DreamMover: Leveraging the Prior of Diffusion Models for Image Interpolation with Large Motion

Liao Shen[1], Tianqi Liu[1], Huiqiang Sun[1], Xinyi Ye[1], Baopu Li[1], Jianming Zhang[2], and Zhiguo Cao[1]*

[1] School of AIA, Huazhong University of Science and Technology
{leoshen,tq_liu,shq1031,xinyiye,zgcao}@hust.edu.cn
bpli.cuhk@gmail.com
[2] Adobe Research
jianmzha@adobe.com

Abstract. We study the problem of generating intermediate images from image pairs with large motion while maintaining semantic consistency. Due to the large motion, the intermediate semantic information may be absent in input images. Existing methods either limit to small motion or focus on topologically similar objects, leading to artifacts and inconsistency in the interpolation results. To overcome this challenge, we delve into pre-trained image diffusion models for their capabilities in semantic cognition and representations, ensuring consistent expression of the absent intermediate semantic representations with the input. To this end, we propose **DreamMover**, a novel image interpolation framework with three main components: 1) A natural flow estimator based on the diffusion model that can implicitly reason about the semantic correspondence between two images. 2) To avoid the loss of detailed information during fusion, our key insight is to fuse information in two parts, high-level space and low-level space. 3) To enhance the consistency between the generated images and input, we propose the self-attention concatenation and replacement approach. Lastly, we present a challenging benchmark dataset called *InterpBench* to evaluate the semantic consistency of generated results. Extensive experiments demonstrate the effectiveness of our method. Our project is available at https://dreamm0ver.github.io.

Keywords: Diffusion models · Image interpolation · Image editing · Short-video generation · Semantic consistency

1 Introduction

With the widespread popularity of short videos on the internet and mobile phone apps such as TikTok and YouTube shorts, people enjoy so much watching short videos. The desire for a more engaging visual experience has led to the exploration of innovative technologies in computer vision and graphics, one of which is

* Corresponding author

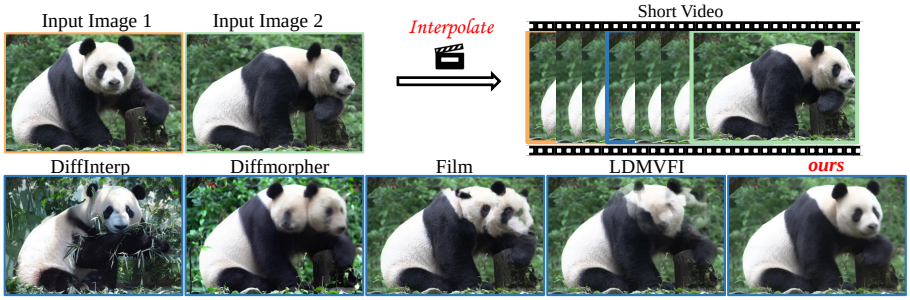

Fig. 1: Given two input images with large motion, our proposed method can generate a short video with high fidelity and semantic consistency compared to previous approaches. To see the dynamic effect of our method, we encourage readers to watch our supplementary video.

image interpolation. Image interpolation refers to the process of generating intermediate images from two given images, and it has been a typical and challenging task for many years, especially when these two images show large motions. Two images with large motion captured at different times in one scene often exhibit great variation, and image interpolation aims to recover the potential dynamic processes, providing viewers with lively and dynamic animations. With the input image pair serving as the starting and ending images, such a process generally produces a consistent sequence of object motion videos with rather high fidelity.

Several existing methods can synthesize intermediate frames from two given images, such as video frame interpolation and image morphing. However, video frame interpolation [18,49,54] is primarily designed to increase video frame rates, which is significantly different from our purpose of generating short videos. Due to the small differences between adjacent frames, these algorithms often neglect the semantic consistency between input video frames and synthesized intermediate frames. LDMVFI [4] struggles in large motion and lacks the ability of semantic cognitive. Film [28] attempts to interpolate frames between two images with relatively large motion. However, it also operates within near-duplicates and does not model the semantic consistency of intermediate frames. On the other hand, image morphing methods [43, 46, 51] can also produce intermediate images from given pairs. However, these models usually focus on the transition between topologically similar objects. In contrast, image interpolation mainly aims to construct semantic consistency for the intermediate and input images, generating realistically consistent videos of object movements from two images. The lack of semantic cognitive in the aforementioned methods results in a tendency to split the complete object during interpolation. When applied to such settings, they often result in severe semantic errors and artifacts, leading to inaccuracies in generating intermediate images (as illustrated in Fig. 1 with the erroneous expression of the panda head).

The rise of diffusion models [10, 38, 39] has made a profound impact on the field of image generation and image editing. Thanks to the powerful architecture

and large aligned image-text datasets [32], the pre-trained generative diffusion models contain rich implicit semantic information. When there is large motion between image pairs, intermediate semantic information may not be present in either of the input images. In order to guarantee a coherence transition from one image to the other, we attempt to leverage the pre-trained diffusion model to express the semantic information of input image pairs, and generate intermediate images with high semantic consistency.

To this end, we propose DreamMover, a novel image interpolation algorithm based on a text-to-image diffusion model, which enables generating large motion videos with semantic consistency from two images. To ensure semantic consistency between the generated and input images, we suggest a new scheme that consists of flow estimation and image fusion. Specifically, we extract feature maps of the input image pair from the up-blocks of U-Net [30] during the noise-adding process. These features are then used to establish pixel correspondences between two images by calculating the cosine distance, further yielding bidirectional optical flow maps. Based on this, we fuse the image pair using softmax splatting [24] and time-weighted interpolation in latent space to generate intermediate images.

