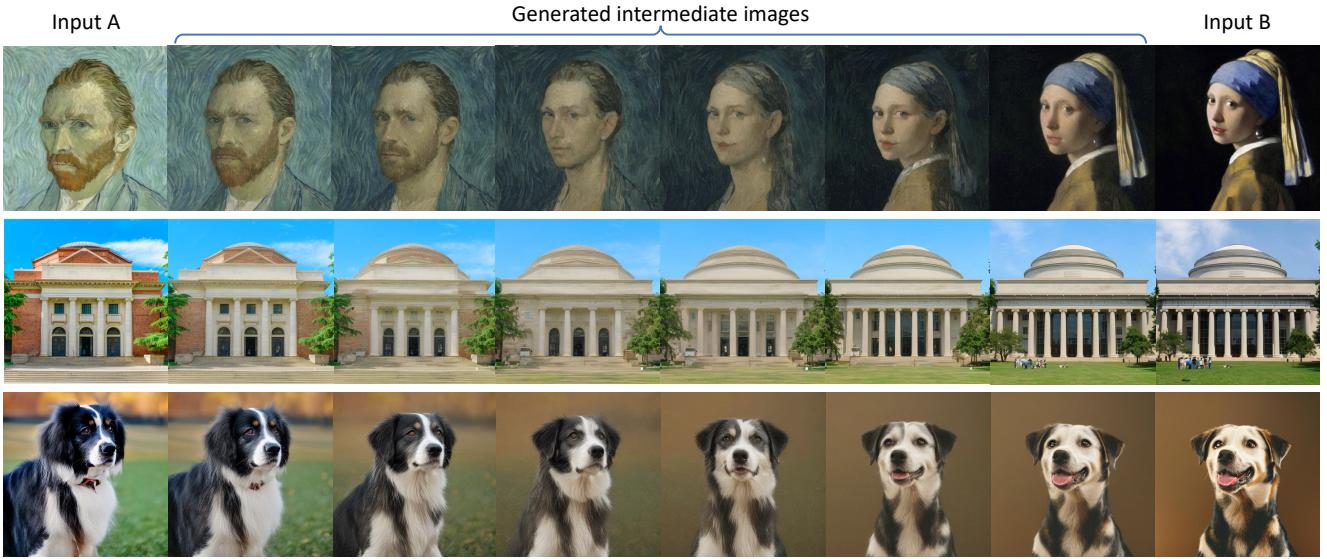
Kaiwen Zhang<sup>1,2\*</sup>Yifan Zhou<sup>3</sup>Xudong Xu<sup>2</sup>Xingang Pan<sup>3</sup> Bo Dai<sup>2</sup><sup>1</sup>Tsinghua University<sup>2</sup>Shanghai AI Laboratory<sup>3</sup>S-Lab, Nanyang Technological UniversityProject page: [https://kevin-thu.github.io/DiffMorpher\\_page/](https://kevin-thu.github.io/DiffMorpher_page/)

Figure 1. Given two input images, our approach can generate a sequence of intermediate images, delivering a smooth and natural transition between them. This is achieved purely by harnessing the prior knowledge of a pre-trained diffusion model, *i.e.*, Stable Diffusion [36].

## Abstract

Diffusion models have achieved remarkable image generation quality surpassing previous generative models. However, a notable limitation of diffusion models, in comparison to GANs, is their difficulty in smoothly interpolating between two image samples, due to their highly unstructured latent space. Such a smooth interpolation is intriguing as it naturally serves as a solution for the image morphing task with many applications. In this work, we address this limitation via DiffMorpher, an approach that enables smooth and natural image interpolation by harnessing the prior knowledge of a pre-trained diffusion model. Our key idea is to capture the semantics of the two images by fitting two LoRAs to them respectively, and interpolate between both the LoRA parameters and the latent noises

to ensure a smooth semantic transition, where correspondence automatically emerges without the need for annotation. In addition, we propose an attention interpolation and injection technique, an adaptive normalization adjustment method, and a new sampling schedule to further enhance the smoothness between consecutive images. Extensive experiments demonstrate that DiffMorpher achieves starkly better image morphing effects than previous methods across a variety of object categories, bridging a critical functional gap that distinguished diffusion models from GANs.

## 1. Introduction

Image morphing [1, 53, 59] is a popular technique for image transformation, lying at the intersection of computer vision and computer graphics with continuous attention over decades. Given two images of topologically similar objects and optionally a set of correspondence key points, a morphing process generates a sequence of reasonable interme-

\*Work done during internship at Shanghai AI Laboratory.

 Corresponding Author.

diary images. When played in succession, the image sequence produces a captivating video of a smooth transition between the two input images. Developed initially for cinematic and visual effects, image morphing has found its applications in various fields like animations, games [1, 59], as well as photo-editing tools [1] for artistic and entertainment purpose to enrich people’s imagination. In the era of deep learning, image morphing can also be used in data augmentation [12].

There are two major concerns in the problem of image morphing, which are hardly balanced in previous studies: the *rationality* of intermediate images and the *smoothness* of the transition video. Classic methods presented in the graphics literature [4, 5, 9, 28, 42] typically involve an image-warping process to align the correspondence points and a cross-dissolution operation to mix the colors. However, the color-space dissolution does not well explain the textural and semantic transition and is prone to undesirable intermediate results like ghosting artifacts. Since the deep learning era, GANs [13] have shown stunning image morphing ability through simple latent code interpolations [6, 22–24, 32, 39, 40]. Despite the smoothness of the transformation process, this method is hard to extend to arbitrary real-world images due to the limited model capacity and the challenge of GAN inversion [54]. Recently, diffusion models [3, 18, 36, 45, 47] have emerged as state-of-the-art generative models, significantly enhancing image synthesis and real image reconstruction. However, initial attempts to apply diffusion models to image morphing suffer from abrupt content changes between consecutive images.

In this work, we are interested to address the image morphing problem by asking the question: *Is it feasible to achieve smooth and natural image interpolation with diffusion models, akin to the capabilities of GANs?* The solution to this problem will immediately serve as an image morphing approach when combined with image reconstruction techniques like DDIM inversion [46]. However, realizing such a reasonable interpolation on diffusion models is non-trivial. Unlike GANs that have a meaningful compact latent space, the latent space of diffusion models is a noise map that lacks semantic meaning, thus random and abrupt content flickering are often observed when naively interpolating in the latent space. How to guarantee smoothness in both high-level semantics and low-level textures remains a key challenge.

To this end, we present *DiffMorpher*, a new approach to achieve *smooth* image interpolation based on diffusion models while maintaining the *rationality* of intermediate images. Since the latent space is non-interpretable, our key idea is to create smooth semantic transition via the *low-rank parameter space*. This is achieved by applying low-rank adaptations (LoRAs) [20] to the two input images separately, encapsulating the corresponding image semantics

in the two groups of LoRA parameters. Thanks to the analogous parameter structures, a linear interpolation between the two sets of LoRA parameters will deliver a smooth transition in the image semantics. Combining spherical interpolation (slerp) between the two latent Gaussian noises associated with the two input images, our approach can create a semantically meaningful transition with high-quality intermediates between them. However, this method does not fully eliminate the low-level abrupt change. To address this, we further introduce a self-attention interpolation and substitution method that ensures smooth transition in low-level textures, and an AdaIN adjustment technique that enhances the coherence in image colors and brightness. Finally, to maintain a homogeneous transition speed in image semantics, we propose a new sampling schedule.

