DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors

Jinbo Xing[1], Menghan Xia[2*], Yong Zhang[2], Haoxin Chen[2], Wangbo Yu[3], Hanyuan Liu[1], Gongye Liu[2], Xintao Wang[2], Ying Shan[2], and Tien-Tsin Wong[1*]

[1] The Chinese University of Hong Kong [2] Tencent AI Lab [3] Peking University
https://doubiiu.github.io/projects/DynamiCrafter

Abstract. Animating a still image offers an engaging visual experience. Traditional image animation techniques mainly focus on animating natural scenes with stochastic dynamics (e.g. clouds and fluid) or domain-specific motions (e.g. human hair or body motions), and thus limits their applicability to more general visual content. To overcome this limitation, we explore the synthesis of dynamic content for open-domain images, converting them into animated videos. The key idea is to utilize the motion prior of text-to-video diffusion models by incorporating the image into the generative process as guidance. Given an image, we first project it into a text-aligned rich image context representation space using a query Transformer, which facilitates the video model to digest the image content in a compatible fashion. However, some visual details still struggle to be preserved in the resultant videos. To supplement with more precise image information, we further feed the full image to the diffusion model by concatenating it with the initial noises. Experimental results show that our proposed method can produce visually convincing and more logical & natural motions, as well as higher conformity to the input image. Comparative evaluation demonstrates the notable superiority of our approach over existing competitors.

1 Introduction

Image animation has been a longstanding challenge in the field of computer vision, with the goal of converting still images into video counterparts that display natural dynamics while preserving the original appearance of the images. Traditional heuristic approaches primarily concentrate on synthesizing stochastic and oscillating motions [42, 45] or customizing for specific object categories [31, 39]. However, the strong assumptions imposed on these methods limit their applicability in general scenarios, such as animating open-domain images. Recently, text-to-video (T2V) generative models have achieved remarkable success in creating diverse videos from text. This inspires us to investigate the potential of leveraging such powerful video generation capabilities for image animation.

* Corresponding Authors

Our key idea is to govern the video generation process of T2V diffusion models by incorporating a conditional image. However, achieving the goal of image animation is still non-trivial, as it requires both visual context understanding (essential for creating dynamics) and detail preservation. Recent studies on multi-modal controllable video diffusion models, such as VideoComposer [76] and I2VGen-XL [36], have made preliminary attempts to enable video generation with visual guidance from an image. Unfortunately, both are incompetent for image animation due to their less comprehensive image injection mechanisms, which results in either abrupt temporal changes or low visual conformity to the input image (see Figure 3). To address this challenge, we propose a dual-stream image injection paradigm, comprised of text-aligned image context projection and visual detail guidance, which ensures that the video diffusion model synthesizes detail-preserved dynamic content in a complementary manner. We call this approach *DynamiCrafter*.

Given an image, we first project it into the text-aligned rich image context representation space through a specially designed context learning network. Specifically, it consists of a pre-trained CLIP image encoder to extract text-aligned image features and a learnable query Transformer to further promote its adaptation to the pre-trained T2V diffusion models. The rich image context features are used by the model via cross-attention layers, which will then be combined with the text-conditioned features through gated fusion. To some extent, the learned image context representation trades visual details with text alignment, which helps facilitate semantic understanding of the image so that reasonable and vivid dynamics can be synthesized. To supplement more precise visual details, we feed the full image to the diffusion model by concatenating it with the initial noise. This *dual-stream image injection* paradigm guarantees both plausible dynamic content and visual conformity to the input image.

Extensive experiments are conducted to evaluate our proposed method, which demonstrates notable superiority over existing competitors, including the latest commercial demos (like Gen-2 [20] and PikaLabs [53]). Furthermore, we offer discussion and analysis on some insightful designs for diffusion-model-based image animation, such as the roles of different visual injection streams, the utility of text prompts and their potential for dynamics control, which may inspire follow-ups to push forward this line of technique. Besides image animation, *DynamiCrafter* can be easily adapted to support various applications like storytelling video generation, looping video generation, and generative frame interpolation. Our contributions are summarized as follows:

- We introduce an innovative approach for animating open-domain images by leveraging video diffusion prior, significantly outperforming contemporary competitors.
- We conduct a comprehensive analysis on the conditional space of text-to-video diffusion models and propose a dual-stream image injection paradigm to achieve the challenging goal of image animation.
- We pioneer the study of text-based motion control for open-domain image animation and demonstrate the proof of concept through experiments.

2 Related Work

2.1 Image Animation

Generating animation from still images is a heavily studied research area. Early physical simulation-based approaches [12, 38] focus on simulating the motion of specific objects, resulting in low generalizability due to the independent modeling of each object category. To produce more realistic motion, reference-based methods [11, 39, 52, 54, 61–63, 79] transfer motion or appearance information from reference signals, such as videos, to the synthesis process. Although they demonstrate better temporal coherence, the need for additional guidance limits their practical application. Additionally, a stream of works based on GAN [26, 40, 59] can generate frames by perturbing initial latents or performing random walk in the latent vector space. However, the generated motion is not plausible since the animated frames are just a visualization of the possible appearance space without temporal awareness. Recently, motion prior-based methods [15, 31, 35, 49, 50, 80, 81, 96] animate still images through explicit or implicit image-based rendering with estimated motion field or geometry priors. Similarly, video prediction [2, 17, 32, 34, 44, 73, 84, 87, 93] predicts future video frames starting from single images by learning spatio-temporal priors from video data.

Although existing approaches have achieved impressive performance, they primarily focus on animating motions in curated domains, particularly stochastic [6, 12, 13, 15, 38, 42, 52, 88] and oscillating [45] motion. Furthermore, the animated objects are limited to specific categories, *e.g.*, fluid [31, 48, 52], natural scenes [11, 38, 45, 59, 84], human hair [81], portraits [21, 70, 77, 79], and bodies [4, 7, 33, 39, 71, 80, 86]. In contrast, our work proposes a generic framework for animating open-domain images with a wide range of content and styles, which is extremely challenging due to the overwhelming complexity and vast diversity.

