Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling

Xiaoyu Shi[1*] Zhaoyang Huang[7*✉] Fu-Yun Wang[1*] Weikang Bian[1*] Dasong Li[1]
Yi Zhang[3] Manyuan Zhang[1] Ka Chun Cheung[2] Simon See[2] Hongwei Qin[3]
Jifeng Dai[4] Hongsheng Li[1,5,6✉]

[1]Multimedia Laboratory, The Chinese University of Hong Kong
[2]NVIDIA AI Technology Center [3]SenseTime Research [4]Tsinghua University
[5]Centre for Perceptual and Interactive Intelligence (CPII) [6]Shanghai AI Laboratory
[7]Avolution AI

xiaoyushi@link.cuhk.edu.hk

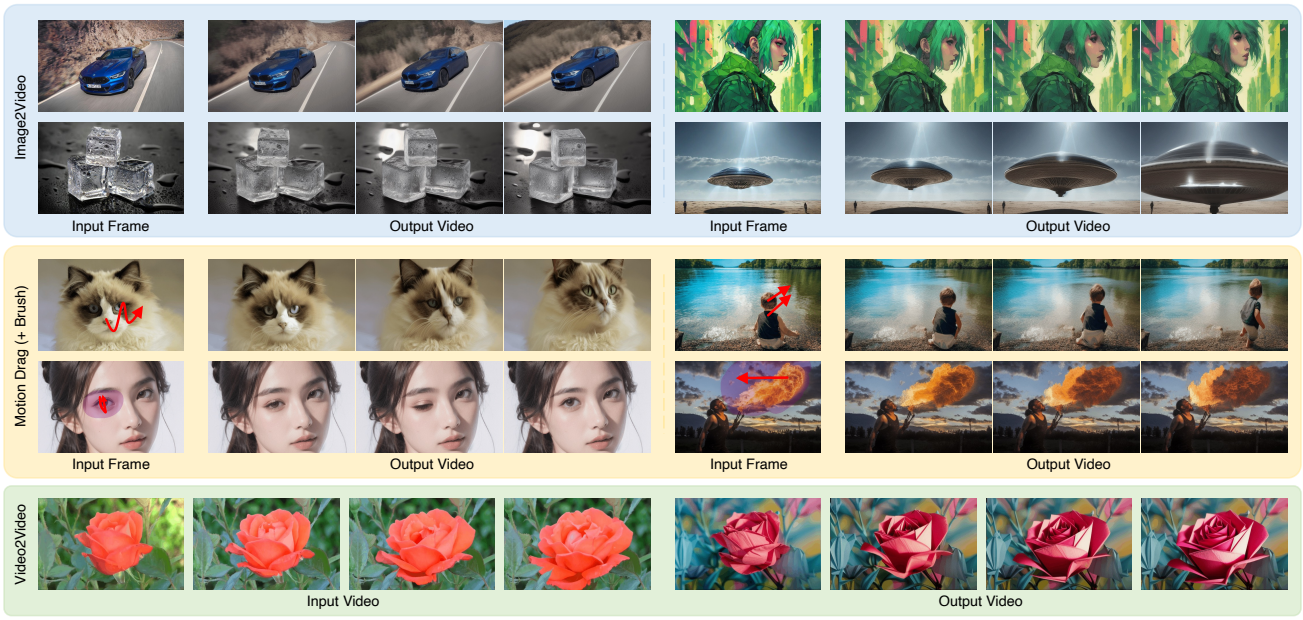


Figure 1: ***Motion-I2V*** *can generate consistent image-to-video results with large motion and viewpoint change. It also naturally supports users to more precisely control the motion trajectories and animated region with sparse trajectories (red curved arrow) and motion brush (purple mask). Additionally, the second stage of Motion-I2V also supports zero-shot video-to-video translation.*

Abstract

We introduce Motion-I2V, a novel framework for consistent and controllable image-to-video generation (I2V). In contrast to previous methods that directly learn the complicated image-to-video mapping, Motion-I2V factorizes I2V into two stages with explicit motion modeling. For the first stage, we propose a diffusion-based motion field predictor, which focuses on deducing the trajectories of the reference image's pixels. For the second stage, we propose motion-augmented temporal attention to enhance the limited 1-D temporal attention in video latent diffusion mod-

els. This module can effectively propagate reference image's feature to synthesized frames with the guidance of predicted trajectories from the first stage. Compared with existing methods, Motion-I2V can generate more consistent videos even at the presence of large motion and viewpoint variation. By training a sparse trajectory ControlNet for the first stage, Motion-I2V can support users to precisely control motion trajectories and motion regions with sparse trajectory and region annotations. This offers more controllability of the I2V process than solely relying on textual instructions. Additionally, Motion-I2V's second stage naturally supports zero-shot video-to-video translation. Both qualitative and quantitative comparisons demonstrate the advantages of Motion-I2V over prior approaches in consistent and controllable image-to-video generation. Please see our project page at `https://xiaoyushi97.github.io/Motion-I2V/`.

1. Introduction

Image-to-video generation (I2V) targets at animating a given image to a video clip with natural dynamics, while preserving the visual appearance. It has widespread applications in fields of film industry, augmented reality, automatic advertising and content creation for social media platforms. Traditional I2V methods, however, focus on specific categories (e.g. human hair [71], fluid [20, 37, 38, 43], portraits [13, 65, 66, 67, 14]). Consequently, such specialization restricts their utility in more diverse, open-domain scenarios.