We extensively evaluate *DiffMorpher* in a wide range of real-world scenarios. A new image morphing benchmark *MorphBench* is created to support quantitative evaluation, where our approach significantly outperforms existing methods in both smoothness and image fidelity. To the best of our knowledge, this is the first time smooth image interpolation can be achieved on diffusion models at a comparable level as GANs. Unlike GANs that struggle with real-world images, *DiffMorpher* can deal with a much wider image scope. The ability to continuously tweak image semantics has empowered GAN for many downstream applications, thus we hope our work will similarly pave the way for new opportunities in diffusion models. For example, our method can augment many image editing methods such as [7, 26, 31, 58] by turning their final images into continuous animations.

## 2. Related Work

### 2.1. Classic Image Morphing

Image morphing is a long-standing problem in computer vision and graphics [1, 53, 59]. Classic graphical techniques [4, 5, 9, 28, 42] typically combine correspondence-driven bidirectional image warping with blending operations to obtain plausible in-betweens in a smooth transition. Although making a smooth morphing between two images, these methods fall short of creating new content beyond the given inputs, thus leading to unsatisfactory results like ghosting artifacts. More recently, the explosion of data volume gave rise to a new data-driven morphing paradigm [2, 12]. Unlike classic approaches, they capitalize on massive images from a specific object class to determine a smooth transition path from the source image to the target one, which contributes to compelling intermediate morphing results. However, the great demand for enormous single-class data impedes their applications in more general scenarios like cross-domain or personalized morphing. In contrast, our model leverages the prior knowledge in diffu-

sion models pre-trained on large-scale images and thus is applicable to diverse object categories.

## 2.2. Image Editing via Diffusion Models

Diffusion models [18, 45, 47] have been a prevalent star in deep generative models in recent years, thanks to their impressive sample quality and scaling ability [11, 19]. By learning to gradually denoise from Gaussian noises with a noise prediction UNet [37], diffusion models can generate high-quality clean data that fits well with real data distribution. Diffusion models trained on large-scale text-image pairs [41], such as Imagen [38] and Stable Diffusion [36], have gained unprecedented success in text-to-image generation. Therefore, they are suitable as a powerful prior for multiple editing tasks, including text-guided [7, 8, 15, 26, 33, 49] and drag-guided [31, 43] image manipulation. Most of these works directly generate the final edited image, while the generation of a continuous animation like image morphing is much less explored in the literature of image diffusion models.

## 2.3. Deep Interpolation

It has been widely demonstrated that Generative Adversarial Networks (GANs) [13] can be used to morph images by interpolating latent codes. Due to their highly continuous and discriminative latent embedding space, a linear interpolation among two latent codes will exhibit impressive image morphing results, as demonstrated in a large body of GAN papers [6, 22–24, 39, 40]. However, to morph between real images, the corresponding latent codes, which are often outside of GAN’s latent distribution, must be obtained with GAN inversion and tuning techniques [32, 35, 54]. Typically, the latent codes obtained struggle to recover the original real images. Although the generator can be tuned to reconstruct the images, the rationality of the intermediates and the correctness of correspondence cannot be guaranteed.

Recent advances in diffusion models [18, 45, 47] also show the potential to generate reasonable intermediate images through latent noise interpolations and text embedding interpolations [3, 51]. However, due to the highly unstructured image distribution learned in diffusion models, the generated transition videos often contain abrupt changes and inconsistent semantic content, which are unacceptable in the image morphing task. Preechakul *et al.* [34] proposed a Diffusion Autoencoder architecture that enables more reasonable image interpolation than the vanilla diffusion models, but this approach cannot be directly applied to the widely used vanilla diffusion models like Stable Diffusion and abrupt changes still remain. In our work, we demonstrate the ability of diffusion models to generate smooth and natural morphing sequences using only the prior knowledge in pretrained text-to-image models.

Recently, a concurrent work [55] has also studied the application of diffusion models for the image morphing task. Compared to their approach, our method incorporates delicately designed self-attention control and AdaIN adjustment, which greatly diminish abrupt changes in textures and improve consistency in colors. Furthermore, our approach fits a single LoRA for each image and interpolates between the LoRA parameters during morphing, thus increasing the versatility and flexibility of our method, such as applying morphing among multiple images.

## 3. Method

Given two images  $\mathcal{I}_0$  and  $\mathcal{I}_1$ , our goal is to obtain an interpolation video  $\mathcal{V} = \{\mathcal{I}_\alpha | \alpha \in (0, 1)\}$  that displays a natural and smooth transition from  $\mathcal{I}_0$  to  $\mathcal{I}_1$ , where the sequence of  $\alpha$  depends on the desired number of frames  $n$  and a specific sampling schedule. A meaningful image morphing should be done between two images with clear correspondence. In our general morphing framework,  $\mathcal{I}_0$  and  $\mathcal{I}_1$  can be either real images or diffusion-generated images with text prompts  $\mathcal{P}_0$  and  $\mathcal{P}_1$ .

In this section, we formally present our *DiffMorpher* approach to address this problem. We first introduce the preliminaries on diffusion models in Sec. 3.1. To capture the identities in  $\mathcal{I}_0, \mathcal{I}_1$  and generate semantic consistent and meaningful in-betweens, we propose LoRA interpolation and latent noise interpolation techniques in Sec. 3.2 and 3.3. To enhance the smoothness of the transition video, we propose the self-attention interpolation and replacement method, a AdaIN adjustment technique and a new reschedule method in Sec. 3.4, Sec. 3.5 and 3.6. An overview of our method with an illustration example is shown in Fig. 2.

### 3.1. Preliminaries on Diffusion Models

Diffusion models [18, 45–47] are a family of latent variable generation models of the form:

$$p_\theta(\mathbf{z}_0) = \int p_\theta(\mathbf{z}_{0:T}) d\mathbf{z}_{1:T} \quad (1)$$

It includes a diffusion process  $\{q(\mathbf{z}_t) | t = 0, 1, \dots, T\}$  that gradually adds noise to the data sampled from the real data distribution  $q(\mathbf{z}_0)$  toward  $q(\mathbf{z}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ , and a corresponding denoising process  $\{p(\mathbf{z}_t) | t = T, T-1, \dots, 0\}$  that generates clean data from the standard Gaussian noise  $\mathbf{z}_T \sim p(\mathbf{z}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ , where  $T$  is the total number of steps. The denoising process is achieved by learning a parameterized joint distribution  $p_\theta(\mathbf{z}_{0:T})$  with a noise prediction network  $\epsilon_\theta$ . Specifically, in the denoising step  $t$ ,  $\epsilon_\theta$  predicts the noise  $\epsilon$  added to  $\mathbf{z}_{t-1}$  according to current noise  $\mathbf{z}_t$ , current time step  $t$  and possible additional condition  $c$ . In practice,  $\epsilon_\theta$  is generally implemented as a UNet [37].