2.2 Video Diffusion Models

Diffusion models (DMs) [28, 65] have recently shown unprecedented generative power in text-to-image (T2I) generation [24, 51, 56–58, 95]. To replicate this success to video generation, numerous video diffusion models (VDMs) [8, 19, 25, 27, 30, 46, 64, 75, 78, 82, 92, 97] are proposed to generate high-quality videos.

Although these models can generate visually pleasing results, they only accept text prompts as the sole semantic guidance, which can be vague and may not accurately reflect users' intentions. Then introducing control signals in T2V, such as structure [16, 83] and pose [41, 47, 94], has been increasingly receiving much attention. However, visual conditions in VDMs [69, 90], such as RGB images, remain under-explored. Most recently, image condition is examined in Seer [23], VideoGen [43], VideoComposer [76], and I2VGen-XL [36] for (text-)image-to-video synthesis. However, they either focus on the curated domain, *i.e.*, indoor objects [23], or fail to generate temporally coherent frames and realistic motions [43, 76] and preserve visual details of the input image [36] due to insufficient context understanding and loss of information of the input image. Our approach

Table 1: Comparison among image-to-video generation frameworks. *: Concurrent.

Method	Image condition injection		Text condition	Pre-trained T2V	T2V prior space
	Detail stream	Context stream			
Seer [23], VideoGen [43]	✓	✗	✓	✗	-
*I2VGen-XL [36]	✗	✓	✗	✗	-
Vid.Comp. [76], *PixelDance [91], *SEINE [10],*EmuVideo [22]	✓	✗	✓	✓	✓
*SVD [5]	✓	✓	✗	✓	✗
DynamiCrafter (Ours)	✓	✓	✓	✓	✓

is built upon text-conditioned VDMs to leverage their rich dynamic prior for animating open-domain images, by introducing *dual-stream image injection* for better semantic/context understanding and conformity to the input image.

Concurrent works. Considering the rapid development in VDMs, numerous concurrent works have emerged, tackling similar image animation problems. Table 1 compares recent and concurrent frameworks with ours. Although SVD [5] adopts a similar dual-stream image condition strategy to ours, it discards the text condition in the underlying pre-trained T2V model, shifting the prior space significantly, which results in less controllability and unrealistic object motions.

3 Method

Given a still image, we aim at animating it to produce a short video, that inherits all the visual content from the image and exhibits implicitly suggested and natural dynamics. Note that the still image can appear in the arbitrary location of the resultant frame sequence. Technically, it can be formulated as image-conditioned video generation that requires strict visual conformity. We tackle this by utilizing the generative priors of pre-trained video diffusion models.

3.1 Preliminary: Video Diffusion Models

Diffusion models [28, 66] are generative models that define a forward diffusion process to convert data $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ into Gaussian noises $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and learn to reverse this process by denoising. The forward process $q(\mathbf{x}_t|\mathbf{x}_0, t)$ contains T timesteps, which gradually adds noise to the data sample $\mathbf{x}_0$ to yield $\mathbf{x}_t$. The denoising process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, t)$ obtains less noisy data $\boldsymbol{x}_{t-1}$ from the noisy input $\boldsymbol{x}_t$ through a denoising network $\epsilon_\theta\left(\mathbf{x}_t, t\right)$, which is supervised by:

$$\min_\theta \mathbb{E}_{t, \mathbf{x}\sim p_{\text{data}}, \epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \|\epsilon - \epsilon_\theta\left(\mathbf{x}_t, t\right)\|_2^2, \tag{1}$$

where ϵ is the sampled ground truth noise and θ indicates the learnable network parameters. Once the model is trained, we can obtain denoised data $\mathbf{x}_0$ from a random noise $\mathbf{x}_T$ through iteratively denoising.

For video generation tasks, Latent Diffusion Models (LDMs) [27] are commonly used to reduce the computation complexity. In this paper, our study is conducted based on an open-source video LDM *VideoCrafter* [9]. Given a

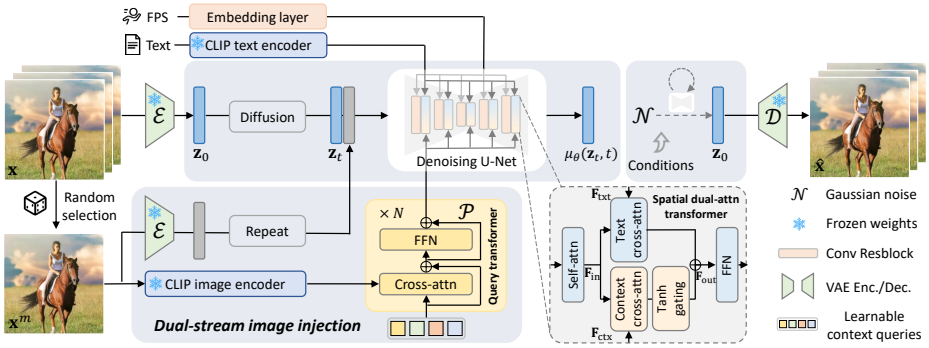

Fig. 1: Flowchart of the proposed *DynamiCrafter*. During training, we randomly select a video frame as the image condition of the denoising process through the proposed dual-stream image injection mechanism to inherit visual details and digest the input image in a context-aware manner. During inference, our model can generate animation clips from noise conditioned on the input still image.

video $\mathbf{x} \in \mathbb{R}^{L\times 3\times H\times W}$, we first encode it into a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x}), \mathbf{z} \in \mathbb{R}^{L\times C\times h\times w}$ frame-by-frame. Then, both the forward diffusion process $\mathbf{z}_t = p(\mathbf{z}_0, t)$ and backward denoising process $\mathbf{z}_t = p_\theta(\mathbf{z}_{t-1}, \mathbf{c}, t)$ are performed in this latent space, where $\mathbf{c}$ denotes possible denoising conditions like text prompt. Accordingly, the generated videos are obtained through the decoder $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$.

3.2 Image Dynamics from Video Diffusion Priors

An open-domain T2V diffusion model is assumed to have diverse dynamic content modeled conditioning on text. To animate a still image with the T2V generative priors, the visual information should be injected into the video generation process in a comprehensive manner. On the one hand, the image should be digested by the T2V model for context understanding, which is important for dynamics synthesis. On the other, the visual details should be preserved in the generated videos. Based on this insight, we propose the dual-stream image injection paradigm, consisting of text-aligned image context projection and visual detail guidance. The overview diagram is illustrated in Fig. 1.