In recent years, diffusion models [49, 51, 41] trained on web-scale image datasets have made impressive strides in producing high-quality and diverse images. Encouraged by this success, researchers have begun extending these models to the realm of I2V, aiming to leverage the strong image generative priors. These works [80, 6, 72, 64, 77] typically equip text-to-image (T2I) models with 1-D temporal attention modules to create video base models. However, I2V presents more challenges compared to static image generation. It requires modeling the complicated spatial-temporal priors. The narrow temporal receptive field of 1-D temporal attention makes it difficult to ensure temporal consistency of the generated videos, especially in the presence of large motion. Another notable shortage of current I2V works is their limited controllability. These models primarily utilize the reference image and textual instructions as the generation conditions, but lack precise and even interactive control of the generated motions. This is in stark contrast to the field of image manipulation, where techniques like drag-based [54, 40, 44] and region-specific [22, 70] controls have demonstrated substantial efficacy.

To remedy the aforementioned issues, we present Motion-I2V, a framework that factorizes image-to-video generation into two stages. The first stage focuses on predicting the plausible motions, in the form of pixel-wise trajectories. With such explicit motion modeling, the second stage is responsible for generating consistent animation with the predicted dynamics from the first stage. Specifically, in the first stage, we tune a pre-trained video diffusion model for motion field prediction. It takes the reference image and textual instruction as conditions, and predicts the trajectories of all pixels in the reference image. In the second stage, we propose a motion-augmented temporal attention to enhance the video diffusion model. The latent features of the reference image are warped according to all pixels' predicted trajectories and act as guidance via adaptively (through cross-attention) injecting into the synthesized frames at multiple scales. This warping operation brings dynamic temporal receptive field and alleviates the pressure of learning the complicated spatial-temporal patterns with only 1-D temporal attention.

Inspired by previous successes of adapting pre-trained large-scale model [12, 76, 10, 11], we also train a ControlNet [75] for motion prediction in the first stage, which takes sparse trajectories as the condition and generates plausible dense trajectories. This design empowers users to manipulate object motions with very sparse trajectory annotations. Our framework also naturally supports region-specific animation (named *motion brush*), enabling users to animate selected image areas with custom motion masks. Moreover, the second stage of Motion-I2V is capable of achieving video-to-video translation, where the trajectories are obtained from the source video. Users can transform the first frame with existing image-to-image tools and consistently propagate the transformed first frame using the second stage of Motion-I2V. These characteristics grant users enhanced controllability over the I2V process.

2. Related Work

2.1. Image Animation

Animating a single image has attracted a lot of attention in the research field. Previous approaches simulate motion for natural dynamics [20, 32, 37, 43, 73, 50, 27], human faces [66, 67, 14] and bodies [69, 67, 55, 29, 3]. Some of the previous methods employ optical flow to model the motion and uses warping-based rendering techniques. We get inspiration from this line of research and introduce explicit motion modeling into modern generative models. Recent developments on image animation are driven by diffusion models [49, 19, 57]. Mahapatra *et al.* [38] transplant the estimated optical flow to artistic paintings with a pre-trained text-to-image diffusion model. Li *et al.* [33] utilize a diffusion model to handle the natural oscillating motions. These animation approaches can only synthesize specific types of content and motion such as time-lapse videos and body ani-

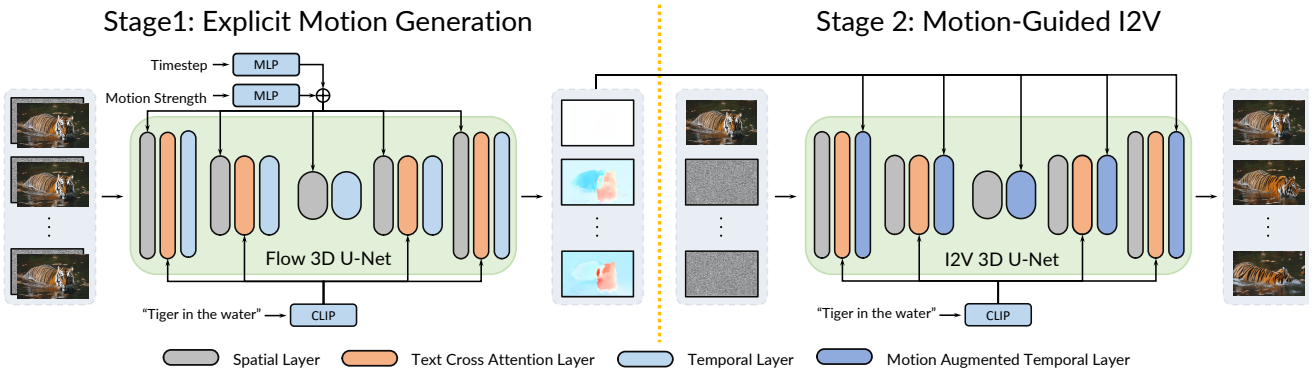


Figure 2: **Overview of Motion-I2V.** The first stage of Motion-I2V targets at deducing the motions that can plausibly animate the reference image. It is conditioned on the reference image and text prompt, and predicts the motion field maps between the reference frame and all the future frames. The second stage propagates reference image's content to synthesize frames. A novel motion-augmented temporal layer enhances 1-D temporal attention with warped features. This operation enlarges the temporal receptive field and alleviates the complexity of directly learning the complicated spatial-temporal patterns.

mation. To solve this problem, some diffusion-based methods [72, 77, 64, 80, 6], are proposed to address the challenge of open-domain image animation. They capitalize on the strong generative priors of pre-trained diffusion models and have achieved unprecedented open-domain animation performance. However, these typically rely on vanilla 1-D temporal attention to learn the complicated image to video mapping. We propose to enlarge the receptive fields with explicit motion prediction.