For image fusion, we observed that directly using weighted average operations in latent space may result in a significant loss of high-frequency information, which is not beneficial to modeling semantic consistency. To address this issue, we divide the noisy latent code into two components: a high-level part for overall spatial layout information and a low-level part representing high-frequency details. For the high-level part, we maintain the fusion method using softmax splatting and time-weighted interpolation. For the low-level part, we employ the Winner-Takes-All (WTA) method for fusion. This approach preserves the correct semantic overall layout in the generated video while effectively retaining high-frequency detail information. During the denoising stage, to further ensure semantic consistency, we concatenate the key and value of the input image pairs and replace those of the intermediate ones. Also, we perform low-rank adaptations (LoRAs) [12] to enhance consistency by fine-tuning the diffusion model.

To the best of our knowledge, we are the first image interpolation method considering semantic consistency, which has a vital impact on video effect. Due to the lack of suitable datasets for image interpolation, we curate a dataset, *InterpBench*, to evaluate the performance of generated videos from image interpolation algorithms. Extensive experiments demonstrate that our approach significantly outperforms the state-of-the-art video frame interpolation and image morphing methods. We also conduct a user study to demonstrate the superiority of our method in the view of humans.

In summary, we propose a novel image interpolation framework that can generate semantic consistent intermediate images from image pairs with large motion, which has the following contributions: 1) a natural optical flow estimator for large motion, 2) a two-level fusion strategy to minimize the loss of high-frequency information, 3) a self-attention concatenation and replacement method to enhance semantic consistency.

2 Related work

Image Interpolation. Previous methods such as video frame interpolation and image morphing can synthesize interpolation images from two given images. Video frame interpolation [6, 15, 17, 26, 36, 49] are commonly used for up-scale frame rates of videos, which mainly exhibit small motion between consecutive video frames. These methods often lack semantic-level cognitive capabilities and are challenging for large motion. LDMVFI [4] trains a diffusion model for video frame interpolation from scratch, but artifacts tend to occur when there is large motion between images. Film [28] attempts to capture relatively large motion in near-duplicate images. However, when the motion is even larger, artifacts and fragmentation often appear. In contrast, our method leverages the prior in pre-trained text-to-image diffusion models and generates reasonable and high-fidelity interpolated images. DiffInterp [43] tries to interpolate images through latent code interpolations and text embedding interpolations. Further, Diffmorpher [51] applies low-rank adaptations (LORA) [12] to two images separately and interpolates between the LoRA parameters for semantic transition. However, they mainly focus on two images of topologically similar objects, but may not work well in the same objects with large motion. Unlike them, we use optical flow to fuse information between two images instead of simply overlaying it.
Controllable Image Editing Controllable image editing based on diffusion model is a challenging task that aims to manipulate and generate novel images according to various conditions, including text-based editing [5, 8, 42], image-based editing [20, 31, 33, 52], point-based editing [23, 25, 35] and motion-based editing [7, 34]. These methods mainly add noise to the clean image using DDIM inversion [38] and denoise by the guidance of various conditions. In this way, diffusion models can generate high-quality new images that fit well with the semantics of the origin image. Most of these works edit a single image and generate semantically consistent edited images, while the generation of intermediate results from two images is much less explored in image diffusion models.
Image-to-Video Diffusion Models Previous works on Image-to-Video Diffusion Models [37,44,47,48] have achieved great success, which contain downstream tasks that can be used for frame interpolation between two images. We differ significantly from these methods in that we edit images to generate intermediate image sequences via the prior pre-trained image diffusion models, but they directly utilize video diffusion models which require more complex architectures and training on large-scale video datasets.

3 Method

Given a pair of images $\mathcal{I}^0$ and $\mathcal{I}^1$ with large motion, we aim to generate intermediate images $\mathcal{I}^\delta$ and yield a semantically consistent video $\mathcal{V} = \{\mathcal{I}^\delta | \delta \in (0, 1)\}$, where the sequence length of time δ depends on the desired number of interpolation images n.

We schematically illustrate our pipeline in Fig. 2. Our method starts by obtaining bidirectional optical flow from correspondence between the feature

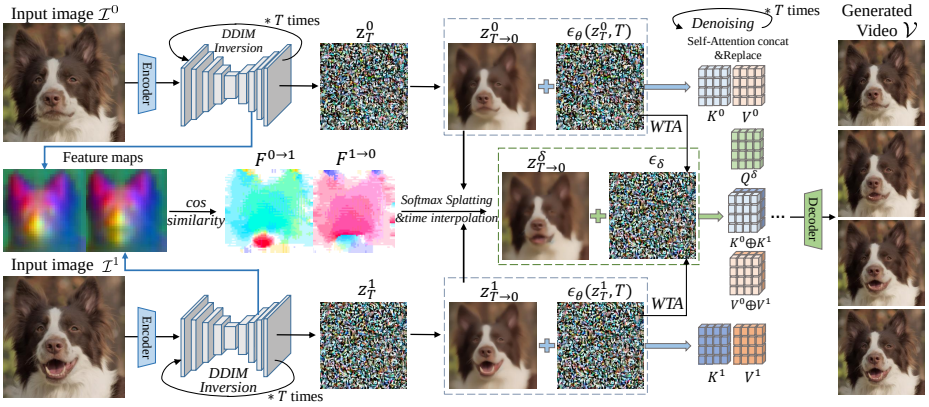

Fig. 2: Overview of our method. Given two input images $\mathcal{I}^0$ and $\mathcal{I}^1$, we extract feature maps and leverage them to obtain the bidirectional optical flow $F^{0\to 1}$ and $F^{1\to 0}$. Next, we decompose the noisy latent code z_T into two-level space and perform softmax splatting and time interpolation for image fusion. For high-frequency information ϵ_θ, we replace all weighted average operations with "Winner-Takes-All"(WTA). In addition, we propose a novel self-attention replacement method for consistency. Finally, our method can generate a sequence of high-fidelity interpolation frames.