Latent Diffusion Model (LDM) [36] is an important variant of diffusion models that achieves a great balance be-

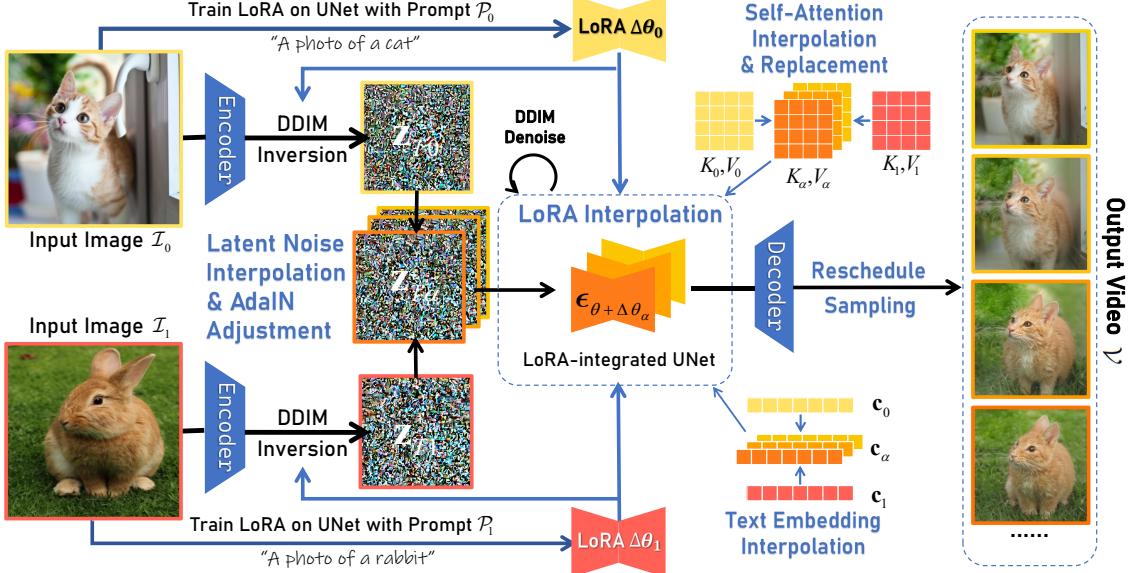


Figure 2. Method pipeline. Given two images  $\mathcal{I}_0$  and  $\mathcal{I}_1$ , two LoRAs are trained to fit the two images respectively. Then the latent noises for the two images are obtained via DDIM inversion. The mean and standard deviation of the interpolated noises are adjusted through AdaIN. To generate an intermediate image, we interpolate between both the LoRA parameters and the latent noises via the interpolation ratio  $\alpha$ . In addition, the text embedding and the  $K$  and  $V$  in self-attention modules are also replaced with the interpolation between the corresponding components. Using a sequence of  $\alpha$  and a new sampling schedule, our method will produce a series of high-fidelity images depicting a smooth transition between  $\mathcal{I}_0$  and  $\mathcal{I}_1$ .

tween image quality and sample efficiency. Based on the LDM framework, a number of powerful pretrained text-to-image models have been available to the public, including the widely-used Stable Diffusion (SD). It involves a variational auto-encoder (VAE) [27] that encodes the images to latent embeddings and trains a text-conditioned diffusion model in the latent space. The denoising UNet  $\epsilon_\theta$  in the SD model is composed of a sequence of basic blocks, each of which includes a self-attention module, a cross-attention module [50], and a residual block [14]. The attention module in UNet can be formulated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where  $Q$  is the query features derived from the spatial features, and  $K, V$  are the key and value features obtained from either the spatial features (in self-attention layers) or the text embedding (in cross-attention layers) with respective projection matrices. Our method in this paper is built upon the SD model.

### 3.2. LoRA Interpolation

Low-Rank Adaption (LoRA) [20] is an efficient tuning technique that was first proposed to fine-tune large language models and recently introduced to the domain of diffusion models. Instead of directly tuning the entire model, LoRA fine-tunes the model parameters  $\theta$  by training a low-rank residual part  $\Delta\theta$ , where  $\Delta\theta$  can be decomposed into low-



Figure 3. Effects of LoRA. A LoRA fit to an image tends to capture its semantic identity, while the layout and appearance are controlled by latent noise.

rank matrices. Besides its inherent advantage in training efficiency, we further discover that LoRA enjoys an impressive capacity to encapsulate high-level image semantics into the low-rank parameter space. By simply fitting a LoRA on a single image, the fine-tuned model can generate diverse samples with consistent semantic identity when traversing the latent noise, as shown in Fig. 3.

Motivated by this observation, we first train two LoRAs  $\Delta\theta_0, \Delta\theta_1$  on the SD UNet  $\epsilon_\theta$  for each of the two images  $\mathcal{I}_0$  and  $\mathcal{I}_1$ . Formally, the learning objective for training  $\Delta\theta_i(i = 0, 1)$  is:

$$\mathcal{L}(\Delta\theta_i) = \mathbb{E}_{\epsilon, t}[\|\epsilon - \epsilon_{\theta+\Delta\theta_i}(\sqrt{\alpha_t}\mathbf{z}_{0i} + \sqrt{1-\alpha_t}\epsilon, t, \mathbf{c}_i)\|^2] \quad (3)$$

where  $\mathbf{z}_{0i} = \mathcal{E}(\mathcal{I}_i)$  is the VAE encoded latent embedding

associated with the input image  $\mathcal{I}_i$ ,  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  is the random sampled Gaussian noise,  $\mathbf{z}_{ti} = \sqrt{\bar{\alpha}_t} \mathbf{z}_{0i} + \sqrt{1 - \bar{\alpha}_t} \epsilon$  is the noised latent embedding at diffusion step  $t$ ,  $\mathbf{c}_i$  is the text embedding encoded from the text prompt  $\mathcal{P}_i$ , and  $\epsilon_{\theta+\Delta\theta_i}$  represents the LoRA-integrated UNet. The fine-tuning objective is optimized separately via gradient descent in  $\Delta\theta_0$  and  $\Delta\theta_1$ .

After fine-tuning,  $\Delta\theta_0$  and  $\Delta\theta_1$  are fixed and stored. When generating the intermediate image  $\mathcal{I}_\alpha$ , we fuse the high-level semantics in  $\mathcal{I}_0$  and  $\mathcal{I}_1$  by applying a linear interpolation in the low-rank parameter space:

$$\Delta\theta_\alpha = (1 - \alpha)\Delta\theta_0 + \alpha\Delta\theta_1 \quad (4)$$

and use the UNet with interpolated LoRA  $\epsilon_{\theta+\Delta\theta_\alpha}$  as the noise prediction network in the denoising steps. Such an interpolated  $\Delta\theta_\alpha$  is meaningful because  $\Delta\theta_0$  and  $\Delta\theta_1$  are moderately fine-tuned from the same initialization and thus are highly correlated. While this idea of deep network interpolation [52] is not new in the literature, this is the first time it has been used for image morphing with diffusion models.

### 3.3. Latent Interpolation

With the noise prediction network, the next step in generating  $\mathcal{I}_\alpha$  is to find the corresponding latent noise  $\mathbf{z}_{T\alpha}$  and the latent text condition  $\mathbf{c}_\alpha$ . To this end, we further introduce latent interpolation.

As tentatively discussed in the DDIM paper [46], a fascinating property of DDIM compared to the original DDPM [18] is its suitability for image inversion and interpolation. Following the idea, we first get the corresponding latent noise  $\mathbf{z}_{T0}, \mathbf{z}_{T1}$  for  $\mathbf{z}_{00}, \mathbf{z}_{01}$  through DDIM inversion, and obtain the intermediate latent noise  $\mathbf{z}_{T\alpha}$  through spherical linear interpolation (slerp) [44]:

$$\mathbf{z}_{T\alpha} = \frac{\sin((1 - \alpha)\phi)}{\sin \phi} \mathbf{z}_{T0} + \frac{\sin(\alpha\phi)}{\sin \phi} \mathbf{z}_{T1} \quad (5)$$

$$\text{where } \phi = \arccos \left( \frac{\mathbf{z}_{T0}^T \mathbf{z}_{T1}}{\|\mathbf{z}_{T0}\| \|\mathbf{z}_{T1}\|} \right).$$

However, the vanilla DDIM inversion is known to suffer from unfaithful reconstruction, especially in real image scenarios [30]. To alleviate this, we utilize LoRA-integrated UNet  $\epsilon_{\theta+\Delta\theta_i}(i = 0, 1)$  when inverse the inputs. Since LoRA has been fine-tuned in the input images, the reconstruction from  $\mathbf{z}_{Ti}$  to  $\mathbf{z}_{0i}$  is much more accurate than before.