2.2. Diffusion Models

Diffusion models (DMs) [19, 57] have recently shown more stable training, better sample quality, and flexibility than VAE [31], GAN [15] and FLow models [7]. DALL-E 2[47], GLIDE [41] and Imagen [51] employ diffusion models for text-to-image generation by conducting the diffusion process in pixel space, guided by language models [45, 46] or classifier-free approaches. Stable diffusion [49] shows unprecendented power on text-to-image generation by performing denoising diffusion on the latent space and supports many downstream applications [79, 78].

Recent attention has also been paid to employing diffusion models [49] for video synthesis. Notably, Imagen-Video [18] and Make-A-Video [56] perform denoising diffusion in video pixel space, while MagicVideo [82] models the video distribution in the latent space. Video-P2P [35] and vid2vid-zero [63] propose to edit the video via cross-attention map manipulation. Text-to-video zero [30] construct latent code to model dynamics to employ stable diffusion models [49] for video generation. Wang *et al.* propose a versatile pipeline for extending the video generation length [62]. VideoComposer [64] adopts textual condition, spatial conditions and temporal conditions on the video diffusion models. Zhang *et al.* propose a cascaded i2vgen-XL [77] to ensure semantically and qualitatively excellent video generation. Dynamicrafter [72] proposes a dual-stream image injection mechanism to utilize the motion prior of text-to-video diffusion models. These methods [64, 72, 77] usually allow the diffusion model to handle motion modeling and video generation simulateously, which leads to unrealistic motions and temporal inconsistent visual details. Our method decouples the motion modeling and video details generation to achieve realistic motions and preserve the pleasant details.

2.3. Motion Modeling

Motion modeling aims to understand and predict the movement of objects. Optical flow is a common approach to represent motion, which estimates the displacement field between two consecutive frames. Early work formulated optical flow estimation as an optimization problem that utilized handcrafted features to maximize the visual similarity between image pairs [21, 2, 5, 58]. Deep learning-based methods have recently revolutionized the field of optical flow estimation. FlowNet [9] was the first to introduce deep learning into end-to-end optical flow estimation, which demonstrated the potential of deep learning in this domain. After that, well-designed neural network architecture and synthetic datasets promoted the progress of optical flow estimation [26, 48, 59, 60, 24, 25, 74]. RAFT [61] adopting iterative refinement with correlation volume significantly improved the performance. FlowFormer [23, 53] successfully applied the attention mechanism. Recently, VideoFlow [52] explored the temporal information between multiple frames and achieved state-of-the-art accuracy. Point tracking is another motion modeling method that computes the trajectory

of the query point throughout the video frames [17, 8, 81]. Context-PIPs [68] and CoTracker [28] further improve the accuracy of point tracking with the context information. In recent work, DOT [39] demonstrated that using sparse point tracing to initialize arbitrary distance optical flow estimation can maintain high accuracy with reasonable computational overhead.

3. Method

Given a reference image I_0 and a text prompt c, image-to-video synthesis (I2V) targets at generating a sequence of subsequent video frames $\{\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_N\}$. The key objective is to ensure that the generated video clip not only exhibits plausible motion but also faithfully preserves the visual appearance of the reference image. By leveraging the strong generative priors of diffusion models, recent methods have shown promising open-domain I2V generalization capacity. However, existing methods struggle to maintain temporal consistency largely due to the limited 1-D temporal attention mechanism. Meanwhile, they offer limited control over the generation results. In view of these limitations, we propose Motion-I2V, a novel framework that factorizes image-to-video generation into two stages, as shown in Fig. 2. The first stage, detailed in Sec. 3.2, focuses on predicting plausible motions in the form of pixel-wise trajectories. Building on the predicted motion field, the second stage, described in Section 3.3, utilizes our proposed warpping-augmented temporal attention to synthesize future frames. We start with introducing the preliminary knowledge of the latent diffusion model [49] and video diffusion model in Sec. 3.1.

3.1. Preliminaries

Latent diffusion model. We choose Latent Diffusion Model [49] (LDM) as the backbone generative model. It conducts the denoising process in the latent space of a Variational Autoencoder (VAE). During training, the input image x_0 is first encoded into a latent representation $z_0 = \mathcal{E}(x_0)$ with the frozen encode $\mathcal{E}(\cdot)$. This latent code z_0 is then perturbed as:

$$z_t = \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\overline{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_t)$ with β_t is the noise strength coefficient at step t, and t is uniformly sampled from the timestep index set $\{1, \ldots, T\}$. This process can be regarded as a Markov chain, which incrementally adds Gaussian noise to the latent code z_0. The denoising model ϵ_θ receives z_t as input and is optimized to learn the latent space distribution with the objective function

$$l_\epsilon = ||\epsilon - \epsilon_\theta(z_t, t, c)||_2^2, \tag{2}$$

where c represents the condition, and is the user-provided text prompt in our case. In this paper, we choose Stable Diffusion 1.5 as the base LDM, where the denoising model ϵ_θ is implemented as a U-Net architecture.

