Follow-Your-Click: Open-domain Regional Image Animation via Motion Prompts

Yue Ma[1*], Yingqing He[1*], Hongfa Wang[2,3*], Andong Wang[2], Leqi Shen[3], Chenyang Qi[1], Jixuan Ying[3], Chengfei Cai[2], Zhifeng Li[2], Heung-Yeung Shum[1,3], Wei Liu[2✉], Qifeng Chen[1✉]

[1]The Hong Kong University of Science and Technology, Hong Kong
[2]Tencent, Hunyuan, China
[3]Tsinghua University, China
ymacn@cse.ust.hk, yhebm@connect.ust.hk, {hongfawang, fletchercai}@tencent.com, wangad@connect.hku.hk, {lunarshen, chenyangqi67, zhifeng0.li}@gmail.com, yingjx23@mails.tsinghua.edu.cn, {hshum, cqf}@ust.hk, wl2223@columbia.edu,

Abstract

Despite recent advances in image-to-video generation, better controllability and local animation are less explored. Most existing image-to-video methods are not locally aware and tend to move the entire scene. However, human artists may need to control the movement of different objects or regions. Additionally, current I2V methods require users not only to describe the target motion but also to provide redundant detailed descriptions of frame contents. These two issues hinder the practical utilization of current I2V tools. In this paper, we propose a practical framework, named Follow-Your-Click, to achieve image animation with a simple user click (for specifying *what* to move) and a motion prompt (for specifying *how* to move). Technically, we propose the first-frame masking strategy, which significantly improves the video generation quality, and a motion-augmented module equipped with a motion prompt dataset to improve the motion prompt following abilities of our model. To further control the motion speed, we propose flow-based motion magnitude control to control the speed of target movement more precisely. Our framework has simpler yet precise user control and better generation performance than previous methods. Extensive experiments compared with 7 baselines, including both commercial tools and research methods on 8 metrics, suggest the superiority of our approach.

Code — https://follow-your-click.github.io/

1 Introduction

Image-to-video generation (I2V) aims to animate an image into a dynamic video clip with reasonable movements. It has widespread applications in the filmmaking industry, augmented reality, and automatic advertising. Traditionally, image animation methods mainly focus on domain-specific categories, such as natural scenes (Cheng, Chen, and Chiu 2020; Li et al. 2023), human hair (Xiao et al. 2023), portraits (Wang et al. 2022) and bodies (Bertiche et al. 2023; Weng, Curless, and Kemelmacher-Shlizerman 2019), limiting their practical application in real world. In recent years, the significant advancements in the diffusion models (Rombach et al. 2022; Saharia et al. 2022) trained on large-scale image datasets have enabled the generation of diverse and realistic images based on text prompts. Encouraged by this success, researchers have begun extending these models to the realm of I2V, aiming to leverage the strong image generation priors for image-to-video generation (Xing et al. 2023; Shi et al. 2024; Chai et al. 2023).

However, existing I2V works (Chai et al. 2023; Xing et al. 2023; Wang et al. 2023b; i2v 2023) have a lack of control over which part of the image needs to be moved, and they produce videos with the movement of the entire scene; And some works such as SVD (Chai et al. 2023) tend to deliver videos always with camera movement, ignoring the more vivid object movement. They cannot achieve regional image animation which is important to human artists (*e.g.*, the user may want to animate the foreground object while keeping the background static). Besides, the typical prompts that users provide to I2V models are the descriptions of the entire scene contents. However, the spatial content is fully described via the input image which is not necessary for users to describe it again. In fact, a more intuitive way is to provide motion-only prompts, but current approaches are less sensitive to motion prompts. A common hypothesis in previous works is that the diffusion model is a prompt-driven framework, and a detailed prompt may enhance the quality of the generated results. However, such a feature dramatically limits the practical application for users in the real world. The existing datasets such as WebVid (Bain et al. 2021) and HDVILA (Xue et al. 2022) mainly focus on describing scenes and events in their captions, while ignoring the motion of the objects. Training on such datasets may result in a decrease in the quality of generated motion and insensitivity towards motion-related keywords.