maps (Sec. 3.2). In order to preserve the details of interpolation images carefully during fusion, we divide the origin latent space into two parts, high-level and low-level space, and operate on each part individually (Sec. 3.3). Finally, to enhance the appearance consistency between the two input images, we propose the self-attention concatenation and replacement during denoising, and perform LoRA for semantic-preserving (Sec. 3.4).

3.1 Premininaries

Latent diffusion model (LDM) [29] stands out as an efficient variant of diffusion models, employing the diffusion process within the latent space. This involves the implementation of both a forward and a backward process. For a given clean latent input z_0, the forward diffusion process gradually adds Gaussian noise at each timestamp t to obtain z_t:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I),$$

where $\{\beta_t\}_{t=1}^T$ represent the scale of noises, and T denotes the number of diffusion timestamps. Then the backward denoising process utilizes a trained U-Net ϵ_θ for denoising:

$$p_\theta(z_{t-1}|z_t) = \mathcal{N}(z_{t-1}; \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)),$$

where μ_θ and Σ_θ are computed by ϵ_θ.

To accurately reconstruct given real images, we employ the deterministic DDIM inversion and sampling [38] to add noise and remove noise. We can simplify the denoising process into the following form to predict the z_{t-1} of previous

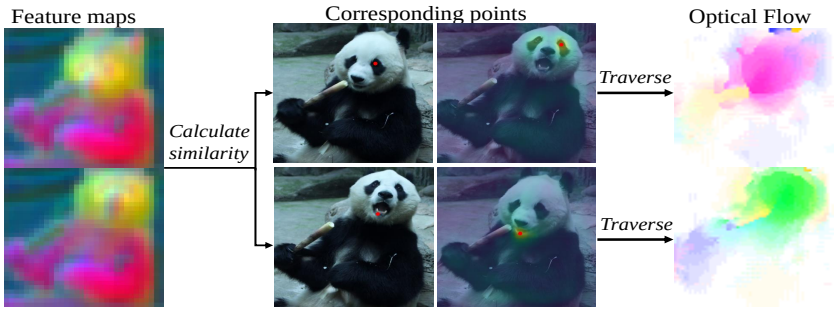

Fig. 3: The potential of diffusion model for optical flow estimation. We perform PCA on the features and observe consistent spatial layouts with input images, and obtain bidirectional optical flow through the correspondence between feature maps.

timestamp:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \cdot z_{t\to 0} + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(z_t, t), \tag{2}$$

$$z_{t\to 0} = \frac{z_t - \sqrt{1-\alpha_t}\epsilon_\theta(z_t, t)}{\sqrt{\alpha_t}}. \tag{3}$$

where t denotes the noisy time, $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$ and $z_{t\to 0}$ means the predicted clean latent code that is directly denoised from z_t.

3.2 Diffusion-aware flow estimation

Given two images $\mathcal{I}^0$ and $\mathcal{I}^1$, optical flow estimation is a key step in image interpolation, which indicates the correspondences of pixels between two images and can be employed to warp pixels to generate the intermediate results. We can warp an image with an optical flow F by softmax splatting method [24]:

$$\overrightarrow{\sigma}(\mathcal{I}, F) = \frac{\sum(exp(M)\cdot\mathcal{I}, F)}{\sum(exp(M), F)}, \tag{4}$$

where M is a metric of brightness constancy [1].

Specifically, we encode $\mathcal{I}_0$ and $\mathcal{I}_1$ into the latent space to get z^0 and z^1. By getting a bidirectional optical flow $F^{0\to 1}$ and $F^{1\to 0}$ from the two images, we can warp z^0 and z^1 to the middle time $\delta \in (0,1)$ using softmax splatting, and get the middle latent code $z^{0\to\delta}$ and $z^{1\to\delta}$ respectively. The final intermediate latent code z^δ can be obtained by fusing them with time-weighted interpolation.

The crux of the matter lies in getting bidirectional optical flow between two images without introducing additional optical flow prediction modules. Drawing inspiration from [19, 40, 50], diffusion model engages in implicit reasoning about image correspondences, yielding remarkably robust and accurate results. Therefore, we use the pre-trained diffusion model to obtain optical flow through semantic correspondence between real images, without the need for additional fine-tuning or supervision. Specifically, we employ the DDIM inversion [38] to

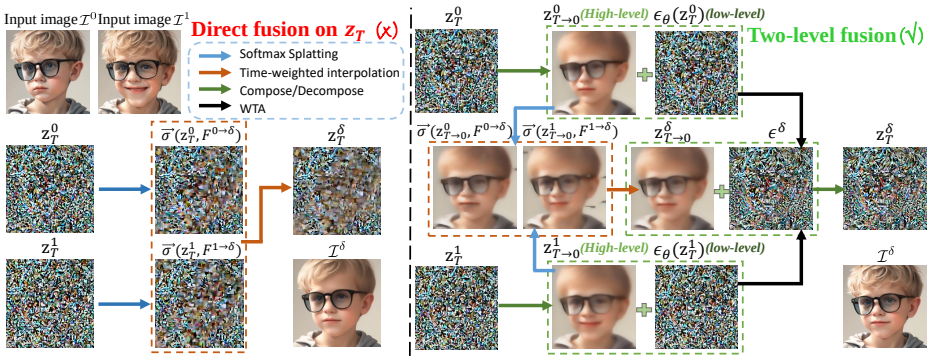

Fig. 4: The process of direct fusion and our proposed two-level fusion. Generally, $z_{T\to 0}$ represents a latent code. Here, for clearer visualization, we illustrate the RGB image decoded from it to emphasize a significant loss of high-frequency information compared to input images.