Text-aligned image context projection. To guide video generation with image context, we propose to project the image into a text-aligned embedding space, so that the video model can utilize the image information in a compatible fashion. Since the text embedding is constructed with pre-trained CLIP [55] text encoder, we employ the image encoder counterpart to extract image feature from the input image. Although the global semantic token $\mathbf{f}_{\text{cls}}$ from the CLIP image encoder is well-aligned with image captions, it mainly represents the visual content at semantic level and fails to capture the image's full extent. To obtain more faithful information, we use the full visual tokens $\mathbf{F}_{\text{vis}} = \{\mathbf{f}^i\}_{i=1}^K$ from the last layer of the CLIP image ViT [14], which demonstrated high-fidelity in conditional image generation works [60,89]. To promote the alignment with text

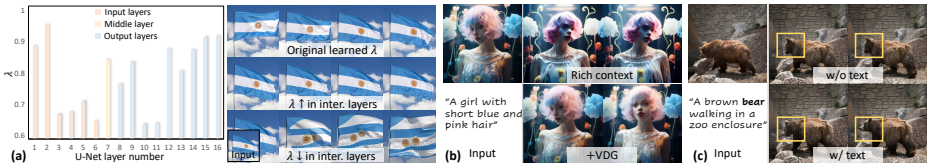

Fig. 2: (a) Visualization of the learned λ across U-Net layers (left), and visual comparisons when manually adjusting λ (right). (b) Comparison of animations produced using rich image context representation solely, and additionally visual detail guidance (VDG). (c) Impact of text with image context representation.

embedding, in other words, to obtain an image context representation that can be interpreted by the denoising U-Net, we utilize a learnable lightweight model $\mathcal{P}$ to translate $\mathbf{F}_{\text{vis}}$ into the final context representation $\mathbf{F}_{\text{ctx}} = \mathcal{P}(\mathbf{F}_{\text{vis}})$. Particularly, $\mathcal{P}$ is realized with a query Transformer architecture [1, 37], which comprises N stacked layers of cross-attention and feed-forward networks (FFN), and is adept at cross-modal representation learning via the cross-attention mechanism.

Subsequently, $\mathbf{F}_{\text{ctx}}$ and the text embedding $\mathbf{F}_{\text{txt}}$ are employed to interact with the U-Net intermediate features $\mathbf{F}_{\text{in}}$ through the dual cross-attention layers:

$$\mathbf{F}_{\text{out}} = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}_{\text{txt}}^{\top}}{\sqrt{d}})\mathbf{V}_{\text{txt}} + \lambda \cdot \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}_{\text{ctx}}^{\top}}{\sqrt{d}})\mathbf{V}_{\text{ctx}}, \tag{2}$$

where $\mathbf{Q} = \mathbf{F}_{\text{in}}\mathbf{W}_{\mathbf{Q}}$, $\mathbf{K}_{\text{txt}} = \mathbf{F}_{\text{txt}}\mathbf{W}_{\mathbf{K}}$, $\mathbf{V}_{\text{txt}} = \mathbf{F}_{\text{txt}}\mathbf{W}_{\mathbf{V}}$, and $\mathbf{K}_{\text{ctx}} = \mathbf{F}_{\text{ctx}}\mathbf{W}'_{\mathbf{K}}$, $\mathbf{V}_{\text{ctx}} = \mathbf{F}_{\text{ctx}}\mathbf{W}'_{\mathbf{V}}$ accordingly. In particular, λ denotes the coefficient that fuses text-conditioned and image-conditioned features, which is achieved through `tanh` gating and adaptively learnable for each layers. This design aims to facilitate the model's ability to absorb image conditions in a layer-dependent manner. As the intermediate layers of the U-Net are more associated with object shapes or poses, and the two-end layers are more linked to appearance [74], we expect that the image features will primarily influence the videos' appearance while exerting relatively less impact on the shape.

Observations and analysis of λ. Fig. 2a (left) illustrates the learned coefficients across different layers, indicating that the image information has a more significant impact on the two-end layers w.r.t. the intermediate layers. To explore further, we manually alter λ in the intermediate layers. As depicted in Fig. 2a (right), increasing λ leads to suppressed cross-frame movements, while decreasing λ poses challenges in preserving the object's shape. This observation not only align with our expectations, but also suggests that in image-conditioned diffusion models, rich-context information influences certain intermediate layers (*e.g.*, layers 7-9) of the U-Net, enabling the model to maintain object shape similar to the input in the presence of motions.

Visual detail guidance (VDG). The rich-informative image context representation enables the video diffusion model to produce videos that closely resemble the input image. However, as shown in Fig. 2b, minor discrepancies may still occur. This is mainly due to the pre-trained CLIP image encoder's limited

capability to fully preserve input image information, as it is designed to align visual and language features. To enhance visual conformity, we propose providing the video model with additional visual details from the image. Specifically, we concatenate the conditional image with per-frame initial noise and feed them to the denoising U-Net as a form of guidance. Therefore, in our dual-stream image injection paradigm, the video diffusion model integrates both global context and local details from the input image in a complementary fashion.

Discussion. (i) *Why are text prompts necessary when a more informative image context is provided?* The text-aligned image representation carries more extensive information than text embedding, which may overburden the T2V model to digest them accurately, *e.g.*, causing shape distortion. Additional text prompts can offer a native global context that enables the model to efficiently utilize image information. Fig. 2c demonstrates how incorporating text can address the issue of shape distortion in the bear's head. Furthermore, as a still image typically contains multiple potential dynamic variations, text prompts can effectively guide the generation of dynamic content tailored to user preferences (see Sec. 5). (ii) *Why is rich image context representation necessary when the visual guidance provides the complete image?* As previously mentioned, the pre-trained T2V model comprises a semantic control space (text embedding) and a complementary random space (initial noise). While the random space effectively integrates low-level information, concatenating the noise of each frame with a fixed image induces spatial misalignment, which may misguide the model in uncontrollable directions. Regarding this, the precise visual context supplied by the image embedding can assist in reliably utilizing visual details (See Sec. 4.5).