In this paper, we aim to devise a more practical and controllable I2V model that can address such problems. To this end, we propose **Follow-Your-Click**, a novel I2V framework that is capable of regional image animation via a user click and following motion prompts. To achieve this simple user interaction mechanism while obtaining good generation performance, we first simply integrate SAM (Cheng et al. 2023) to convert user clicks to binary regional masks, which serve as one of our network conditions. Then to better learn the temporal correlation correctly, we introduce an effective

*Equal contribution.

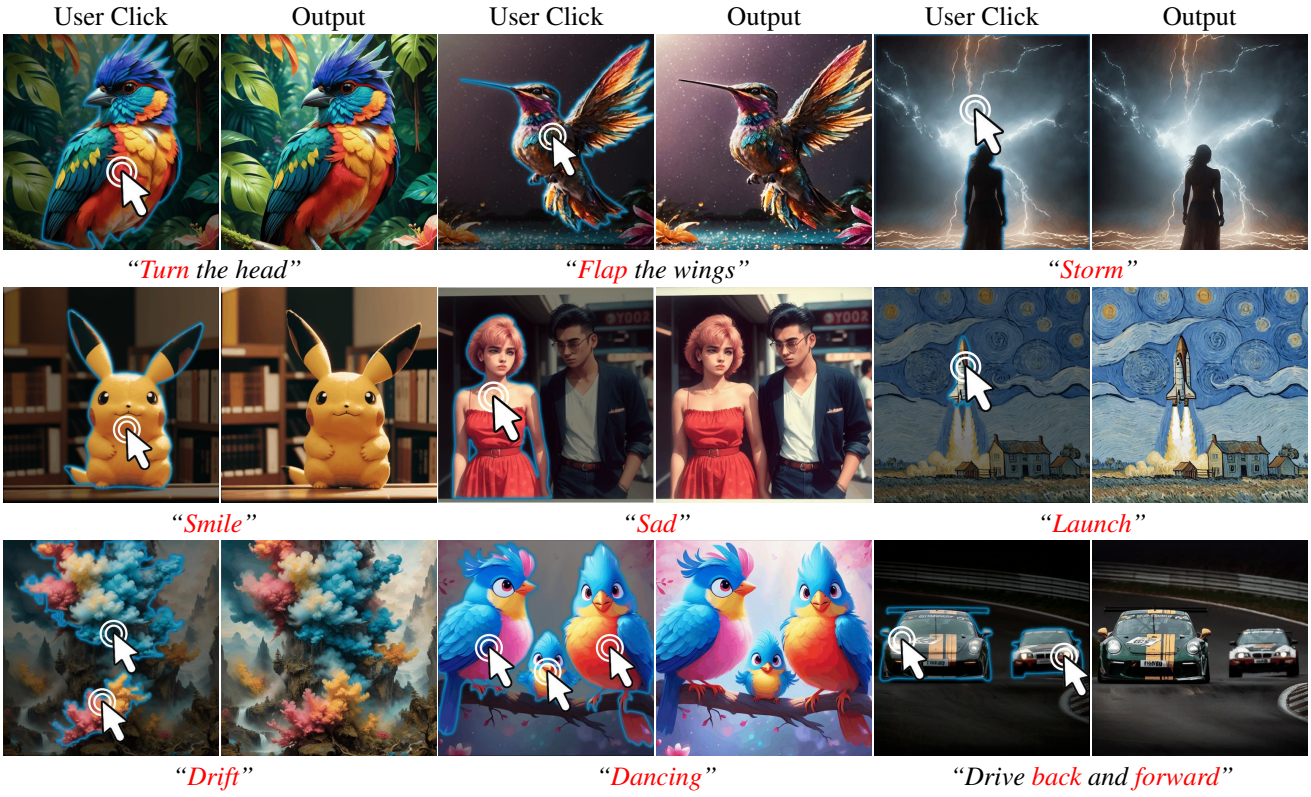

Figure 1: **Regional Image Animation using a Click and a Motion Prompt**. We present a novel framework that facilitates locally aware image animation via a user-provided click (*where* to move) and a motion prompt (*how* to move). Our framework can provide vivid object movement, background movement (e.g., storm), and multiple object movements. *Best viewed with Acrobat Reader*, which supports clicking on the video to play the animations. *Static frames and videos of all results are provided in supplementary materials*.

first-frame masking strategy and observe a large margin of performance gains. To achieve the motion prompt following abilities, we construct a dataset referred to as *WebVid-Motion*, which is built by leveraging a large language model (LLM) for filtering and annotating the video captions, emphasizing human emotion, action, and common motion of objects. We then design a *motion-augmented module* to better adapt to the dataset and enhance the model's response to motion-related words and understand motion prompt instructions. Furthermore, we also observe that different object types may exhibit varied motion speeds. In previous works (Xing et al. 2023), frame rate per second (FPS) primarily serves as a global scaling factor to indirectly adjust the motion speed of multiple objects. However, it cannot effectively control the speed of moving objects. For instance, a video featuring a sculpture may have a high FPS but zero motion speed. To enable accurate learning of motion speed, we propose a novel *flow-based motion magnitude control*.