send z_0^0 and z_0^1 into the U-Net for adding noise, where z_t^δ represents the latent code of intermediate time δ after t steps of noise-adding. Meanwhile, feature maps f^0 and f^1 are extracted from up-blocks of U-Net. As illustrated in Fig. 3, the spatial layout between the feature maps is highly similar to that of the original images, which provides us with the possibility for optical flow prediction of two images in latent space. By traversing the points in one feature map, we can select the pixel in the other map with the highest cosine similarity as its corresponding location, thereby obtaining the bidirectional optical flow maps:

$$F^{0\to 1}(x,y) = \arg\max_{i,j} \langle f^0(x,y), f^1(i,j) \rangle, \tag{5}$$

where (x,y) and (i,j) are the indexes of I^0 and I^1, and $\langle\ ,\ \rangle$ denotes the calculating of cosine similarity. Likewise, we can derive the optical flow $F^{1\to 0}$ from I^1 to I^0. Lastly, we can gain the optical flow $F^{0\to\delta} = \delta \cdot F^{0\to 1}$ from time 0 to the intermediate time δ and the flow $F^{1\to\delta} = (1-\delta)\cdot F^{1\to 0}$ from time 1 to the intermediate time δ.

3.3 Latent space fusion

In order to combine the effective semantic information of the input image pairs, we apply softmax splatting [24] to respectively warp these two images to middle time δ based on the bi-directional flow, and acquire the middle results $z^{0\to\delta} = \overrightarrow{\sigma}(z^0, F^{0\to\delta})$ and $z^{1\to\delta} = \overrightarrow{\sigma}(z^1, F^{1\to\delta})$. Further, we can fuse them together using time-weighted interpolation. However, due to the large object motion between two input images, the intermediate semantic information may be absent in input images. Hence, we leverage the potential semantic capability of image diffusion to yield reasonable results and maintain semantic consistency.

Direct fusion. We first add noise through DDIM inversion to z_0^0 and z_0^1 in order to obtain z_T^0 and z_T^1, where T is the total number of noise addition steps.

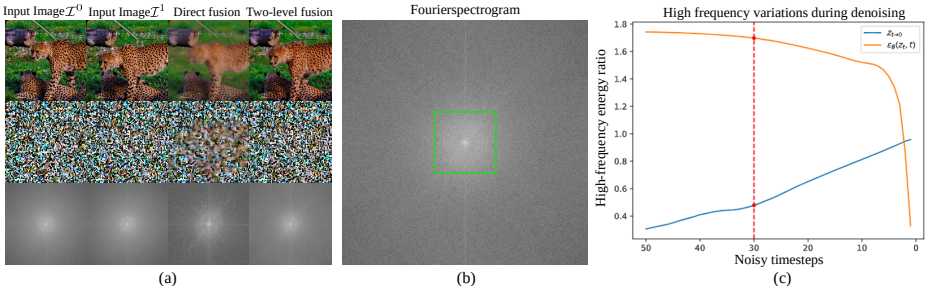

Fig. 5: (a) Effects of fusion in different space. Compared to direct fusion, our strategy better preserves details in the RGB image and maintains more high-frequency energy in the Fourier spectrograms. **(b) Definition of high-frequency region.** We define it as the part of the spectrogram beyond the centre 1/4. **(c) High-frequency variations during denoising.**

Then, we simply warp them to the middle and combine them according to time-weighted interpolation to obtain the noisy latent code for the middle time δ:

$$z_T^\delta = (1-\delta)\cdot\overrightarrow{\sigma}(z_T^0, F^{0\to\delta}) + \delta\cdot\overrightarrow{\sigma}(z_T^1, F^{1\to\delta}). \tag{6}$$

Finally, we transmit z_T^δ to the diffusion model for denoising and desire to directly generate a reasonable and fidelity result. Nevertheless, as demonstrated in Fig. 4, the results of direct fusion exhibit noticeable blurriness. Intuitively, both softmax splatting and interpolation will introduce average operations, leading to the loss of high-frequency information. We provide mathematical explanations and conduct a thorough analysis in supplementary material.

Two-level fusion. Revisiting the denoising process of diffusion as shown in Eq. 2, it has two components $z_{t\to0}$ and $\epsilon_\theta(z_t, t)$. For $z_{t\to0}$, since it is a predicted clean latent by a one-step denoising rather than a multi-step progressive denoising, it can only capture certain high-level context information while lacking high-frequency details. On the other hand, the component $\epsilon_\theta(z_t, t)$ serves to complement low-level textures during denoising. we show quantitatively that $\epsilon_\theta(z_t, t)$ has more high-frequency components than $z_{t\to0}$ and that our strategy retains more high-frequency information in Fig. 5(c).We define high-frequency energy by the sum of the amplitudes in the high-frequency region of the Fourier spectrogram. For the high-frequency region, we define it as the part of the spectrogram beyond the centre 1/4. Specifically, the portion outside the green square in Fig. 5(b). We take the ratio of their respective high-frequency energies to that of input image as a metric(y-axis). We perform fusion operations at the 30th noisy step, and $\epsilon_\theta(z_t, t)$ has more high-frequency information compared to $z_{t\to0}$.