3.3 Training Paradigm

The conditional image is integrated through two complementary streams, which play roles in context control and detail guidance, respectively. To modulate them in a cooperative manner, we devise a dedicated training strategy consisting of three stages, *i.e.*, (i) training the image context projection network $\mathcal{P}$, (ii) adapting $\mathcal{P}$ to the T2V model, and (iii) joint fine-tuning with VDG.

Specifically, we propose to train a context projection network $\mathcal{P}$ to extract text-aligned visual information from the input image. Considering the fact that $\mathcal{P}$ takes numerous optimization steps to converge, we propose to train it based on a lightweight T2I model instead of a T2V model, allowing it to focus on image context learning, and then adapt it to the T2V model by jointly training $\mathcal{P}$ and spatial layers (in contrast to temporal layers) of the T2V model. After establishing a compatible image context conditioning branch for T2V, we concatenate the input image with per-frame noise for joint fine-tuning to enhance visual conformity. Here we only fine-tune $\mathcal{P}$ and the VDM's *spatial layers* to avoid disrupting the pre-trained T2V model's temporal prior knowledge with dense image concatenation, which could lead to significant performance degradation and contradict our original intention. Additionally, we randomly select a video frame as the image condition based on two considerations: (i) to prevent the network from learning a shortcut that maps the concatenated image to a

Table 2: Quantitative comparisons with state-of-the-art open-domain image-to-video generation methods on UCF-101 and MSR-VTT for the zero-shot setting.

Method	UCF-101			MSR-VTT		
	FVD ↓	KVD ↓	PIC ↑	FVD ↓	KVD ↓	PIC ↑
VideoComposer	576.81	65.56	0.5269	377.29	26.34	0.4460
I2VGen-XL	571.11	**58.59**	0.5313	289.10	14.70	0.5352
Ours	**429.23**	62.47	**0.6078**	**234.66**	**13.74**	**0.5803**

frame in the specific location, and (ii) to force the image context to be more flexible to avoid offering the over-rigid information for a specific frame, *i.e.*, the objective in the context learning based on T2I. It has been verified in Sec. 4.5.

4 Experiment

4.1 Implementation Details

Our development is mainly based on the open-source T2V model VideoCrafter [9] (@256×256 resolution) and T2I model Stable-Diffusion-v2.1 (SD) [57]. We firstly train $\mathcal{P}$ and the newly injected image cross-attention layers based on SD, with 1000K steps on the learning rate 1×10^{-4} and valid mini-batch size 64. Then we replace SD with VideoCrafter and further fine-tune $\mathcal{P}$ and spatial layers with 30K steps for adaptation, and additional 100K steps with image concatenation on the learning rate 5×10^{-5} and valid mini-batch size 64. Our DynamiCrafter was trained on WebVid-10M [3] dataset by sampling 16 frames with dynamic FPS at the resolution of 256×256 in a batch. The experiments are primarily conducted at this resolution unless otherwise specified. At inference, we adopt DDIM [67] with multi-condition classifier-free guidance [29] (see *Supplement*).

4.2 Quantitative Evaluation

Metrics and datasets. To evaluate the quality and temporal coherence of synthesized videos in both the spatial and temporal domains, we report Fréchet Video Distance (FVD) [72] as well as Kernel Video Distance (KVD) [72]. Following [8, 97], we evaluate the zero-shot generation performance of all the methods on UCF-101 [68] and MSR-VTT [85]. To further investigate the perceptual conformity between the input image and the animation results, we introduce Perceptual Input Conformity (PIC), computed by $\frac{1}{L}\sum_l(1 - D(\mathbf{x}^{\text{in}}, \mathbf{x}^l))$, where $\mathbf{x}^{\text{in}}, \mathbf{x}^l$, L are input image, video frames, and video length, respectively, and we adopt perceptual distance metric DreamSim [18] as the distance function $D(\cdot, \cdot)$.

We evaluate our method against VideoComposer [76] and I2VGen-XL [36] (Concurrent work SVD [5] is excluded due to the mismatch between model configuration and evaluation dataset), with the quantitative results presented in Table 2. According to the results, our proposed method significantly outperforms

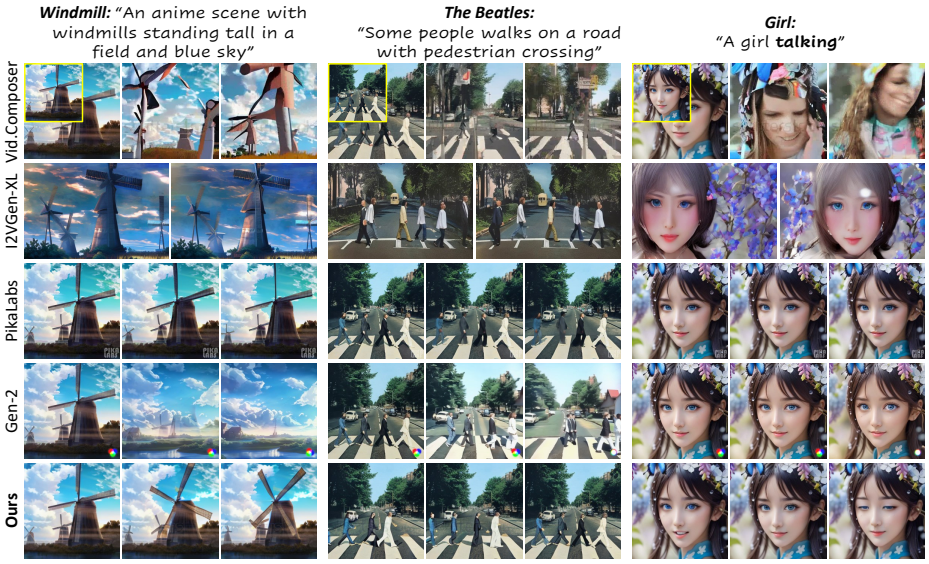

Fig. 3: Visual comparisons of image animation results from VideoComposer, I2VGen-XL, PikaLabs, Gen-2, and our DynamiCrafter.

previous approaches in all evaluation metrics, except for KVD on UCF-101, thanks to the effective dual-stream image injection design for fully exploiting the video diffusion prior. More details of the evaluation are in the *Supplement*.