With our design, we achieve remarkable results on eight various evaluation metrics. Our method can also facilitate the control of *multiple* object and moving types via multiple clicks. Besides, it is easy to integrate our approach with controlling signals, such as human skeletons, to achieve a more fine-grained motion control.

Our contributions can be summarized as follows: (1) To the best of our knowledge, Follow-Your-Click is the first framework supporting a simple *click* and *motion prompt* for regional image animation. (2) To achieve such a user-friendly and controllable I2V framework, technically, we propose the *first-frame masking* to enhance the general generation quality, a *motion-augmented module* with an equipped *motion prompt dataset* for motion prompt following, and a *flow-based motion magnitude* for a more accurate motion speed control. (3) We conducted extensive experiments and user studies to evaluate our approach, which shows our method achieves state-of-the-art performance.

2 Related Work

2.1 Text-to-Video Generation

Text-to-video generation is a popular topic with extensive research in recent years. The emergency of diffusion models (Song, Meng, and Ermon 2020) delivers higher quality and more diverse results. While they provide the potential to control appearance and motion separately, they still face the challenge of video regional control. Even though these models can produce high-quality videos, they mainly rely on textual prompts for semantic guidance, which can be am-

biguous and may not precisely describe users' intentions. To address such a problem, many control signals such as structure (Gao et al. 2023), pose (Ma et al. 2024a; Zhang et al. 2023; Wang et al. 2023a), and Canny edge (Zhang et al. 2023) are applied for controllable video generation. Many recent and concurrent methods in Dynamicrafter (Xing et al. 2023), VideoComposer (Wang et al. 2023b), and I2VGen-XL (i2v 2023) explore RGB images as a condition to guide video synthesis. However, they fail to generate temporally coherent frames and realistic motions while preserving the details of the input image. Most of the prompts are used to describe the image content, users can not animate the image according to their intent. Our approach is based on text-conditioned VDMs and leverages their powerful generation ability to animate the objects in the images while preserving the consistency of background.

2.2 Image Animation

Image-to-video generation involves an important demand: maintaining the identity of the input image while creating a coherent video (Mei, Dong, and Xu 2024; Yu et al. 2024; Shen et al. 2024b; Shen and Tang 2024; Shen et al. 2024a; Feng et al. 2024; Zhu et al. 2024; Wang et al. 2024; Shen et al. 2024d; Chen et al. 2024; Ma et al. 2024c, 2023, 2022a; Shen et al. 2024c; Ma et al. 2024b,a). This presents a significant challenge in striking a balance between preserving the image's identity and the dynamic nature of video generation. Currently, mainstream works based on diffusion (Weng, Curless, and Kemelmacher-Shlizerman 2019; Gao et al. 2024) can generate frames using the video diffusion model. But they only focus on the curated domain and fail to generate temporally coherent real frames. The controllability of these approaches is poor. Additionally, Some commercial large-scale models, Gen-2 (gen 2023a), Genmo (gen 2023b), and Pika Labs (pik 2023) deliver impressive results in the realistic image domain in its November 2023 update. However, these works cannot achieve regional image animation and accurate control. They still face the challenge of synthesizing realistic motion (see Fig. 3) and cannot support the user click and motion prompt interactions. As a commercial tool, Gen-2 will not release technical solutions and checkpoints for research. Furthermore, previous methods (Chen et al. 2023; Dai et al. 2023; Shi et al. 2024) have utilized various motion factors to control motion strength. LivePhoto (Chen et al. 2023) uses structural similarity, Animate-anything (Dai et al. 2023) relies on frame difference, and Motion-I2V (Shi et al. 2024) employs frame stride. Differing from these approaches, we introduce a novel method using optical flow magnitude to more effectively regulate motion intensity. In contrast, our method holds unique advantages in its simple interactions, motion-augmented learning, and better generation quality.