To integrate the effective information of two images while preserving high-frequency details, we propose a two-level fusion strategy. Specifically, for high-level information, we perform fusion in the $z_{T\to0}$ space. Per Eq. 3, $z^0_{T\to0}$ and $z^1_{T\to0}$ can be obtained from z^0_T and z^1_T, and the fused result is:

$$z^\delta_{T\to0} = (1-\delta)\cdot\overrightarrow{\sigma}(z^0_{T\to0}, F^{0\to\delta}) + \delta\cdot\overrightarrow{\sigma}(z^1_{T\to0}, F^{1\to\delta}). \tag{7}$$

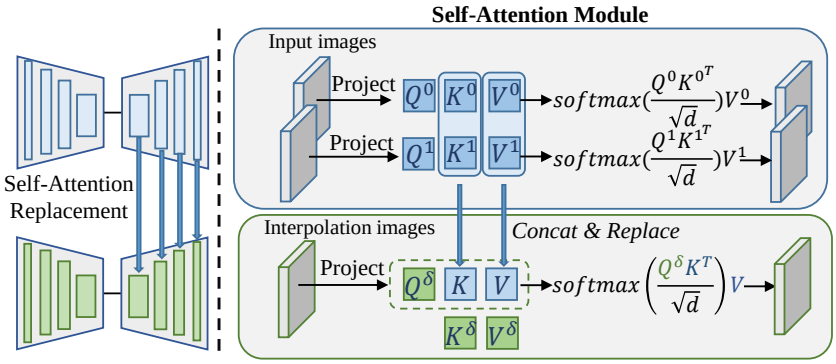

Fig. 6: Self-attention concatenation and replacement.

For low-level information, we perform fusion on ϵ_θ space. To mitigate the loss of high-frequency information caused by average operation in softmax splatting and time interpolation, we apply the "Winner-Takes-All"(WTA) operation, i.e, taking values of the highest weights, to obtain the fused result $\epsilon^\delta = WTA(\epsilon_\theta(z_T^0), \epsilon_\theta(z_T^1))$. After obtaining the separately fused results for two levels, we backtrack to obtain z_T^δ,

$$z_T^\delta = \sqrt{\alpha_T} \cdot z_{T\to 0}^\delta + \sqrt{1-\alpha_T} \cdot \epsilon^\delta. \tag{8}$$

Finally, based on Eq. 2, we can yield a clean latent z_0^δ by performing T times of denoising, and then send it to the decoder of diffusion to get the intermediate image $\mathcal{I}^\delta$. As shown in Fig. 5(a), the quality of images obtained through our proposed two-level fusion is significantly superior to the one of direct fusion.

3.4 Reference-guided consistency

Although the intermediate results are reasonable regarding spatial layout, we observe inconsistent changes in the generated images. We posit that this issue arises due to the absence of adequate guidance from the original input images during the denoising process. To solve this problem, we draw inspiration from attention control techniques in previous image editing research [2,3,13,14,27,42] and propose a novel self-attention concatenation and replacement method, which introduces the attention features of the input image pair during denosing into the denoising process of the intermediate image. Specifically, we can use the query features in the self-attention module of interpolation images to query the corresponding key and value features in input image pairs.

As shown in Fig. 6, in the denoising steps, we feed the noisy latent code of the input two images into the U-Net to obtain the key and value matrices $K^i, V^i (i = 0, 1)$ in the self-attention modules of U-Net up-sampling blocks. In order to generate a reliable intermediate image I_δ, we replace its key and value by concatenating K^i and V^i:

$$Q = Q^\delta, \quad K = (K^0 \oplus K^1), \quad V = (V^0 \oplus V^1); \tag{9}$$

$$Attention^{\delta} = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{10}$$

where $\oplus$ denotes the concatenation operation. Thus, intermediate latent code can query correlated local structures and textures from both input images to further enhance consistency.

In addition, we conduct Low-Rank Adaption (LoRA) [12] to further improve the semantic consistency of the intermediate images with input images. Unlike Diffmorpher [51], which requires adapting LoRAs to the input two images respectively, our method simply fits a single LoRA for the image pair. Finally, the fine-tuned model can generate samples with consistent semantic identity.

4 Experiments

4.1 Implementation Details

In all of our experiments, we adopt the Stable Diffusion 1.5 [29] as our diffusion model and the number of interpolation images is 32. During the latent optimization stage, we schedule 50 steps for DDIM and optimize the diffusion latent at the 30th noisy step unless specified otherwise, and we extract the output of the second up-block of the UNet at the 14th noisy step as feature maps used for flow estimation. In addition, we set the rank of LoRA to 16. We fine-tune the LoRA using the AdamW optimizer with a learning rate of 5×10^{-4} for 80 steps, and it takes ~ 40 seconds on a single NVIDIA RTX 3090 GPU. It is noteworthy that in both DDIM inversion and denoising, we do not apply classifier-free guidance (CFG) [11]. This is because CFG tends to accumulate numerical errors and cause supersaturation problems [21].