Regarding the latent text conditions  $\mathbf{c}_\alpha$ , we find that linear interpolations between aligned input condition  $\mathbf{c}_0$  and  $\mathbf{c}_1$  can serve as meaningful intermediate conditions:

$$\mathbf{c}_\alpha = (1 - \alpha)\mathbf{c}_0 + \alpha\mathbf{c}_1 \quad (6)$$

For example, an interpolation between “day” and “night” will show a gradual transition from daylight to darkness.

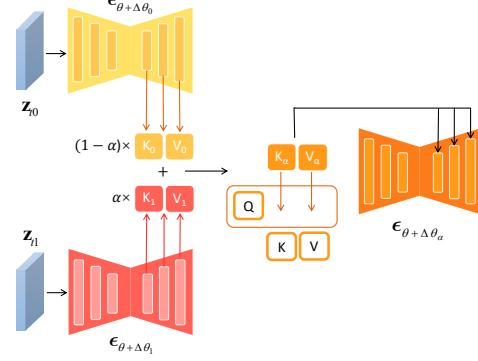


Figure 4. Self-Attention Interpolation and Replacement.

After getting latent noises  $\mathbf{z}_{T\alpha}$  and latent condition  $\mathbf{c}_\alpha$ , we then denoise  $\mathbf{z}_{T\alpha}$  with LoRA-integrated UNet  $\epsilon_{\theta+\Delta\theta_\alpha}$  using the DDIM schedule, and obtain semantically meaningful intermediate images with natural spatial transitions.

### 3.4. Self-Attention Interpolation and Replacement

Despite the semantic rationality of the intermediate results  $\{\mathcal{I}_\alpha\}$ , we still observe unsatisfactory changes in low-level textures in the generated video  $\mathcal{V}$ . We attribute this problem to the highly nonlinear properties introduced in the multi-step denoising process. To address this, we draw inspiration from attention control techniques in previous image editing studies [8, 15, 33, 43, 49], and propose a novel self-attention interpolation and replacement method that introduces linearly changing attention features to the denoising process and greatly reduces abrupt changes in the generated video.

As illustrated in Fig. 4, in the denoising step  $t$ , we first feed the latents of the input images  $\mathbf{z}_{ti}(i = 0, 1)$  into the LoRA-integrated UNet  $\epsilon_{\theta+\Delta\theta_i}$ , to obtain the key and value matrices  $K_i, V_i(i = 0, 1)$  in the self-attention modules of the UNet upsampling blocks. In order to generate an intermediate image  $\mathcal{I}_\alpha$ , we linearly interpolate the matrices to get intermediate matrices:

$$\begin{aligned} K_\alpha &= (1 - \alpha)K_0 + \alpha K_1 \\ V_\alpha &= (1 - \alpha)V_0 + \alpha V_1 \end{aligned} \quad (7)$$

and replace the corresponding matrices in intermediate UNet  $\epsilon_{\theta+\Delta\theta_\alpha}$  with them. Thus, in denoising steps, intermediate latents can query correlated local structures and textures from both input images to further enhance consistency and smoothness.

In particular, we find that replacing attention features in all denoising steps may lead to blurred image textures. Therefore, we only replace the features in the early  $\lambda T(\lambda \in (0, 1))$  steps and leave the self-attention modules unchanged in the remaining steps, to add high-quality details to the images. Empirically, we find that setting  $\lambda$  to  $0.4 \sim 0.6$  works well in most cases.



Figure 5. Qualitative evaluation. Our method generates intermediate images that are significantly more natural and smoother compared to those produced by previous methods.

### 3.5. AdaIN Adjustment

To ensure the coherence in color and brightness between generated images and input images, we additionally introduce the Adaptive Instance Normalization (AdaIN) [21] adjustment for interpolated latent noise  $\mathbf{z}_{0\alpha} (\alpha \in (0, 1))$  before denoising.

Specifically, we calculate the mean  $\mu_i$  and standard deviation  $\sigma_i (i = 0, 1)$  for each channel of latent noises  $\mathbf{z}_{00}, \mathbf{z}_{01}$ , and interpolate between  $\mu_i, \sigma_i$  as the adjustment target of intermediate noises:

$$\mu_\alpha = (1 - \alpha)\mu_0 + \alpha\mu_1 \quad (8)$$

$$\sigma_\alpha = (1 - \alpha)\sigma_0 + \alpha\sigma_1 \quad (9)$$

$$\tilde{\mathbf{z}}_{0\alpha} = \sigma_\alpha \left( \frac{\mathbf{z}_{0\alpha} - \mu(\mathbf{z}_{0\alpha})}{\sigma(\mathbf{z}_{0\alpha})} \right) + \mu_\alpha \quad (10)$$

and replace the intermediate latent noise  $\mathbf{z}_{0\alpha}$  with the adjusted one  $\tilde{\mathbf{z}}_{0\alpha}$  in the denoising process. As demonstrated in Fig. 8, the color and brightness are more coherent after AdaIN adjustment.

### 3.6. Reschedule Sampling

With all the methods introduced above, we can generate a smooth transition video between two input images with natural and high-quality in-betweens. However, we observe that using a naive linear sampling schedule for  $\alpha$  may result in an uneven transition rate in image content. To achieve a homogeneous transition rate, we further introduce a new reschedule method.

Formally, assuming  $D(\mathcal{I}_i, \mathcal{I}_j) (i, j \in [0, 1])$  is the perceptual distance between  $\mathcal{I}_i$  and  $\mathcal{I}_j$ , given the number of frames  $n$ , we want the variance of  $\{D(\mathcal{I}_i, \mathcal{I}_{i+\frac{1}{n}})|i = 0, \frac{1}{n}, \dots, 1 - \frac{1}{n}\}$  to be as small as possible. The reschedule sampling starts with approximating the gradient of the relative perceptual distance  $\Delta D$  with respect to  $\alpha$ :

$$\Delta D(\alpha) = \begin{cases} D(\mathcal{I}_0, \mathcal{I}_{\frac{1}{n}})/\bar{D} & \text{if } 0 \leq \alpha < \frac{1}{n}, \\ D(\mathcal{I}_{\frac{1}{n}}, \mathcal{I}_{\frac{2}{n}})/\bar{D} & \text{if } \frac{1}{n} \leq \alpha < \frac{2}{n}, \\ \vdots & \vdots \\ D(\mathcal{I}_{1-\frac{1}{n}}, \mathcal{I}_1)/\bar{D} & \text{if } 1 - \frac{1}{n} \leq \alpha \leq 1, \end{cases} \quad (11)$$