4.3 Qualitative Evaluation

In addition to the aforementioned approaches, we include two more proprietary commercial products, *i.e.*, PikaLabs [53] and Gen-2 [20], for qualitative comparison. Fig. 3 presents the visual comparison of image animation results with various content and styles. Among all compared methods, our approach generates temporally coherent videos that adhere to the input image condition. In contrast, VideoComposer struggles to produce consistent video frames, as subsequent frames tend to deviate from the initial frame due to inadequate semantic understanding of the input image. I2VGen-XL can generate videos that semantically resemble the input images but fails to preserve intricate local visual details and produce aesthetically appealing results. As commercial products, PikaLabs and Gen-2 can produce appealing high-resolution and long-duration videos. However, Gen-2 suffers from sudden content changes (the 'Windmill' case) and content drifting issues ('The Beatles' and 'Girl' cases). PikaLabs tends to generate still videos with less dynamic and exhibits blurriness when attempting to produce larger dynamics ('The Beatles' case). It is worth noting that our method allows dynamic control through text prompts while other methods suffer from neglecting the text modality (*e.g.*, *talking* in the 'Girl' case).

Fig. 4: Visual comparisons. All generated results are with a resolution of 1024×576.

User study. We conduct a user study to evaluate the perceptual quality. The participants are asked to choose the best in terms of motion quality and temporal coherence, and to select the results with good visual conformity to the input. The statistics from 49 participants' responses are presented in Table 3 (left). Our method demonstrates significant superiority over other open-source methods and achieves comparable, or even better, performance than commercial products.

4.4 Adaption for High-resolution Image Animation

To demonstrate the expandability and generalizability of our methodology and fulfill a broader range of practical applications in line with cutting-edge com-

Table 3: User study statistics of the preference rate for Motion Quality (M.Q.) & Temporal Coherence (T.C.), and selection rate for visual conformity to the input image (I.C.=Input Conformity). Left: 256×256 resolution setting; Right: 1024×576 setting.

Property	Proprietary		Open-source			Proprietary		Open-source	
	PikaLabs	Gen-2	Vid.Com.	I2VGen-XL	Ours	PikaLabs	Gen-2	SVD	Ours
M.Q. ↑	28.60%	22.91%	2.09%	7.56%	**38.84**%	19.13%	17.39%	27.39%	**36.09**%
T.C. ↑	32.09%	26.05%	2.21%	6.51%	**33.14**%	**30.43**%	15.65%	24.35%	29.57%
I.C. ↑	79.07%	64.77%	18.14%	15.00%	**79.88**%	74.78%	62.61%	83.91%	**86.96**%

mercial [20, 53] and academic [5] image-to-video generation techniques, we further train a high-resolution (*i.e.*, DynamiCrafter$_{1024}$ @1024×576) version via multi-stage fine-tuning, *i.e.*, training a DynamiCrafter$_{512}$ @512×320 based on VideoCrafter1 @512×320, and then fine-tuning it into DynamiCrafter$_{1024}$ (See *Supplement*). Fig. 4 displays the qualitative comparisons, illustrating that our method can generate temporally coherent and realistic dynamics adhering to the input images, *e.g.*, the rising beer level in the glass mug (the 'Beer' case) and walk (the 'Robot' case). Conversely, PikaLabs and Gen-2 produce still (the 'Robot' case) or temporally inconsistent videos (the 'Beer' case). In addition, Gen-2 experiences significant color-shifting issues compared to the input image. While SVD tends to animate images using camera motions and fails to produce natural object motions, despite being trained on a larger curated proprietary dataset. Moreover, SVD does not support text prompts as input, rendering the animation results less controllable. The user study with the same configuration as the above, except for high-resolution samples shown in Table 3 (right) demonstrates the superiority of our DynamiCrafter$_{1024}$ in motion quality & input conformity.

4.5 Ablation Studies

Dual-stream image injection. To investigate the roles of each image conditioning stream, we examine two variants: **i). Ours w/o ctx**, by removing the image context projection stream, **ii). Ours w/o VDG**, by removing the visual detail guidance stream. Table 4 compares our full method and these variants. The performance of 'w/o ctx' declines significantly due to its inability to semantically comprehend the input image without injection of rich image context, leading to temporal inconsistencies in the generated videos (2nd row in Fig. 5a). Although removing the VDG (w/o VDG) can yield better FVD scores, it causes severe shape distortions and exhibits limited motion magnitude, as the remaining image context condition can only provide semantic-level image information. Moreover, while it achieves a higher PIC score, it fails to capture all the visual details of the input image, as evidenced by the 3rd row in Fig. 5a.

We then study several key designs in the image context projection stream: adaptive gating λ and full visual tokens in CLIP image encoder. Eliminating the adaptive gating λ (w/o λ) leads to a slight decrease in model performance. This is because, without considering the nature of the denoising U-Net layers, context

Table 4: Ablation study on the dual-stream image injection and training paradigm.

Metric	Ours	Dual-stream image injection				Training paradigm	
		w/o ctx	w/o VDG	w/o λ	Ours$_{\mathrm{G}}$	Ft. ent.	1st frame
FVD ↓	234.66	372.80	159.24	241.38	286.84	364.11	309.23
PIC ↑	0.5803	0.4916	0.6945	0.5708	0.5717	0.5564	0.5673