3 Follow-Your-Click

3.1 Problem Formulation

Given a still image, our goal is to animate user-selected regions, creating a short video clip that showcases realistic motion while keeping the rest of the image static. Formally, given an input image $\mathcal{I}$, a point prompt p, and a motion-related verb description of the desired motion t, our approach produces a target animated video $\mathcal{V}$. We decompose this task into several sub-problems including improving the generation quality of local-aware regional animation, achieving motion prompt controlled generation, and motion magnitude controllable generation. Note that the target region is utilized for selecting the animated object rather than limiting the motion of the generated object in subsequent frames. In other words, the object is not constrained to remain within the specified areas and can move outside of them if necessary.

User Interaction and Control. Given an input image that the user wants to animate. An intuitive way is first to choose which part of the image needs to move, then use the text prompt to describe the desired moving pattern. Current approaches, such as research works I2VGen-XL, SVD, dynamicrater, and commercial tools like Pika Lab and Genmo, lack the ability of regional control. The motion brush of Gen-2 (gen 2023a) and animate-anything (Dai et al. 2023) can achieve such a goal but the motion mask needs to be provided or drawn by users, which is not efficient and intuitive for users. Thus, to provide a user-friendly control, we design to use a *point prompt* instead of a binary mask. Furthermore, current image-to-video methods require the input prompt to describe the entire scene and frame content, which is tedious and unnecessary. On the contrary, we simplify this procedure with a motion prompt. To achieve this, we integrate a promptable segmentation tool SAM (Cheng et al. 2023) to convert the point to prompt p to a high-quality object mask $\mathcal{M}$. The masked-controlled regional animation will be introduced in 3.2. To achieve the motion prompt following, we propose a motion-augmented module described in 3.3.

3.2 Regional Image Animation

Optical flow-based motion mask generation. Training on public datasets such as WebVid (Bain et al. 2021) and HDVILA (Xue et al. 2022) directly is challenging to achieve regional image animation due to the lack of corresponding binary mask guidance for regions with large movement. To solve this issue, we utilize the optical flow prediction model to automatically generate the mask indicating the moving regions. Specifically, give training video frames $\{x_0, x_1..., x_{L-1}\}$, we utilize an open-sourced optical flow estimator $\mathcal{E}_{flow}$ (Teed and Deng 2020) to extract the optical flow map $\mathcal{F}_i$ of each two consecutive frame pairs, where i is the frame index of the video. For each flow map $\mathcal{F}_i$, we threshold the map into a binary one $\mathcal{M}_i$ via a threshold calculated via its average magnitude. Finally, we take the union of all masks $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_{L-1}$ to get the final mask $\mathcal{M}_{final}$ to represent area of motion. Formally, the motion area guidance is implemented as

$$\begin{aligned} \mathcal{F}_i &= \mathcal{E}_{flow}(x_i, x_{i-1}), \\ \mathcal{M}_i = \text{Binarize}(\mathcal{F}_i, \text{Avg}(\|\mathcal{F}_i\|)),\ \mathcal{M}_{final} &= \bigcup_{i=0}^{L-1} (\mathcal{M}_i). \end{aligned} \tag{1}$$