Video latent diffusion model. We follow the previous works [16, 4, 72] to expand the image LDM by incorporating temporal modules to create the video latent diffusion model (VLDM). Specifically, the spatial modules from the original image LDM are initialized with the pre-trained weights and are frozen during training. This is to inherit the generative priors from the image LDM. Temporal modules l_ϕ^i comprise of efficient 1-D temporal attention for efficiency are inserted after each spatial attention block l_θ^i. Given a 5-D video tensor z of shape $batch \times channels \times frames \times height \times width$, passed through the temporal modules, its spatial dimensions $height$ and $width$ are reshaped to the batch dimension, yielding 1-D $(batch \times height \times width)$ feature sequences of length $frames$, and are transformed by the self-attention blocks. Such temporal modules are responsible for capturing the temporal dependencies between features of the same spatial location across different frames. For clarity and ease of discussion, we refer to such a VLDM variant with 1-D temporal attentions as the vanilla VLDM.

3.2. Motion Prediction with Video Diffusion Models

The first stage of our proposed image-to-video generation framework targets at deducing the motions that can plausibly animate the reference image. As the latest large-scale diffusion models have been trained on web-scale text-image data, they contain rich knowledge of visual semantics. Such semantic knowledge can greatly benefit motion prediction once the model is trained to associate motion distributions with corresponding objects. Therefore, we choose to adapt the pre-trained stable diffusion model for video motion fields prediction to capitalize on the strong generative priors.

Motion fields modeling. We denote the predicting target of the first stage, the motion fields that animate the reference image, as a sequence of 2D displacement maps $\{f_{0\to i} | i = 1, \ldots, N\}$, where each $f_{0\to i} \in \mathbb{R}^{2\times H\times W}$ is the optical flow between the reference frame and future frame at timestep i. With such a motion fields representation, for each source pixel $\mathbf{p} \in \mathbb{I}^2$ of the reference image I_0, we can easily determine its corresponding coordinate $\mathbf{p}'_i = \mathbf{p} + f_{0\to i}(\mathbf{p})$ on the target image I_i at timestep i.

Training a motion field predictor. To learn a motion field prediction VLDM, we propose a three-step fine-tuning strategy. Initially, we tune a pre-trained LDM to predict a single displacement field conditioned on the reference image and text prompt. Subsequently, we freeze the tuned LDM parameters and integrate the vanilla temporal modules (as described in Sec. 3.1) to create an VLDM for training. This integration allows the model to learn the video's temporal motion distribution by jointly denoising the whole

sequence of the motion fields. After training the temporal modules, we proceed to finetune the entire VLDM model to obtain the final motion fields predictor. We use FlowFormer++ [53] and DOT [39] to estimate optical flow and multi-frame trajectories as ground truth during training, respectively.

Encoding motion fields and conditonal image. As we choose the latent diffusion model for its computational efficiency, we encode each flow map $f_{0\to i} \in \mathbb{R}^{2\times H\times W}$ into a latent representation $z_{0\to i,0} = \mathcal{E}_{flow}(f_i) \in \mathbb{R}^{4\times h\times w}$ using an optical flow VAE encoder, where $h = H/8$ and $w = W/8$. The optical flow autoencoder mirrors the LDM image autoencoder's structure, except that it receives and outputs 2-channel optical flow map rather than 3-channel RGB images. To support image conditioning, we concatenate the latent code of clean reference image $\mathcal{E}(I_0) \in \mathbb{R}^{4\times h\times w}$ along the channel dimension. We initialize all available LDM weights from the SD 1.5 checkpoint, and set weights for the newly added 4 input channels to zero. Additionally, frame stride i is embedded using a two-layer MLP and is added to the time embeddings, serving as a motion strength condition.

3.3. Video Rendering with Predicted Motion

The second stage of Motion-I2V targets at propagating the content of the reference image according to the predicted motion fields from stage 1 to synthesized frames, maintaining the fidelity and temporal consistency. We propose a motion-augmented temporal attention to enhance the vanilla 1-D temporal attention, guided by the predicted motion fields from the first stage. This operation enlarges the temporal receptive field and alleviates the pressure of directly predicting the complicated spatial-temporal patterns from a single image.

Motion-augmented temporal attention. We enhance vanilla VLDM's 1-D temporal attention with the proposed motion-augmented temporal attention and keep its other modules as is. Consider a latent feature $z \in \mathbb{R}^{(1+N)\times C_l\times h_l\times w_l}$ in the $l-th$ temporal layer l^i_ϕ, where c_l, h_l, w_l represent the channel dimension, height and width of the feature, respectively. We omit the batch dimension for brevity. Here we use $z[0] \in \mathbb{R}^{1\times C_l\times h_l\times w_l}$ to denote the feature map corresponding to the reference frame, and $z[1:N] \in \mathbb{R}^{N\times C_l\times h_l\times w_l}$ for the subsequent frames. With the predicted motion fields $\{f_{0\to i} | i = 1,\dots,N\}$ (assuming resized to align the spatial shape) from the first stage, we forward-warp [42] $z[0]$ according to each of the motion field $f_{0\to i}$ as:

$$z[i]' = \mathcal{W}(z[0], f_{0\to i}). \tag{3}$$

These warped feature maps $z[i]'$ are interleaved with original feature maps along the temporal dimension to create augmented features $z_{aug} = [z[0], z[1]', z[1], ..., z[N]', z[N]] \in \mathbb{R}^{(1+2\times N)\times C_l\times h_l\times w_l}$. Then z and z_{aug} are reshaped to $z' \in \mathbb{R}^{(h_l\times w_l)\times(1+N)\times C_l}$ and $z'_{aug} \in \mathbb{R}^{(h_l\times w_l)\times(1+2\times N)\times C_l}$, respectively. In other words, the spatial dimensions are shifted to batch axis and they are treated as 1-D tokens. The reshaped feature maps will be projected and go through the 1-D temporal attention layer:

$$z'' = \text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}(\text{QK}^{\text{T}})\text{V}, \tag{4}$$

where $Q = W^Q z'$, $K = W^K z'_{aug}$ and $V = W^V z'_{aug}$ are the three projections. Notably, the warped feature maps only serve as key and value features. Additionally, we add sinusoidal position encoding to z and z_{aug} to make the network aware of the temporal order the interleaved augmented feature maps. This operation enlarges the receptive fields of the temporal modules guided by the predicted motion fields from the first stage.

Selective noising. At each timestep t of the denoising process, we always concatenate the $clean$ reference image's latent code $z_{ref} \in \mathbb{R}^{1\times 4\times h\times w}$ with other noisy latent codes $z_{0:N,t} \in \mathbb{R}^{N\times 4\times h\times w}$ along the temporal axis. This guarantees that the reference image's content is faithfully preserved in the generation process.

4. Fine-grained Control of Motion-I2V Generation

Relying solely on textual prompt can lead to a lack of fine-grained control of the generation results. This limitation often results in users engaging in multiple rounds of trial and error to achieve their desired creation. In this section, we show that, by virtue of the explicit motion modeling, our Motion-I2V naturally supports fine-grained controls over the I2V process.

4.1. Sparse Trajectory Guided I2V

We propose sparse trajectory guided I2V as an extension of our Motion-I2V framework. Specifically, given an image, users can draw one or multiple pixel-wise trajectories to precisely specify the desired motion of target pixels. Our network is designed to interpret these sparse trajectory inputs and transform them into plausible dense displacement fields with generative priors. Subsequently, these dense motion fields are utilized as inputs for the second stage of Motion-I2V. This strategy effectively enables users to interactively control the I2V process, as shown in Fig. 4.

To achieve this intuitive setting, we train a ControlNet for the first stage, as shown in Fig. 3. Specifically, we clone the down-sample and middle blocks of the 3D-Unet in the first stage as the ControlNet branch. This trainable ControlNet branch is connected to the frozen main branch with zero-initialized convolution layers following [75]. The ControlNet additionally takes sparse displacement fields $f_{sparse} \in \mathbb{R}^{N\times 2\times H\times W}$ and a binary mask

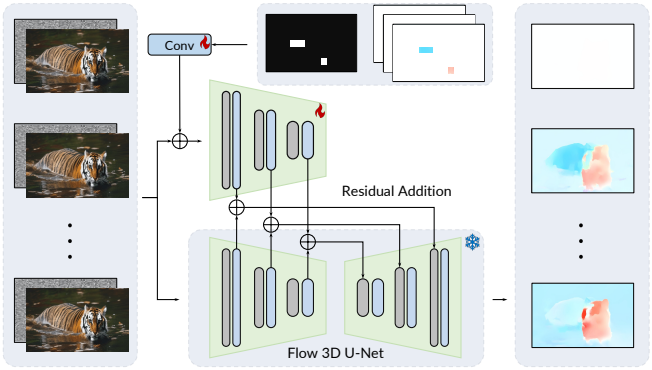


Figure 3: **Overview of trajectory ControlNet.** We train a Trajectory ControlNet based on the pre-trained stage 1 of Motion-I2V. It takes sparse trajectories and corresponding binary mask as additional conditions, and output dense optical flow maps.

$m \in \{0,1\}^{H\times W}$ as conditions, where 1 indicates pixels with given motion. We use a shallow 3D Conv network to encode the concatenation of f_{sparse} and m into 4-dim feature maps and add to the noisy latents as residuals. Please refer to Supplementary for training details.

Input+Drag Frame 2 Frame 9 Frame 16

Figure 4: **Examples of sparse trajectory guided I2V.** Users can precisely control the synthesized motions by drawing one or multiple trajectories (red curved arrow).

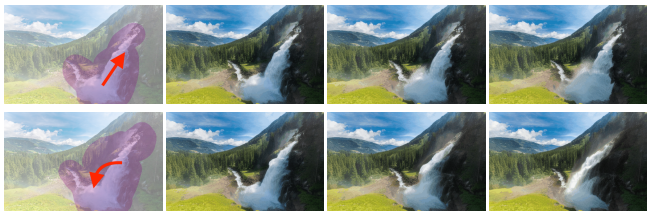
4.2. Region-Specific I2V

Our framework also naturally supports region-specific I2V, where only user-specified regions of the reference image are animated, as shown in Fig. 5. It also can be used in combination with the sparse trajectory guidance, as shown in Fig. 6, for more controllability.

This is a natural extension of the aforementioned sparse trajectory guided I2V. Specifically, the input f_{sparse} is set to all-zero maps. As for the mask m, user-specified regions are set to 0 and other areas as 1. Intuitively, this setting requires the ControlNet to keep the non-specified regions static while infer plausible motions for the user-specified regions.