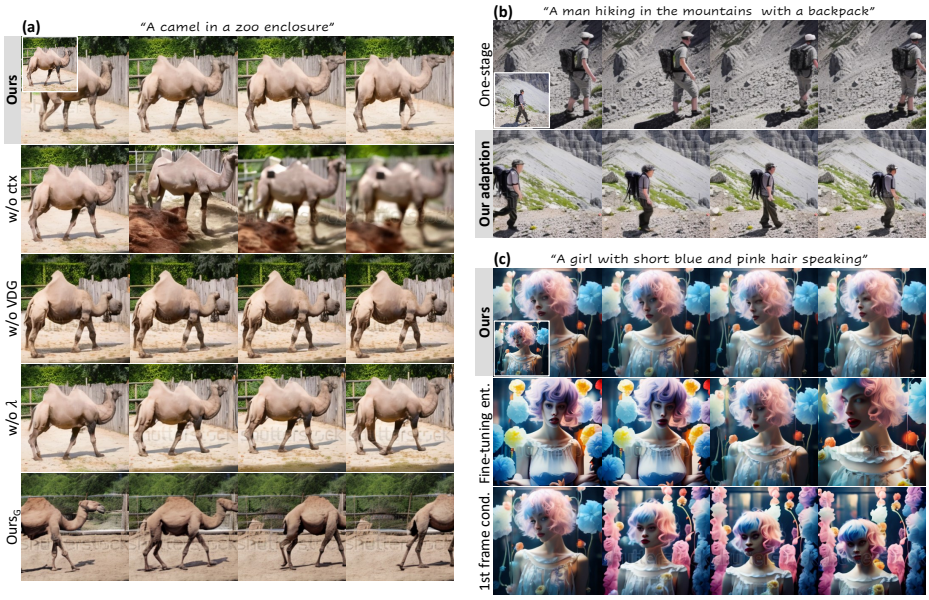

Fig. 5: Visual comparisons of (a) different variants of our method, (b) the image context projection stream learned in one-stage and our two-stage adaption strategy, and (c) different training paradigms.

information cannot be adaptively integrated into the T2V model, resulting in shaky generated videos and unnatural motions (4th row in Fig. 5a). On the other hand, using a strategy (Ours$_{\mathrm{G}}$) like I2VGen-XL that utilizes a single CLIP global token may generate results that are only semantically similar to the input due to the absence of full image extent. In contrast, our full method effectively leverages the video diffusion prior for image animation with natural motion, coherent frames, and visual conformity to the input image.

Training paradigm. We further examine the specialized training paradigm to ensure the model works. We firstly construct a baseline by training the context projection network $\mathcal{P}$ based on the pre-trained T2V and keeping other settings unchanged. As illustrated in Fig. 5b, this baseline (one-stage) converges at a significantly slow pace, resulting in only coarse-grained context conditioning with the same optimization steps. This may make it challenging for the T2V model to harmonize the dual-stream conditions after incorporating VDG.

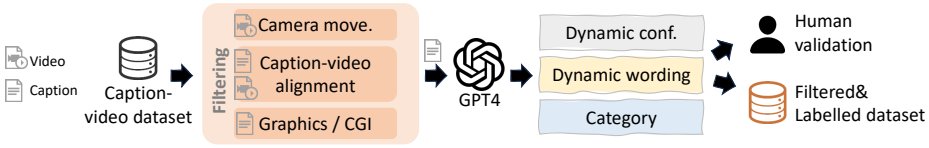

Fig. 6: Illustration of dataset filtering and annotation process.

Fig. 7: Image animation results from different methods with motion control using text.

After obtaining a compatible image context projection stream $\mathcal{P}$, we further incorporate image concatenation with per-frame noise to enhance visual conformity by jointly fine-tuning $\mathcal{P}$ and *spatial layers* of the T2V model. We construct a baseline by fine-tuning the entire T2V model, and Table 4 (Ft. ent.) shows that this baseline results in an unstable model that is prone to collapse, disrupting the temporal prior. Additionally, to study the effectiveness of our random selection conditioning strategy, we train a baseline (1st frame cond.) that consistently uses the first video frame as the conditional image. Table 4 reveals its inferior performance in terms of both FVD and PIC, which can be attributed to the "content sudden change" effect observed in the generated videos (Fig. 5c (bottom)). We hypothesize that the model discovers a suboptimal shortcut for mapping the concatenated image to the first frame while neglecting other frames.

5 Discussions on Motion Control using Text

Since images are typically associated with multiple potential dynamics, text can complementarily guide the generation of dynamic content. However, captions in existing datasets often consist of a combination of scene descriptive words and less dynamic descriptions, potentially causing the model to overlook dynamics/motions during learning. For image animation, pure motion descriptions should be treated as text conditions to train the model in a *decoupled* manner.

Dataset construction. To enable the decoupled training, we construct a dataset by filtering and re-annotating the WebVid10M dataset, as illustrated in Fig. 6. The constructed dataset contains captions with purer *dynamic wording*, such as "Man doing push-ups.", and categories, *e.g.*, `human`.

We then train a model DynamiCrater$_{\text{DCP}}$ using the dataset and validate its effectiveness with 40 image-prompt testing cases featuring human figures with

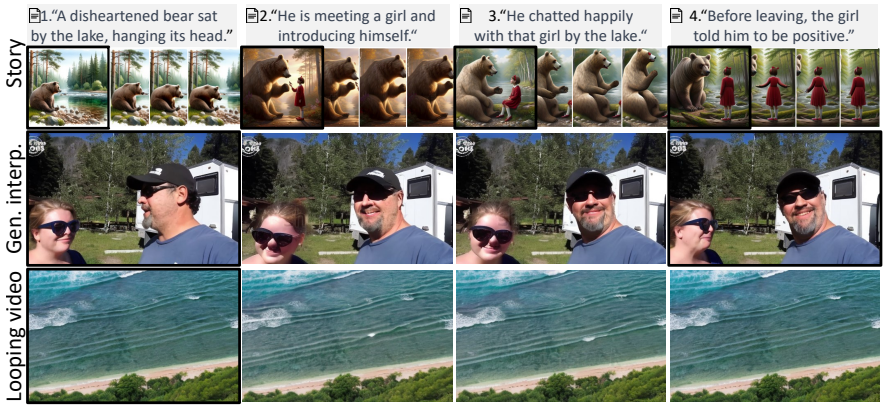

Fig. 8: Applications of our *DynamiCrafter*. □: input images.

ambiguous potential actions, and prompts describing various motions (*e.g.*, "Man waving hands" and "Man clapping"). We measure the average CLIP similarity (CLIP-SIM) between the prompt and video results, and DynamiCrater$_{\text{DCP}}$ improves the performance from 0.17 to 0.19 in terms of CLIP-SIM score. The visual comparison in Fig. 7 shows that Gen-2 and PikaLabs cannot support motion control using text, while our DynamiCrafter reflects the text prompt and is further enhanced in DynamiCrafter$_{\text{DCP}}$ with the proposed decoupled training.