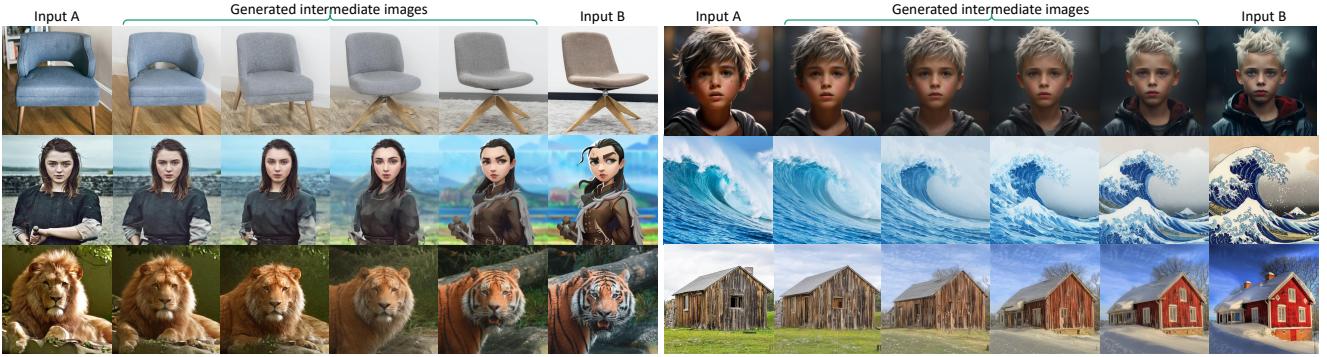


Figure 6. Qualitative results. Our approach achieves visually pleasing image morphing across diverse object categories.

where  $\bar{D} = \sum_{i=0}^{1-\frac{1}{n}} D(\mathcal{I}_i, \mathcal{I}_{i+\frac{1}{n}})$  is the sum of perceptual distance between all adjacent frames. Then the relative perceptual distance to the first frame for every  $\alpha$  can be estimated with

$$D_0(\alpha) = \int_0^{\alpha} \Delta D(x) dx. \quad (12)$$

Finally, by utilizing  $D_0$  and its inversion function  $D'_0$ , we can determine the rescheduled interpolation parameters  $\alpha_i$  as  $\{\alpha_i = D'_0(y) | y = 0, \frac{1}{n}, \dots, 1\}$ . As demonstrated in Fig. 7, the new sampling schedule ensures a more uniform transition rate in the image content.

## 4. Experiment

### 4.1. Implementation Details

In all of our experiments, we use the publicly available state-of-the-art Stable Diffusion v2.1-base as our diffusion model. When training LoRA, to achieve a balance between efficiency and quality and avoid overfitting the single image, we only fine-tune the projection matrices  $Q, K, V$  in the attention modules of the diffusion UNet. Additionally, we set the rank of LoRA to 16, and train for 200 steps using AdamW optimizer [29] with a learning rate of  $2 \times 10^{-4}$ . In this setting, training a LoRA for a  $512 \times 512$  image requires only  $\sim 20$ s on a NVIDIA A100 GPU.

During the inversion and denoising process, we adopt the DDIM schedule of 50 steps distilled from entire diffusion steps  $T = 1000$ . It's noteworthy that we do not apply classifier-free guidance (CFG) [17] in both DDIM inversion and denoising. This is because CFG tends to accumulate numerical errors and cause supersaturation problems, which is also observed in [30, 43]. For attention control, we only perform the feature injection in the upsampling blocks in the self-attention module of the diffusion UNet, and set the hyperparameter  $\lambda$  to 0.6 by default.

### 4.2. MorphBench

Conventional image morphing techniques in computer graphics generally require tedious manual labeling of correspondences, and general image morphing is rarely explored

in depth in the area of generative models. Therefore, there is a lack of specific evaluation benchmarks for this task. To comprehensively evaluate the effectiveness of our methods, we present *MorphBench*, the first benchmark dataset for assessing image morphing of general objects. We collect 90 pairs of pictures of diverse content and styles, and divide them into two categories: i) *metamorphosis* between different objects (66 pairs) and ii) *animation* of the same objects (24 pairs). The latter is obtained using off-the-shelf image editing tools such as DragDiffusion [43], Imagic [26], and MasaCtrl [8]. We hope *MorphBench* can also promote future studies on this important problem.

### 4.3. Qualitative Evaluation

To demonstrate the superiority of our methods, we provide a visual comparison of the results produced by the previous methods and ours. We extensively compare our outcomes with five representative image morphing methods for general objects, including the following three types: i) classic graphical morphing technique [5] based on warping and blending; ii) GAN-based deep interpolation methods DGP [32] and StyleGAN-XL [40] trained on large-scale general image dataset [10]; iii) diffusion-based deep interpolation methods DDIM [46] and Diff.Interp. [51] based on Stable Diffusion v2.1-base [36]. More details about the baselines we use can be found in the supplementary material.

As demonstrated in Fig. 5, our *DiffMorpher* outperforms all previous approaches in terms of image fidelity, semantic consistency, and transformation smoothness, whether used to morph between different objects or animate the same object. We observe that previous approaches suffer from artifacts of different characteristics, while the results of our method are much more visually pleasing. More qualitative results are presented in Fig. 6. With a single diffusion model, our approach well handles diverse object categories and image styles. It is worth mentioning that the generated images accurately reflect the dense correspondence between the two input images, although no such annotation is provided. We recommend readers refer to the supplementary



Figure 7. Ablation study. The four settings are the same as in Table 2: (a) DDIM baseline, (b) + LoRA interpolation, (c) + attention interpolation and replacement, (d) + reschedule (Ours).

Table 1. Quantitative evaluation on *MorphBench*. We report FID ( $\downarrow$ ), PPL ( $\downarrow$ ), and perceptual distance variance (PDV,  $\downarrow$ ) to evaluate the fidelity, smoothness, and speed homogeneity of the transition video respectively.

Method	Metamorphosis			Animation			Overall		
	FID	PPL	PDV	FID	PPL	PDV	FID	PPL	PDV
Warp & Blend	79.63	15.97	4.64	56.86	9.58	0.99	67.57	14.27	3.67
DGP	150.29	29.65	79.40	194.65	27.50	34.21	138.20	29.08	67.35
StyleGAN-XL	122.42	41.94	181.50	133.73	33.43	37.95	112.63	39.67	143.22
DDIM	95.44	27.80	302.83	174.31	18.70	249.16	101.68	25.38	288.51
Diff.Interp.	169.07	108.51	135.95	148.95	96.12	49.27	146.66	105.23	112.84
Ours	70.49	18.19	22.93	43.15	5.14	3.55	54.69	21.10	21.42

material for video results.

#### 4.4. Quantitative Evaluation

To quantitatively evaluate the quality of intermediate images and the smoothness of the transition video, we follow the metrics adopted in the baseline Diff.Interp. [51]:

(1) Frechet inception distance (FID,  $\downarrow$ ) [16]: We compute the FID score between the distribution of the input images and the distribution of the generated images. To estimate the distribution of generated images, we randomly sample two images from the interpolation video 10 times and calculate the mean FID score as an index of the rationality and fidelity of intermediate images.

(2) Perceptual path length (PPL,  $\downarrow$ ) [25]: We compute the sum of the perceptual loss [57] between adjacent images in 17-frame sequences, as an index of the smoothness and consistency of the transition video.

Furthermore, in order to measure the homogeneity of the video transition rate, we introduce a new metric:

(3) Perceptual distance variance (PDV,  $\downarrow$ ): We compute the perceptual loss between consecutive images in 17-frame sequences just like PPL, and then calculate the variance of these distances in the sequence. The average distance variance of all sequences from the test set is taken as the PDV index. This can be a natural measurement of the homogeneity of the video transition rate, where a lower PDV indicates a more uniform speed.

The quantitative results of all approaches are presented in Table 1. Our approach achieves significantly lower FID in both *metamorphosis* and *animation* scenarios, showing better image fidelity and consistency with the input images.

Table 2. Ablation study. We study the effects of each proposed component in our method.