Input+Brush Frame 2 Frame 9 Frame 16

Figure 5: **Examples of region-specific I2V.** Users can precisely Specify the animated regions by motion brush (purple mask). Unmasked regions remains static.



Input+Drag Frame 2 Frame 9 Frame 16

Figure 6: **Combination of motion trajectories and motion brush**. Motion-I2V supports the combined usage of motion brush and trajectory guidance.

4.3. Zero-Shot Video-to-Video Translation

Our framework also naturally supports video-to-video translation (V2V), where a given video is rendered into a new video of another artistic expression specified by the text prompt, as shown in Fig 7. Specifically, users can utilize existing image-to-image tools to convert the first frame into the target style. Then the displacement fields of the source video can be predicted using off-the-shelf dense point tracker and are used to animate the converted first frame with the second stage of our Motion-I2V.

5. Experiments

5.1. Experimental Setup

Training. We choose Stable Diffusion v1.5 as the base LDM model for the first stage and AnimateDiff v2 as the base VLDM for the second stage. All models are trained on WebVid-10M [1], a large scale text-video dataset. During training, we randomly sample 16-frame video clips with a stride of 8. We employ the AdamW [36] optimizer with a constant learning rate of 3×10^{-5} for training all models. Stage 1 is trained with videos of resolution of 320×512 and stage 2 is 320×320. All experiments are conducted on 32 NVIDIA A100 GPU. Please refer to supplementary for more training details.

Frame 1 Frame 6 Frame 11 Frame 16

Figure 7: **Example of video-to-video translation.** The second stage of Motion-I2V can be used for zero-shot video-to-video translation. The first frame of source video is transformed into the target style using existing image-to-image tools. Then the transformed image can be animated using the second stage of Motion-I2V guided by the motions from the source video.

Evaluation. There are a few image-to-video benchmarks, but they are limited to specific domains. For extensive evaluation, we build a test set that covers various categories, such as human activity, animals, vehicle, natural scenes and AI-generated images. It contains 80 images downloaded from the copyright-free website Pixabay. We use ChatGPT-4V to generate prompts for the image content and possible motion. We use CLIP text-image logits to measure the prompt consistency between prompt and generated frames. We calculate the cosine similarity between consecutive generated frames in the CLIP embedding space to measure the temporal consistency. We further estimate the optical flows between the first and subsequent generated frames to show the motion magnitude.

5.2. Comparison with Other Methods

For quantitative evaluation, we compare our method with the open-sourced state-of-the-art methods VideoComposer [64], I2VGen-XL [77] and DynamiCrafter [72]. Detailed results are shown in Table. 1. Our Motion-I2V outperforms other methods in the prompt-following metric. Additionally, Motion-I2V can generate more consistent videos even with larger motions.

For qualitative comparison, due to limited space, we compare with the quantitatively second best method DynamiCrafter, together with two commercial products Pika1.0 and Gen2. Results are shown in Fig. 8. We observe that DynamiCrafter is not sensitive to motion prompts and tends to generate videos with small motion. This observation is in line with the quantitative results. Pika 1.0 shares similar limited motions but offers better visual quality. Gen2 can generate motions as large as those produced by Motion-I2V, but it suffers from severe distortion. These results verify that Motion-I2V has superior performance in generating consistent results even at the presence of large motions.

Method	Prompt Consistency ↑	Frame Consistency ↑	Average Displacement
VideoComposer	32.62	0.9393	67.15
I2VGen-XL	33.69	0.9650	17.70
DynamiCrafter	34.60	0.9860	3.31
Ours	34.86	0.9871	20.06

Table 1: **Quantitative comparison.** Motion-I2V shows best instruction-following ability and temporal consistency. Meanwhile, Motion-I2V generates relatively large motions.

5.3. Ablation Study

We conduct ablation studies to evaluate the effects of critical design choices. We first train a model without stage 1 (first row of Table. 2) where temporal dependencies are solely learned by the 1-D temporal module. We observe that this model is unstable during inference and easy to generate crashed results. This is in line with the low consistency scores and extremely large motions. Then we add stage 1 but utilize the predicted motion fields in a naive way: directly adding the warped feature maps $z[i]'$ to $z[i]$ rather than using attention to adaptively inject the warped feature to subsequent frames. As shown in the second row of Table. 2, this additional motion information stabilizes the prediction, leading to higher consistency score and more vivid motions. By further changing the fusion type to attention as Equation 3.3, we obtain the final model that achieves the highest consistency score.

Stage 1?	Fusion Type	Prompt Consis. ↑	Frame Consis. ↑	Average Displacement
✗	-	32.95	0.9505	66.44
✓	Addition	33.99	0.9542	48.91
✓	Attention	34.86	0.9871	20.06

Table 2: **Ablation study.** Utilizing the motion fields from stage 1 can significantly stabilize the prediction. Additionally, using attention to adaptively inject the warped features into synthesized frames can further increase consistency and avoid extreme distortions.