6 Applications

Storytelling with shots. First, we utilize ChatGPT to generate a story script and corresponding shots (images). Then storytelling videos can be generated by animating those shots, as displayed in Fig. 8 (top). **Generative frame interpolation.** With minor modifications, our framework can be adapted to facilitate generative frame interpolation. Specifically, we provide both $\mathbf{x}^1$ and $\mathbf{x}^L$ as visual detail guidance and leave middle frames empty during training. During inference, we provide two images for interpolation. We experiment with building this on top of VideoCrafter1 @512×320. The interpolation results are shown in Fig. 8 (middle). **Looping video generation.** The modified model enables looping video generation by inputting the same images for $\mathbf{x}^1$ and $\mathbf{x}^L$ (Fig. 8 (bottom)).

7 Conclusion

We introduced *DynamiCrafter*, an effective framework for animating open-domain images by leveraging video diffusion priors with the proposed dual-stream image injection mechanism and dedicated training paradigm. Our experimental results highlight the effectiveness and superiority of our approach compared to existing methods. Lastly, we explored text-based dynamic control for image animation and demonstrated the versatility of our framework across various applications.

Acknowledgements

This project is partially supported by Hong Kong Innovation and Technology Fund (ITF) (ref: ITS/307/20FP). We thank the anonymous reviewers for their constructive feedback.

References

1. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
2. Babaeizadeh, M., Finn, C., Erhan, D., Campbell, R., Levine, S.: Stochastic variational video prediction. In: ICLR (2018)
3. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV (2021)
4. Bertiche, H., Mitra, N.J., Kulkarni, K., Huang, C.H.P., Wang, T.Y., Madadi, M., Escalera, S., Ceylan, D.: Blowing in the wind: Cyclenet for human cinemagraphs from still images. In: CVPR (2023)
5. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
6. Blattmann, A., Milbich, T., Dorkenwald, M., Ommer, B.: ipoke: Poking a still image for controlled stochastic video synthesis. In: ICCV (2021)
7. Blattmann, A., Milbich, T., Dorkenwald, M., Ommer, B.: Understanding object dynamics for interactive image-to-video synthesis. In: CVPR (2021)
8. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023)
9. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
10. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: Seine: Short-to-long video diffusion model for generative transition and prediction. In: ICLR (2024)
11. Cheng, C.C., Chen, H.Y., Chiu, W.C.: Time flies: Animating a still image with time-lapse video as reference. In: CVPR (2020)
12. Chuang, Y.Y., Goldman, D.B., Zheng, K.C., Curless, B., Salesin, D.H., Szeliski, R.: Animating pictures with stochastic motion textures. In: ACM SIGGRAPH (2005)
13. Dorkenwald, M., Milbich, T., Blattmann, A., Rombach, R., Derpanis, K.G., Ommer, B.: Stochastic image-to-video synthesis using cinns. In: CVPR (2021)
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
15. Endo, Y., Kanamori, Y., Kuriyama, S.: Animating landscape: self-supervised learning of decoupled motion and appearance for single-image video synthesis. ACM TOG **38**(6), 1–19 (2019)

16. Esser, P., Chiu, J., Atighehchian, P., Granskog, J., Germanidis, A.: Structure and content-guided video synthesis with diffusion models. In: ICCV (2023)
17. Franceschi, J.Y., Delasalles, E., Chen, M., Lamprier, S., Gallinari, P.: Stochastic latent residual video prediction. In: ICML (2020)
18. Fu, S., Tamir, N., Sundaram, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In: NeurIPS (2023)
19. Ge, S., Nah, S., Liu, G., Poon, T., Tao, A., Catanzaro, B., Jacobs, D., Huang, J.B., Liu, M.Y., Balaji, Y.: Preserve your own correlation: A noise prior for video diffusion models. In: ICCV (2023)
20. Gen-2: Gen-2. *Gen-2*. Accessed Nov.1, 2023, Feb.01, 2024 [Online] `https://research.runwayml.com/gen2`, `https://research.runwayml.com/gen2`
21. Geng, J., Shao, T., Zheng, Y., Weng, Y., Zhou, K.: Warp-guided gans for single-photo facial animation. ACM TOG **37**(6), 1–12 (2018)
22. Girdhar, R., Singh, M., Brown, A., Duval, Q., Azadi, S., Rambhatla, S.S., Shah, A., Yin, X., Parikh, D., Misra, I.: Emu video: Factorizing text-to-video generation by explicit image conditioning. arXiv preprint arXiv:2311.10709 (2023)
23. Gu, X., Wen, C., Song, J., Gao, Y.: Seer: Language instructed video prediction with latent diffusion models. arXiv preprint arXiv:2303.14897 (2023)
24. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. arXiv preprint arXiv:2310.07702 (2023)
25. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arXiv preprint arXiv:2211.13221 (2022)
26. Hinz, T., Fisher, M., Wang, O., Wermter, S.: Improved techniques for training single-image gans. In: WACV (2021)
27. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
28. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
29. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
30. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022)
31. Holynski, A., Curless, B.L., Seitz, S.M., Szeliski, R.: Animating pictures with eulerian motion fields. In: CVPR (2021)
32. Höppe, T., Mehrjou, A., Bauer, S., Nielsen, D., Dittadi, A.: Diffusion models for video prediction and infilling. TMLR (2022)
33. Hu, L., Gao, X., Zhang, P., Sun, K., Zhang, B., Bo, L.: Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117 (2023)
34. Hu, X., Huang, Z., Huang, A., Xu, J., Zhou, S.: A dynamic multi-scale voxel flow network for video prediction. In: CVPR (2023)
35. Hu, Y., Luo, C., Chen, Z.: Make it move: controllable image-to-video generation with text descriptions. In: CVPR (2022)
36. I2VGen-XL: I2vgen-xl. *ModelScope*. Accessed Oct.15, 2023 [Online], `https://modelscope.cn/models/damo/Image-to-Video/summary`
37. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: ICML (2021)