Method	LoRA Interp.	Attention Interp.	AdaIN & Reschedule	FID	PPL	PDV
DDIM	-	✓		101.68	25.38	288.51
	-	✓	✓	44.40	21.81	249.33
Ours	✓	✓	✓	44.90	19.86	157.73
				54.69	21.10	21.42

Although the classic Warp & Blend approach shows better PPL and PDV scores, this is due to the smooth and linear nature of the warping and blending operation which is prone to ghosting artifacts as can be seen in Fig. 5. Among all the deep interpolation methods, our approach has far lower PPL and PDV than others, demonstrating smoother transition video and more homogeneous speed of content change. These results are consistent with the qualitative comparison.

#### 4.5. Ablation Study

To verify the effectiveness of each proposed component, we perform an ablation study and show the results in Table 2 and Fig. 7. The most critical component is LoRA interpolation, which fixes the corrupted images of DDIM to be high-fidelity and semantically smooth images, thus reducing FID, PPL, and PDV. However, abrupt content changes can still be observed, such as the 7th and 8th images of Fig. 7 (b). The attention interpolation and replacement technique effectively eliminates such abrupt changes and makes the image sequence much smoother as shown in Fig. 7 (c), further improving PPL and PDV. Despite so, the speed of content change is still uneven, e.g., the first three images or the last three images of Fig. 7 (c) are almost the same while the content change during 7-9th images is much faster. As shown in Fig. 7 (d), this problem is addressed with our new sampling schedule, which redistributes the content change to be balanced among all consecutive images and thus cuts down PDV by a large margin. Note that this leads to slightly higher FID, because results without rescheduling are biased toward the two ends and thus are closer to the two input images. Lastly, after applying AdaIN adjustment to the latent

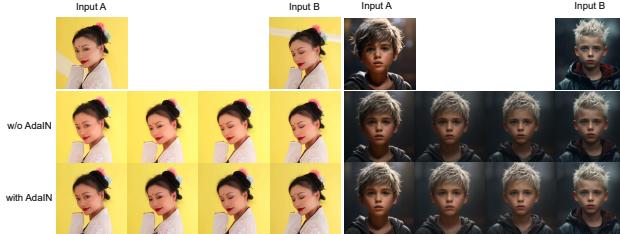


Figure 8. Effects of AdaIN adjustment. The colors and brightness of the intermediate images become more consistent with the input images after AdaIN adjustment.

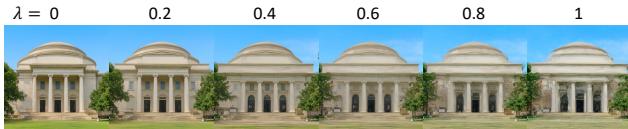


Figure 9. Effects of  $\lambda$ . We show an intermediate image of the second example in Fig. 1 with different  $\lambda$ . The image starts to get blurry when  $\lambda > 0.6$ .

Table 3. Effects of  $\lambda$ .

$\lambda$	0	0.2	0.4	0.6	0.8	1
FID	53.78	52.99	53.45	54.69	52.47	55.84
PPL	23.85	23.25	22.26	21.10	19.49	17.85
PDV	81.15	62.64	36.26	21.42	15.79	12.86

noises, the colors and brightness are more consistent than before, as shown in Fig. 8.

In our method,  $\lambda$  is used to control the strength of attention replacement. We further study its effects in Table 3 and Fig. 9. As  $\lambda$  increases, more attention replacement is involved in the denoising steps, thus improving smoothness and reducing PPL and PDV. However, using interpolated attentions in the later denoising steps can harm the generation of low-level textures and blurry artifacts may emerge, as demonstrated in Fig. 9. We found that setting  $\lambda = 0.6$  achieves a good balance between video smoothness and image quality.

## 5. Conclusion

We have presented *DiffMorpher*, an image morphing approach that only relies on the prior knowledge of a pre-trained text-to-image diffusion model. Our method is able to generate a sequence of visually pleasing images that deliver a smooth transition between two input images. This is achieved by capturing the semantics of the two images via two LoRAs, and interpolating in both the LoRA parameter space and the latent noise to produce a smooth semantic interpolation. An attention interpolation and injection method, an AdaIN adjustment technique, and a new sampling schedule are further introduced to motivate smoothness between consecutive images. We have demonstrated that our approach significantly advances the state of the art in image morphing, uncovering the large potential of diffusion models in this task.

## References

- [1] Alyaa Aloraibi. Image morphing techniques: A review. *Technium: Romanian Journal of Applied Sciences and Technology*, 9:41–53, 2023. [1](#), [2](#), [11](#)
- [2] Hadar Averbuch-Elor, Daniel Cohen-Or, and Johannes Kopf. Smooth image sequences for data-driven morphing. In *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics*, page 203–213, 2016. [2](#)
- [3] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shiliang Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, 2023. [2](#), [3](#)
- [4] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, number 8, page 35–42, 1992. [2](#)
- [5] Bhumika G. Bhatt. Comparative study of triangulation based and feature based image morphing. *Signal & Image Processing : An International Journal*, 2:235–243, 2011. [2](#), [7](#), [11](#)
- [6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. [2](#), [3](#), [11](#)
- [7] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. [2](#), [3](#)
- [8] Ming Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *ArXiv*, abs/2304.08465, 2023. [3](#), [5](#), [7](#)
- [9] Soheil Darabi, Eli Shechtman, Connally Barnes, Dan B. Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM TOG*, 31(4), 2012. [2](#)
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. [7](#)
- [11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794, 2021. [3](#)
- [12] Noa Fish, Richard Zhang, Lilach Perry, Daniel Cohen-Or, Eli Shechtman, and Connally Barnes. Image morphing with perceptual constraints and stn alignment. *Computer Graphics Forum*, 39, 2020. [2](#)
- [13] 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. [2](#), [3](#)
- [14] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [4](#)
- [15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. [3](#), [5](#)
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. [8](#)
- [17] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*,

- abs/2207.12598, 2022. 7
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3, 5
- [19] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47:1–47:33, 2021. 3
- [20] 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*, abs/2106.09685, 2021. 2, 4
- [21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 6
- [22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 2, 3
- [23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4396–4405, 2018.
- [24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. pages 8107–8116, 2019. 2, 3
- [25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2019. 8
- [26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-Tang Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2022. 2, 3, 7
- [27] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. 4
- [28] Jing Liao, Rodolfo S. Lima, Diego Nehab, Hugues Hoppe, Pedro V. Sander, and Jinhui Yu. Automating image morphing using structural similarity on a halfway domain. *ACM TOG*, 33(5), 2014. 2
- [29] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. 7
- [30] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2022. 5, 7
- [31] Chong Mou, Xintao Wang, Jie Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *ArXiv*, abs/2307.02421, 2023. 2, 3
- [32] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 2020. 2, 3, 7, 11
- [33] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 3, 5
- [34] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongs, and Supasorn Suwajanakorn. Diffusion autoen-
- coders: Toward a meaningful and decodable representation. In *CVPR*, 2022. 3
- [35] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 2021. 3, 11
- [36] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2021. 1, 2, 3, 7
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 3
- [38] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 3
- [39] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. In *International Conference on Machine Learning*, 2023. 2, 3, 11
- [40] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. 2022. 2, 3, 7, 11
- [41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 3
- [42] Eli Shechtman, Alex Rav-Acha, Michal Irani, and Steve Seitz. Regenerative morphing. In *CVPR*, pages 615–622, 2010. 2
- [43] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *ArXiv*, abs/2306.14435, 2023. 3, 5, 7
- [44] Ken Shoemake. Animating rotation with quaternion curves. *SIGGRAPH Comput. Graph.*, 19(3):245–254, jul 1985. 5
- [45] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, page 2256–2265, 2015. 2, 3
- [46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 2, 5, 7, 11
- [47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 2, 3
- [48] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jegou. Training data-efficient image transformers & distillation through at-