4.2 Baselines and Evaluation metrics

To evaluate the effectiveness of our method, we extensively compare our outcomes with two image morphing techniques and two video interpolation methods. Diffinterp [43] and Diffmorpher [51] are diffusion model-based methods that can generate a sequence of intermediary images for two given images of topologically similar objects. Film [28] trained on multi-scale video interpolation datasets attempts to handle frame interpolation for large motion. LDMVFI [4] trains a latent diffusion model for video interpolation, since this method just obtains one intermediate image, we iterate to generate a sequence of interpolated images.

To quantitatively evaluate the quality of interpolation images and the generated videos, we adopt Fréchet Inception Distance (FID) [9], Perceptual Similarity (LPIPS) [53], Warping Error (WE) [16], and WE_{mid} as our metrics. We use FID and LPIPS to evaluate the fidelity and rationality of all methods, and utilize WE to evaluate the temporal coherency of the generated videos. In addition, we employ WE_{mid} to measure whether the middle-most image is consistent with the input image pair.

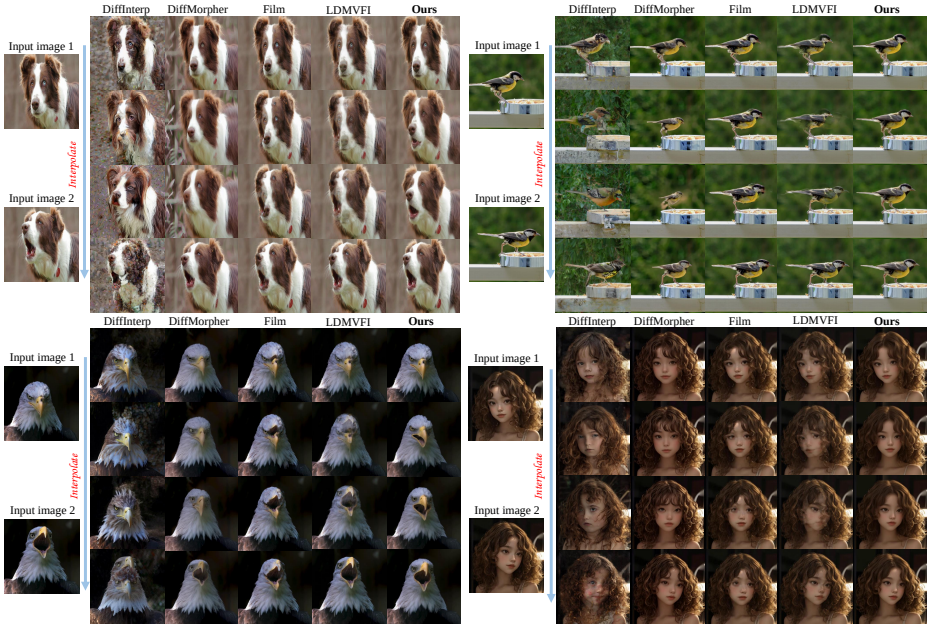

Fig. 7: Qualitative comparisons of baselines and our method on *InterpBench*. For each scenario, from left to right we show four methods: Diffinterp, DiffMorpher, Film, and ours, and from top to bottom we show four images interpolated from each method.

4.3 InterpBench

Due to the lack of discussion on semantic consistency modeling for image interpolation, there are currently no datasets suitable for our task. Existing video frame interpolation datasets [1, 17, 22, 45] provide triplets of input image pairs and intermediate images, but they are designed for scenarios where the motion between two input images is minimal. As such, discussing the semantic consistency of the intermediate image may be meaningless and not applicable to evaluating the performance of image interpolation algorithms. To meet the demand for performance evaluation of image interpolation algorithms, we introduce *InterpBench*, the first benchmark dataset tailored for image interpolation. *InterpBench* is a diverse compilation encompassing various large motions of objects and we collected 100 pairs of pictures in total. Details of our dataset can be found in the supplementary materials.

4.4 Qualitative Evaluation

Visual qualitative comparisons are shown in Fig. 7. Our method outperforms all other baselines in terms of image fidelity, image detail, and semantic consistency

Fig. 8: More Visualization Comparison of baselines and our method. We show the middle-most image obtained by all methods. Our approach generates intermediate results that maintain the best semantic consistency.

of interpolated images. In particular, our method can generate reasonable and realistic intermediate results, such as a puppy and an eagle slowly opening their mouths. However, Diffinterp cannot produce results that are consistent with the input and the results are full of flickering artifacts. Diffmorpher cannot handle correct semantic transitions with large object motion, resulting in low-quality and distorted images. The results of Film and LDMVFI produce artifacts and give the impression of fragmentation. Also note the legs of the bird in the top right of Fig. 7, only our method retains the details. Meanwhile, we can observe the girl in the bottom right, our results have the best quality and consistency.

For more comparison results, please see Fig. 8 and the supplementary material. We show a variety of scenarios to demonstrate the superiority of our approach in both image details and semantic consistency. Furthermore, we hope readers to watch the supplementary video for a better dynamic comparison.

4.5 Quantitative Evaluation

As shown in Table 1, our method outperforms all baselines across most metrics by a large margin. Specifically, our approach produces higher-quality images with fewer artifacts, resulting in significantly better FID than other approaches. Additionally, thanks to our effective modeling of semantic representation and consistency, the images generated by our method exhibit higher consistency with input images, achieving optimal LPIPS and WE_{mid} metrics, which measure the consistency of high-level semantic and low-level detail information, respectively. Film [28] and LDMVFI [4] achieve better temporal consistency metrics, but the

Table 1: Quantitative comparisons against all baselines on *InterpBench*. The better approach favors lower FID, LPIPS and WE metrics. The best performance is in **bold**.