6. Limitations and Conclusions

We observe that our method tends to generate videos of medium brightness. This is likely because the noise schedule does not enforce the last timestep to have zero signal-to-noise (SNR) ratio as discussed in [34]. This flawed schedule leads to training-test discrepancy and limit the model's generalization. We believe using the latest Zero-SNR schedulers can alleviate this issue. To conclude, in this paper we propose a novel I2V framework that factorizes the difficult image to video generation task into two stages. In the first stage, we train a diffusion-based motion fields predictor that focuses on deducing the plausible motions. It shows great

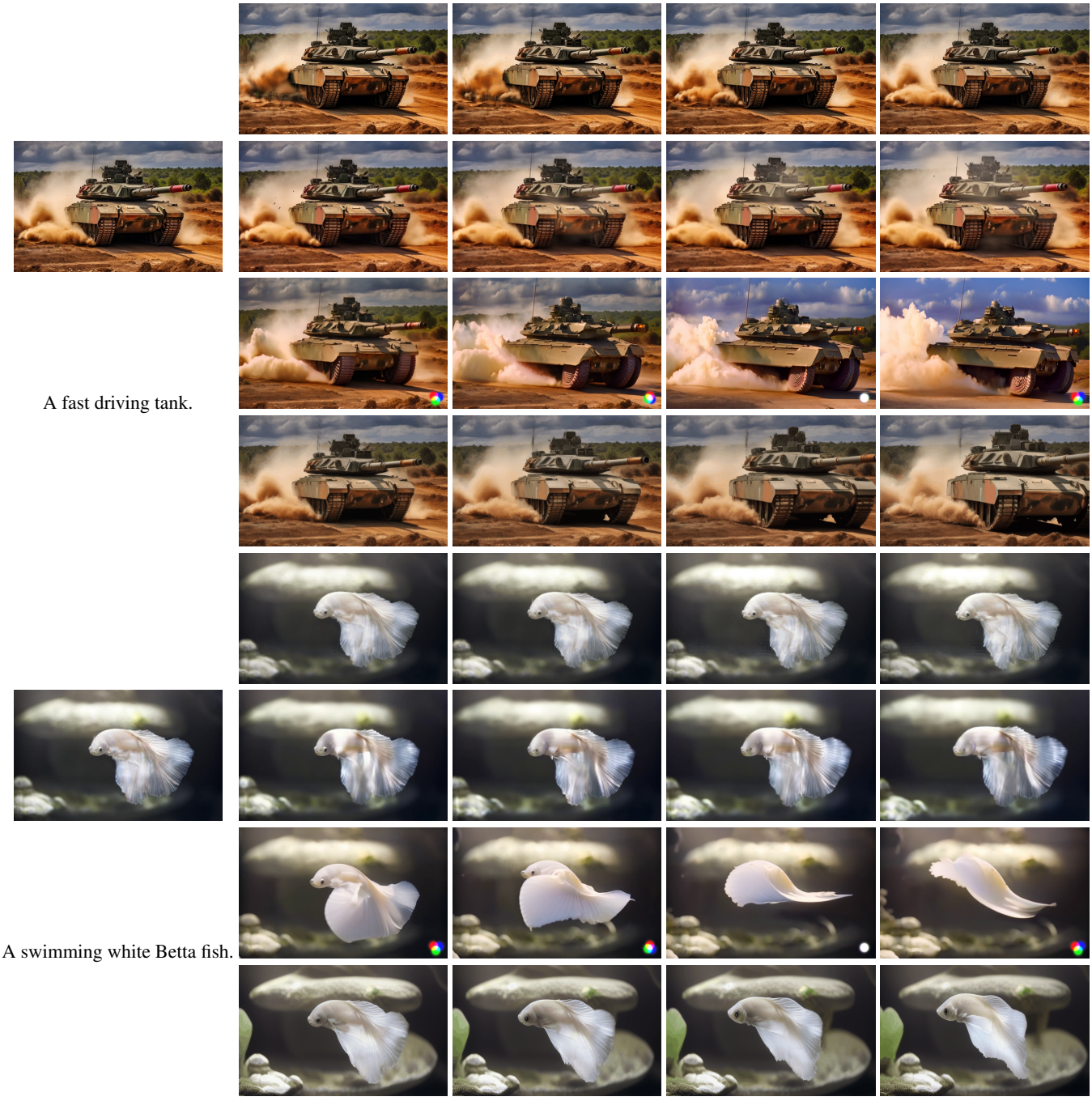


Figure 8: **Qualitative comparison.** DynamiCrafter and Pika 1.0 tend to generate videos of very small motions. Gen2 can generate as large motion as our method, but fails to preserve the identity of the reference image. Our Motion-I2V can synthesize temporally consistent videos in the presence of large motions.

motion generative capacity. In the second stage of video rendering, we identify that the naive 1-D temporal attention limits the temporal modeling capacity. To effectively enlarge temporal receptive field, we propose the motion-guided temporal attention. We further explore to provide more controls over the I2V generation process by training a ControlNet for the first stage. We believe controllability of I2V will obtain more attention from the community in the

future.

7. Acknowledgements

This project is funded in part by National Key R&D Program of China Project 2022ZD0161100, and in part by General Research Fund of Hong Kong RGC Project 14204021.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[2] Michael J Black and Padmanabhan Anandan. A framework for the robust estimation of optical flow. In *1993 (4th) International Conference on Computer Vision*, pages 231–236. IEEE, 1993.

[3] Andreas Blattmann, Timo Milbich, Michael Dorkenwald, and Björn Ommer. Understanding object dynamics for interactive image-to-video synthesis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5167–5177, 2021.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[5] Andrés Bruhn, Joachim Weickert, and Christoph Schnörr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International journal of computer vision*, 61(3):211–231, 2005.