- tention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 11
- [49] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, June 2023. 3, 5
- [50] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4
- [51] Clinton J. Wang and Polina Golland. Interpolating between images with diffusion models, 2023. 3, 7, 8, 11
- [52] Xintao Wang, K. Yu, Chao Dong, Xiaoou Tang, and Chen Change Loy. Deep network interpolation for continuous imagery effect transition. In *CVPR*, pages 1692–1701, 2018. 5
- [53] George Wolberg. Image morphing: a survey. *The Visual Computer*, 14:360–372, 1998. 1, 2, 11
- [54] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE TPAMI*, 45:3121–3138, 2021. 2, 3, 11
- [55] Zhaoyuan Yang, Zhengyang Yu, Zhiwei Xu, Jaskirat Singh, Jing Zhang, Dylan Campbell, Peter Tu, and Richard Hartley. Impus: Image morphing with perceptually-uniform sampling using diffusion models. *ArXiv*, abs/2311.06792, 2023. 3
- [56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 11
- [57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8
- [58] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *CVPR*, pages 6027–6037, 2023. 2
- [59] Bhushan Zope and Soniya B. Zope. A survey of morphing techniques. *International Journal of Advanced engineering, Management and Science*, 3:81–87, 2017. 1, 2, 11

## A. More Details of Baselines

In Sec. 4, we comprehensively compare our method with previous state-of-the-art methods, including graphical, GAN-based and diffusion-based techniques. We offer more details of the baselines that we use here:

- Warp & Blend [1, 53, 59]: Conventional graphical techniques usually involve bidirectional image warping based on correspondence point pairs with blending operations to achieve morphing effects. We select the representative triangulation-based method [5] as our baseline, which is also widely used in standard libraries such as OpenCV. It divides the images into triangles by performing Delaunay triangulation on user-defined corresponding points, and then morphs between the triangle pairs. Thus, the quantity and quality of the manually labeled pair of points greatly affect the generated results. Since all the other methods do not require correspondence annotations, for the sake of fairness, we adopt the automatic version of this approach <https://github.com/jankovicsandras/autonomousvision/styleganmorph> that selects 50 morph-points automatically using OpenCV.
- Deep Generative Prior (DGP) [32]: DGP is an image manipulation method based on BigGAN [6], which is suitable for general image morphing. We adopt the official code <https://github.com/XingangPan/deep-generative-prior> with its default hyperparameters and the pretrained BigGAN model trained on ImageNet [6] as our baseline.
- StyleGAN-XL [40]: Since the pretrained checkpoint of StyleGAN-T [39] is not publicly available, we use the alternative state-of-the-art GAN model StyleGAN-XL <https://github.com/autonomousvision/stylegan-xl> as our another baseline. Similarly to DGP, the model is trained on ImageNet. We obtain the latent codes of input images by GAN inversion [54] and tune the generator by PTI [35] for better reconstruction results, and interpolate both the latent codes and the generator parameters to get intermediate images. For both GAN-based methods, we use the ImageNet classifier DeiT [48] to automatically determine the class label.
- Denoising Diffusion Implicit Model (DDIM) [46]: We implement a naive diffusion-based interpolation method through DDIM inversion and latent interpolation as our baseline, as discussed in the DDIM paper. As with our approach, the underlying model used is also Stable Diffusion v2.1-base <https://huggingface.co/stabilityai/stable-diffusion-2-1-base>.
- Diff.Interp. [51]: *Interpolating between Images with Diffusion Models* is a recent state-of-the-art image interpolation method based on diffusion models. Besides latent interpolation, it further introduced pose guidance based on ControlNet [56] to encourage more reasonable intermediate results. However, the smoothness of the morph-

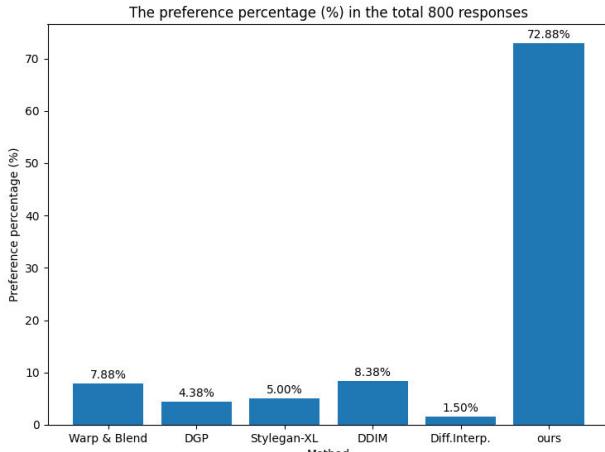


Figure 10. User study result. Our method surpasses all the previous methods by a large margin in terms of user preference.

ing video was not considered in this work, and the generated video is full of flickering artifacts. We employ the official code <https://github.com/clintonjwang/ControlNet> with default settings and pretrained Stable Diffusion v2.1-base model as our baseline. For all three diffusion-based methods, the prompts for each test case are shared.

## B. User Study

To assess the quality of image morphing from a human perspective, we invite 40 volunteers to conduct a user study. Each participant are shown 20 groups of morphing videos created by our approach and five baseline methods, chosen at random. They are asked to evaluate the image morphing quality from the perspective of intermediate image fidelity and video smoothness, and to select the one with the best quality for each question. An example of the questionnaire is shown in Fig. 15. In total, we collect 800 responses and summarize the results in Fig. 10. As we can see, our approach is significantly more preferred by users than any of the prior methods.

## C. Limitations

One of the limitations of our approach is that we have to train a LoRA for each input image before morphing, which costs additional time ( $\sim 20$  s on a single NVIDIA A100 GPU for a  $512 \times 512$  image). Another limitation of text-guided diffusion models is that the user must input aligned text prompts in addition to images. Besides, our approach occasionally fails in difficult cases where the correspondence between two input images is not clear enough, and produces relatively unreasonable intermediate images, as shown in Fig. 11.



Figure 11. Some relatively unsuccessful cases where the correspondence between two images is not clear enough.

## D. More Qualitative Results

Here we present more qualitative results to demonstrate the effectiveness of our *DiffMorpher*. Fig. 12 gives more examples to illustrate the superiority of our approach compared to previous methods in diverse scenarios, and Fig. 13 and Fig. 14 provide additional qualitative results generated by our method that further demonstrate its versatility in real-world applications.