Method	FID↓	LPIPS↓	WE ↓	WE_{mid} ↓
DiffInterp	185.7836	0.5375	0.5112	0.9573
Diffmorpher	68.2286	0.3061	0.2673	0.7784
Film	54.2792	0.2313	**0.1244**	0.4176
LDMVFI	48.3469	0.2347	0.1453	0.4373
Ours	**43.1798**	**0.2227**	0.2069	**0.3687**

Table 2: User study. Pairwise comparison results indicate that users prefer our method as better quality and fidelity.

Comparison	Human preference
Diffinterp / **Ours**	5.61% /**94.39%**
Diffmorpher / **Ours**	20.36% /**79.64%**
Film / **Ours**	24.77% /**75.23%**
LDMVFI/ **Ours**	13.65% /**86.35%**

content they generate does not guarantee semantically consistent representations of objects, which can seriously affect the quality of the generated video. We encourage readers to watch the video of the supplementary material, which can reflect the superiority of our method more intuitively.

4.6 User Study

We further conducted a user study to investigate the performance of our method compared to all baselines from a human perspective. Specifically, we collected 30 pairs of images from *InterpBench*. We used different approaches to generate videos with identical settings. During the study, we showed participants with the input image pair and two interpolation videos, one generated by our method and another randomly selected approach, in random order. 139 volunteers were invited to choose the method with better perceptual quality and realism. We report the results in Table 2, which indicates that our method outperforms alternative approaches by a significant margin.

4.7 Ablation Study

Each component of our system plays an important role in improving the generation quality. To justify our design choices, we conduct quantitative ablation studies, as presented in Table 3. Visual results of the ablation study are shown in Fig. 9. In the "w/o our estimator" experiment, we employ "off-the-shelf" optical estimator RAFT [41] as an alternative to obtain the flow. However, on one hand, directly acquiring the optical flow in the RGB space and using it in the latent space generates errors. On the other hand, it is difficult for general optical flow estimators to capture large motions. Therefore, as shown in Fig. 9(a), the rock on the left incorrectly moves with the bear since input image pairs are not derived with accurate correspondence by optical flow. In the "w/o two-level fusion", we directly fuse on z_T space, and we can observe the results in Fig. 9(b) and Fig. 5 are blurry. In the "w/o replace attention" and "w/o lora" experiments, the interpolation images are not consistent with the input two images. In the supplementary material, we discuss the effects of alternative fusion strategies and different total noisy timestamp T.

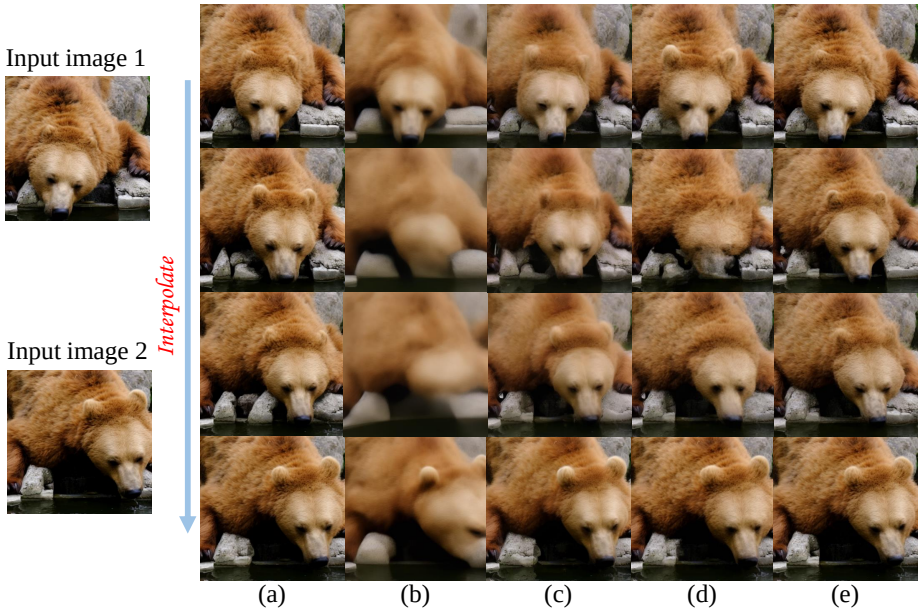

Fig. 9: Visual examples of the ablation study. Each row shows the results of four intermediate images in different settings.

Table 3: Ablation Study on each component of our method.

Method	FID ↓	LPIPS ↓	WE ↓	WE_{mid} ↓
(a) w/o our estimator	53.8510	0.2334	0.2682	0.5229
(b) w/o two-level fusion	138.9329	0.4286	0.2693	0.6510
(c) w/o replace attention	69.4221	0.2919	0.2254	0.4865
(d) w/o lora	76.7142	0.2595	0.2332	0.4744
(e) Full model	**43.1798**	**0.2227**	**0.2069**	**0.3687**

5 Conclusion

In this paper, we present a novel approach for image interpolation with large motion while ensuring the preservation of semantic consistency in the generated results. By leveraging the prior knowledge of a pre-trained text-to-image diffusion model, we propose a natural optical flow estimator, a novel two-level fusion strategy, and a self-attention concatenation and replacement method to generate intermediate images. We conduct extensive experiments to verify the effectiveness of our method. We hope that our work can bring large motion interpolation into the sight of a broader community and motivate further research.

Limitations. Our method leverages the prior of the pre-trained diffusion model, but meanwhile inherits its limitations. Since we employ the diffusion model at low resolution latent space, it may cause texture sticking and be difficult to capture slight motion. We plan to explore more effective solutions in future work.