[6] Zuozhuo Dai, Zhenghao Zhang, Yao Yao, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Fine-grained open domain image animation with motion guidance, 2023.

[7] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: non-linear independent components estimation. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[8] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *arXiv preprint arXiv:2211.03726*, 2022.

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[10] Rongyao Fang, Peng Gao, Aojun Zhou, Yingjie Cai, Si Liu, Jifeng Dai, and Hongsheng Li. Feataug-detr: Enriching one-to-many matching for detrs with feature augmentation. *arXiv preprint arXiv:2303.01503*, 2023.

[11] Rongyao Fang, Shilin Yan, Zhaoyang Huang, Jingqiu Zhou, Hao Tian, Jifeng Dai, and Hongsheng Li. Instructseq: Unifying vision tasks with instruction-conditioned multi-modal sequence generation. *arXiv preprint arXiv:2311.18835*, 2023.

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023.

[13] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Trans. Graph.*, 37(6), dec 2018.

[14] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Trans. Graph.*, 37(6), dec 2018.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[16] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[17] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 59–75. Springer, 2022.

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey A. Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *ArXiv*, abs/2210.02303, 2022.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

[20] Aleksander Holynski, Brian L. Curless, Steven M. Seitz, and Richard Szeliski. Animating pictures with eulerian motion fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5819, June 2021.

[21] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[22] Nisha Huang, Fan Tang, Weiming Dong, Tong-Yee Lee, and Changsheng Xu. Region-aware diffusion for zero-shot text-driven image editing. *arXiv preprint arXiv:2302.11797*, 2023.

[23] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. *arXiv preprint arXiv:2203.16194*, 2022.

[24] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8981–8989, 2018.

[25] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2555–2569, 2020.

[26] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[27] Wei-Cih Jhou and Wen-Huang Cheng. Animating still landscape photographs through cloud motion creation. *IEEE Transactions on Multimedia*, 18(1):4–13, 2016.

[28] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

[29] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion, 2023.

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

[31] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[32] Xingyi Li, Zhiguo Cao, Huiqiang Sun, Jianming Zhang, Ke Xian, and Guosheng Lin. 3d cinemagraphy from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4595–4605, June 2023.

[33] Zhengqi Li, Richard Tucker, Noah Snavely, and Aleksander Holynski. Generative image dynamics, 2023.

[34] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5404–5411, 2024.

[35] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control, 2023.

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[37] Aniruddha Mahapatra and Kuldeep Kulkarni. Controllable animation of fluid elements in still images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[38] Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee, Sergey Tulyakov, and Jun-Yan Zhu. Text-guided synthesis of eulerian cinemagraphs. *arXiv preprint arXiv:2307.03190*, 2023.

[39] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: Connecting the dots. *arXiv preprint arXiv:2312.00786*, 2023.

[40] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023.

[41] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021.

[42] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2020.

[43] Makoto Okabe, Ken ichi Anjyo, Takeo Igarashi, and Hans-Peter Seidel. Animating pictures of fluid using video examples. *Computer Graphics Forum*, 28, 2009.

[44] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, New York, NY, USA, 2023. Association for Computing Machinery.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020.

[47] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[48] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4161–4170, 2017.

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[50] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Computer Vision (ICCV), IEEE International Conference on*, 2019.

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[52] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei

Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. *arXiv preprint arXiv:2303.08340*, 2023.

[53] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. *arXiv preprint arXiv:2303.01237*, 2023.

[54] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023.

[55] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021.

[56] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

[57] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[58] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.

[59] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018.

[60] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021.

[61] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

[62] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[63] Wen Wang, kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023.

[64] Xiang* Wang, Hangjie* Yuan, Shiwei* Zhang, Dayou* Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. 2023.

[65] Yaohui WANG, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[66] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1149–1158, 2020.

[67] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022.

[68] BIAN Weikang, Zhaoyang Huang, Xiaoyu Shi, Yitong Dong, Yijin Li, and Hongsheng Li. Context-pips: Persistent independent particles demands context features. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[69] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5901–5910, 2018.

[70] Changming Xiao, Qi Yang, Xiaoqiang Xu, Jianwei Zhang, Feng Zhou, and Changshui Zhang. Where you edit is what you get: Text-guided image editing with region-based attention. *Pattern Recognition*, 139:109458, 2023.

[71] Wenpeng Xiao, Wentao Liu, Yitong Wang, Bernard Ghanem, and Bing Li. Automatic animation of hair blowing in still portrait photos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22963–22975, 2023.

[72] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

[73] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[74] Gengshan Yang and Deva Ramanan. Volumetric correspondence networks for optical flow. *Advances in neural information processing systems*, 32:794–805, 2019.

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[76] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[77] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qing, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. 2023.

[78] Yi Zhang, Dasong Li, Xiaoyu Shi, Dailan He, Kangning Song, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Kbnet: Kernel basis network for image restoration. *arXiv preprint arXiv:2303.02881*, 2023.

[79] Yi Zhang, Xiaoyu Shi, Dasong Li, Xiaogang Wang, Jian Wang, and Hongsheng Li. A unified conditional framework for diffusion-based image restoration. *arXiv preprint arXiv:2305.20049*, 2023.

[80] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models, 2023.

[81] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023.

[82] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models, 2023.