38. Jhou, W.C., Cheng, W.H.: Animating still landscape photographs through cloud motion creation. IEEE TMM **18**(1), 4–13 (2015)
39. Karras, J., Holynski, A., Wang, T.C., Kemelmacher-Shlizerman, I.: Dreampose: Fashion image-to-video synthesis via stable diffusion. arXiv preprint arXiv:2304.06025 (2023)
40. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
41. 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. arXiv preprint arXiv:2303.13439 (2023)
42. Lee, A.X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., Levine, S.: Stochastic adversarial video prediction. arXiv preprint arXiv:1804.01523 (2018)
43. Li, X., Chu, W., Wu, Y., Yuan, W., Liu, F., Zhang, Q., Li, F., Feng, H., Ding, E., Wang, J.: Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. arXiv preprint arXiv:2309.00398 (2023)
44. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Flow-grounded spatial-temporal video prediction from still images. In: ECCV (2018)
45. Li, Z., Tucker, R., Snavely, N., Holynski, A.: Generative image dynamics. In: CVPR (2024)
46. Luo, Z., Chen, D., Zhang, Y., Huang, Y., Wang, L., Shen, Y., Zhao, D., Zhou, J., Tan, T.: Videofusion: Decomposed diffusion models for high-quality video generation. In: CVPR (2023)
47. Ma, Y., He, Y., Cun, X., Wang, X., Shan, Y., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. arXiv preprint arXiv:2304.01186 (2023)
48. Mahapatra, A., Kulkarni, K.: Controllable animation of fluid elements in still images. In: CVPR (2022)
49. Mallya, A., Wang, T.C., Liu, M.Y.: Implicit warping for animation with image sets. In: NeurIPS (2022)
50. Ni, H., Shi, C., Li, K., Huang, S.X., Min, M.R.: Conditional image-to-video generation with latent flow diffusion models. In: CVPR (2023)
51. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022)
52. Okabe, M., Anjyo, K., Igarashi, T., Seidel, H.P.: Animating pictures of fluid using video examples. In: CGF. vol. 28, pp. 677–686 (2009)
53. PikaLabs: Pikalabs. *PikaLabs*. Accessed Nov.1, 2023, Feb.01, 2024 [Online] `https://www.pika.art/`, `https://www.pika.art/`
54. Prashnani, E., Noorkami, M., Vaquero, D., Sen, P.: A phase-based approach for animating images using video examples. In: CGF. vol. 36, pp. 303–311 (2017)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
56. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
57. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
58. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS (2022)

59. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: ICCV (2019)
60. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411 (2023)
61. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: CVPR (2019)
62. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. In: NeurIPS (2019)
63. Siarohin, A., Woodford, O.J., Ren, J., Chai, M., Tulyakov, S.: Motion representations for articulated animation. In: CVPR (2021)
64. 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. In: ICLR (2023)
65. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
66. Sohl-Dickstein, J., Weiss, E.A., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
67. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
68. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
69. Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. In: NeurIPS (2023)
70. Tao, J., Gu, S., Li, W., Duan, L.: Learning motion refinement for unsupervised face animation. In: NeurIPS (2023)
71. Tao, J., Wang, B., Ge, T., Jiang, Y., Li, W., Duan, L.: Motion transformer for unsupervised image animation. In: ECCV (2022)
72. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. In: ICLR workshop (2019)
73. Voleti, V., Jolicoeur-Martineau, A., Pal, C.: Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. In: NeurIPS (2022)
74. Voynov, A., Chu, Q., Cohen-Or, D., Aberman, K.: $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522 (2023)
75. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571 (2023)
76. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. arXiv preprint arXiv:2306.02018 (2023)
77. Wang, Y., Bilinski, P., Bremond, F., Dantcheva, A.: Imaginator: Conditional spatio-temporal gan for video generation. In: WACV (2020)
78. Wang, Y., Chen, X., Ma, X., Zhou, S., Huang, Z., Wang, Y., Yang, C., He, Y., Yu, J., Yang, P., et al.: Lavie: High-quality video generation with cascaded latent diffusion models. arXiv preprint arXiv:2309.15103 (2023)
79. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. In: ICLR (2021)
80. Weng, C.Y., Curless, B., Kemelmacher-Shlizerman, I.: Photo wake-up: 3d character animation from a single photo. In: CVPR (2019)
81. Xiao, W., Liu, W., Wang, Y., Ghanem, B., Li, B.: Automatic animation of hair blowing in still portrait photos. In: ICCV (2023)
82. Xing, J., Liu, H., Xia, M., Zhang, Y., Wang, X., Shan, Y., Wong, T.T.: Tooncrafter: Generative cartoon interpolation. arXiv preprint arXiv:2405.17933 (2024)

83. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. arXiv preprint arXiv:2306.00943 (2023)
84. Xiong, W., Luo, W., Ma, L., Liu, W., Luo, J.: Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In: CVPR (2018)
85. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016)
86. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498 (2023)
87. Xue, T., Wu, J., Bouman, K., Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In: NeurIPS (2016)
88. Xue, T., Wu, J., Bouman, K.L., Freeman, W.T.: Visual dynamics: Stochastic future generation via layered cross convolutional networks. IEEE TPAMI **41**(9), 2236–2250 (2018)
89. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
90. Yin, S., Wu, C., Liang, J., Shi, J., Li, H., Ming, G., Duan, N.: Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089 (2023)
91. 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)
92. Zhang, D.J., Wu, J.Z., Liu, J.W., Zhao, R., Ran, L., Gu, Y., Gao, D., Shou, M.Z.: Show-1: Marrying pixel and latent diffusion models for text-to-video generation. arXiv preprint arXiv:2309.15818 (2023)
93. Zhang, J., Xu, C., Liu, L., Wang, M., Wu, X., Liu, Y., Jiang, Y.: Dtvnet: Dynamic time-lapse video generation via single still image. In: ECCV (2020)
94. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)
95. Zhang, Y., Xing, J., Lo, E., Jia, J.: Real-world image variation by aligning diffusion inversion chain. arXiv preprint arXiv:2305.18729 (2023)
96. Zhao, J., Zhang, H.: Thin-plate spline motion model for image animation. In: CVPR (2022)
97. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)