References

1. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. International journal of computer vision **92**, 1–31 (2011)
2. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
3. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
4. Danier, D., Zhang, F., Bull, D.: Ldmvfi: Video frame interpolation with latent diffusion models. arXiv preprint arXiv:2303.09508 (2023)
5. Epstein, D., Jabri, A., Poole, B., Efros, A.A., Holynski, A.: Diffusion self-guidance for controllable image generation. arXiv preprint arXiv:2306.00986 (2023)
6. Figueirêdo, P., Paliwal, A., Kalantari, N.K.: Frame interpolation for dynamic scenes with implicit flow encoding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 218–228 (2023)
7. Geng, D., Owens, A.: Motion guidance: Diffusion-based image editing with differentiable motion estimators. arXiv preprint arXiv:2401.18085 (2024)
8. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
12. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
13. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
14. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15954–15964 (2023)
15. Kong, L., Jiang, B., Luo, D., Chu, W., Huang, X., Tai, Y., Wang, C., Yang, J.: Ifrnet: Intermediate feature refine network for efficient frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1969–1978 (2022)
16. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 170–185 (2018)
17. Liu, Z., Yeh, R.A., Tang, X., Liu, Y., Agarwala, A.: Video frame synthesis using deep voxel flow. In: Proceedings of the IEEE international conference on computer vision. pp. 4463–4471 (2017)

18. Lu, L., Wu, R., Lin, H., Lu, J., Jia, J.: Video frame interpolation with transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3532–3542 (2022)
19. Luo, G., Dunlap, L., Park, D.H., Holynski, A., Darrell, T.: Diffusion hyperfeatures: Searching through time and space for semantic correspondence. arXiv preprint arXiv:2305.14334 (2023)
20. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
21. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6038–6047 (2023)
22. Montgomery, C., et al.: Xiph. org video test media (derf's collection), the xiph open source community, 1994. Online, https://media. xiph. org/video/derf **3**(5) (2021)
23. Mou, C., Wang, X., Song, J., Shan, Y., Zhang, J.: Dragondiffusion: Enabling drag-style manipulation on diffusion models. arXiv preprint arXiv:2307.02421 (2023)
24. Niklaus, S., Liu, F.: Softmax splatting for video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5437–5446 (2020)
25. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
26. Park, J., Lee, C., Kim, C.S.: Asymmetric bilateral motion estimation for video frame interpolation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14539–14548 (2021)
27. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023)
28. Reda, F., Kontkanen, J., Tabellion, E., Sun, D., Pantofaru, C., Curless, B.: Film: Frame interpolation for large motion. In: European Conference on Computer Vision. pp. 250–266. Springer (2022)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
31. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022)
32. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
33. Shen, L., Li, X., Sun, H., Peng, J., Xian, K., Cao, Z., Lin, G.: Make-it-4d: Synthesizing a consistent long-term dynamic scene video from a single image. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 8167–8175 (2023)

34. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. arXiv preprint arXiv:2401.15977 (2024)
35. Shi, Y., Xue, C., Pan, J., Zhang, W., Tan, V.Y., Bai, S.: Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. arXiv preprint arXiv:2306.14435 (2023)
36. Sim, H., Oh, J., Kim, M.: Xvfi: extreme video frame interpolation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14489–14498 (2021)
37. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
38. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
39. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
40. Tang, L., Jia, M., Wang, Q., Phoo, C.P., Hariharan, B.: Emergent correspondence from image diffusion. arXiv preprint arXiv:2306.03881 (2023)
41. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402–419. Springer (2020)
42. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1921–1930 (2023)
43. Wang, C.J., Golland, P.: Interpolating between images with diffusion models. arXiv preprint arXiv:2307.12560 (2023)
44. Xing, J., Xia, M., Zhang, Y., Chen, H., Wang, X., Wong, T.T., Shan, Y.: Dynamicrafter: Animating open-domain images with video diffusion priors. arXiv preprint arXiv:2310.12190 (2023)
45. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. International Journal of Computer Vision **127**, 1106–1125 (2019)
46. Yang, Z., Yu, Z., Xu, Z., Singh, J., Zhang, J., Campbell, D., Tu, P., Hartley, R.: Impus: Image morphing with perceptually-uniform sampling using diffusion models. arXiv preprint arXiv:2311.06792 (2023)
47. Yu, J., Cun, X., Qi, C., Zhang, Y., Wang, X., Shan, Y., Zhang, J.: Animatezero: Video diffusion models are zero-shot image animators. arXiv preprint arXiv:2312.03793 (2023)
48. Zeng, Y., Wei, G., Zheng, J., Zou, J., Wei, Y., Zhang, Y., Li, H.: Make pixels dance: High-dynamic video generation. arXiv preprint arXiv:2311.10982 (2023)
49. Zhang, G., Zhu, Y., Wang, H., Chen, Y., Wu, G., Wang, L.: Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5682–5692 (2023)
50. Zhang, J., Herrmann, C., Hur, J., Cabrera, L.P., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. arXiv preprint arXiv:2305.15347 (2023)
51. Zhang, K., Zhou, Y., Xu, X., Pan, X., Dai, B.: Diffmorpher: Unleashing the capability of diffusion models for image morphing. arXiv preprint arXiv:2312.07409 (2023)

52. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
53. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
54. Zhewei, H., Tianyuan, Z., Wen, H., Boxin, S., Shuchang, Z.: Rife: Real-time intermediate flow estimation for video frame interpolation. arXiv preprint arXiv:2011.06294 (2020)