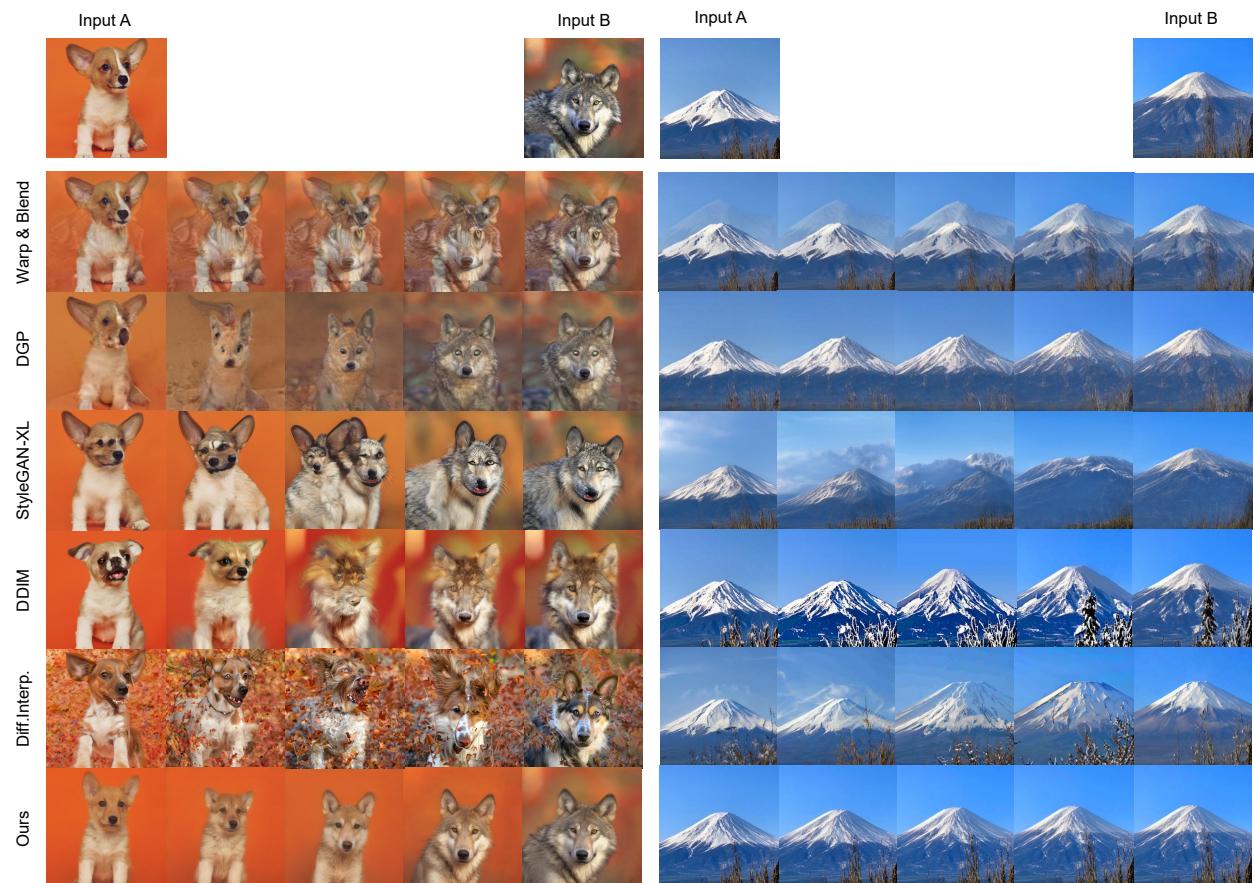


Figure 12. More qualitative comparison results.

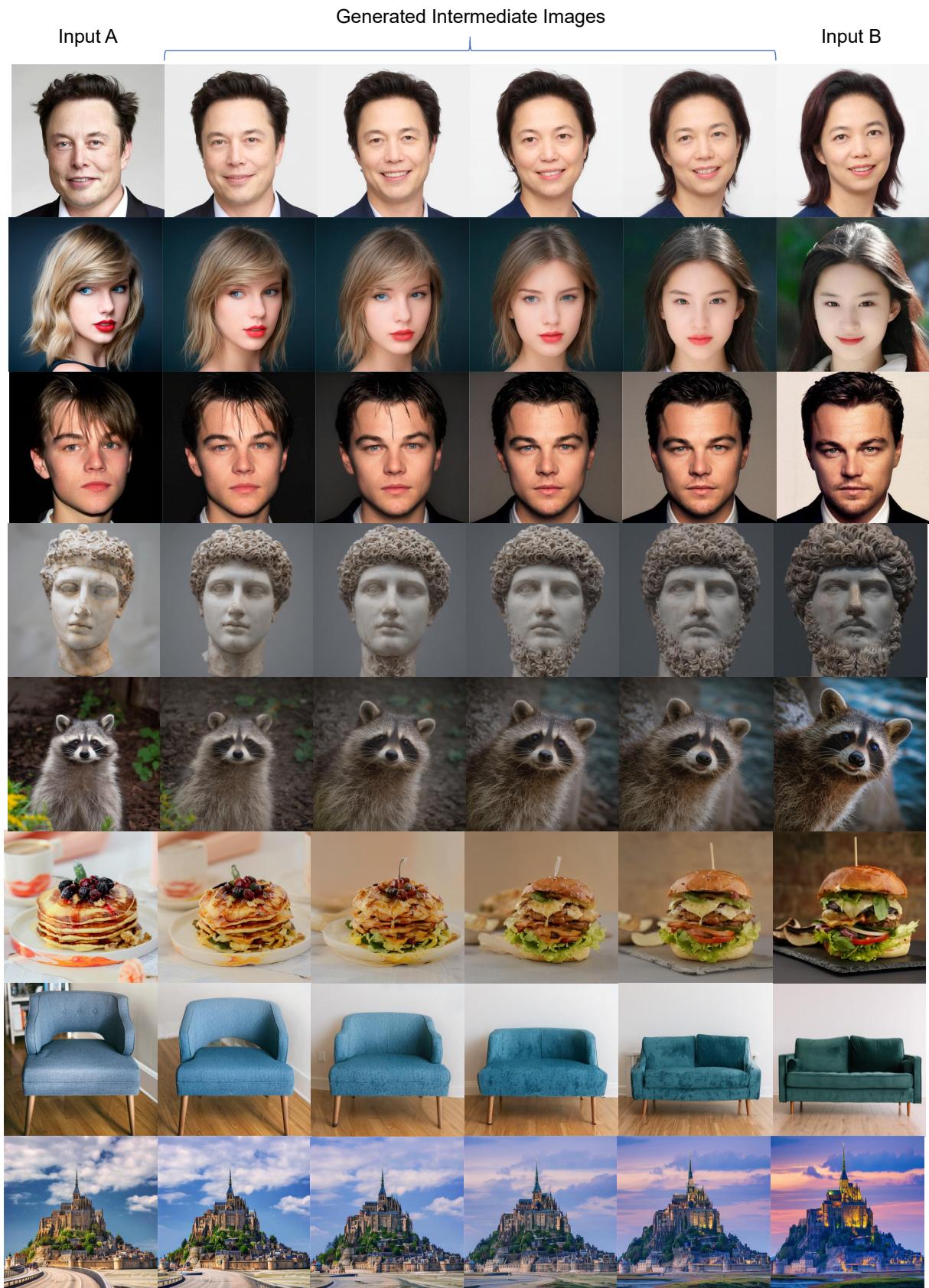


Figure 13. More qualitative results of our approach.



Figure 14. More qualitative results of our approach.

cat\_rabbit \*

Please select the one with the best image morphing quality from the perspective of intermediate image fidelity and video smoothness.



Input A



Input B



1



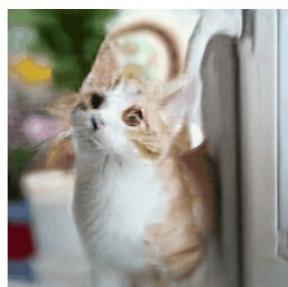
3



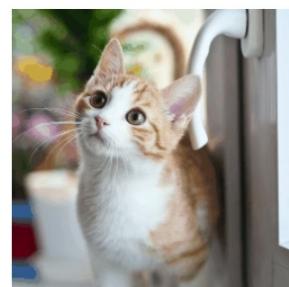
5



2



4



6

- Result 1
- Result 2
- Result 3
- Result 4
- Result 5
- Result 6

Figure 15. An example of the questionnaire we used in the user study. Note that all the results shown here are videos.