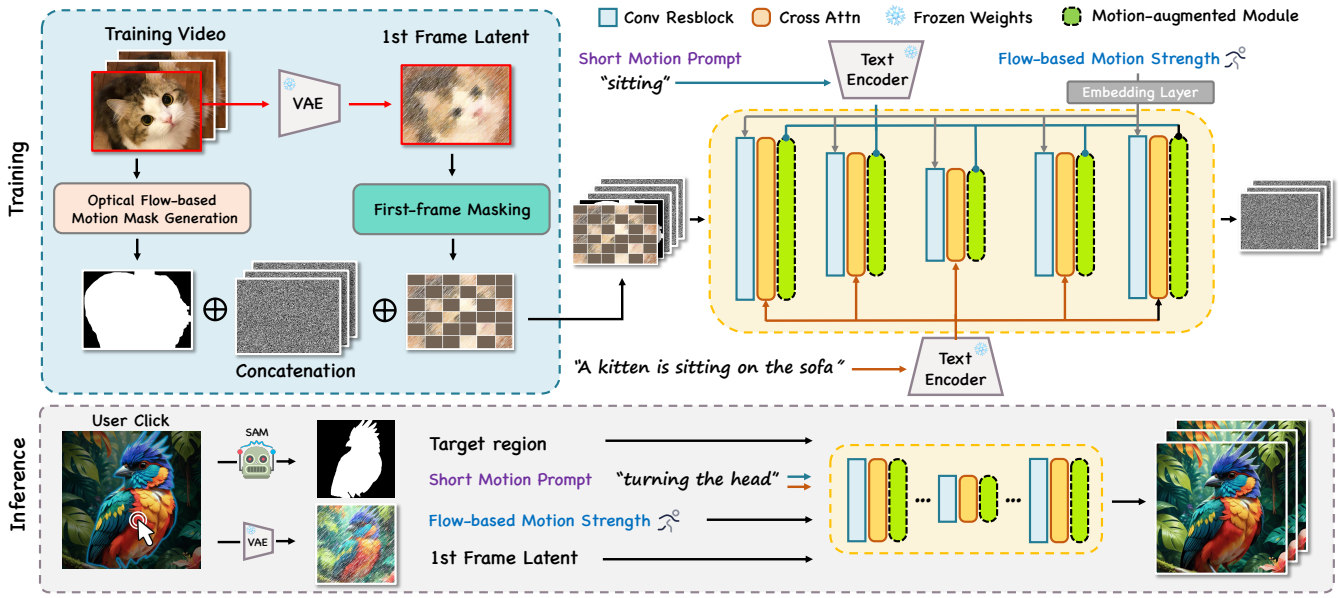

Figure 2: **Framework overview.** The key components of our framework are the first-frame masking, motion-augmented module for motion prompt following, and flow-based motion strength control. During inference, the regional animation can be achieved by user clicks and motion prompts.

where $i = 1, 2, 3, \ldots, L$, Binarize$(\cdot, \cdot)$ is the binarization operation and $\|\cdot\|$ denotes magnitude of optical flow in each pixel. During training, we use $\mathcal{M}_{final}$ to represent the motion area of ground truth videos. During inference, we transfer the user clicks into the binary mask via the promptable image segmentation tool SAM (Cheng et al. 2023) and then feed the binary mask to our network. We also study the generalization ability of conditional masks in supplementary materials.

First-frame masking training. After obtaining the moving region mask $\mathcal{M}_{final}$, we concatenate the downsampled version, the first frame latent z_0, and random noise in the channel dimension in the latent space, obtaining input with size $[9, L, h, w]$ and then fed it into the network. z_0 is the latent of the first frame x_0 which is encoded via the VAE encoder $\mathcal{E}$. The $\mathcal{M}_{final}$ is downsampled to match the resolution of the frame latent. The mask of the target generated frame $\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_{L-1}$ is set to zero, and the first frame serves as guidance and is repeated to L frames. The 9 channels consist of 4 channels of input image latent, 4 channels of the generated frames, and 1 channel of the binary mask. We adopt the $\mathbf{v}$-prediction parameterization proposed in (Salimans and Ho 2022) for training since it has better sampling stability when a few of the inference steps. However, we observe that training directly in this manner exhibits temporal structure distortion issues. Inspired by the recent masked strategy works (He et al. 2022; Feichtenhofer et al. 2022; Ma et al. 2022b), we hypothesize that augmenting the condition information in training can help the model to learn the temporal correlation better. Therefore, we randomly mask the latent embedding of the input image z_0 by a ratio of $\mathcal{R}$, setting the masked region to 0. As shown in Fig. 2, the masked first frame latent, along with the downsampled $\mathcal{M}_{final}$ and noisy video latent $\mathbf{z}$, are concatenated and fed into the network for optimization. Empirically, we discover that randomly masking the input image latent can significantly improve the quality of the generated video clip. In Sec. 4.3, we conduct a detailed analysis of the selection of mask ratio.

3.3 Temporal Motion Control

Motion caption construction. We discover that captions in current extensive datasets always comprise numerous scene descriptive terms alongside fewer dynamic or motion-related descriptions. To enable the achieve better motion prompt following, we construct the WebVid-Motion dataset, a dataset by filtering and re-annotating the WebVid-10M dataset using GPT4 (gpt 2023). In particular, we construct 50 samples to achieve in-context learning of GPT4. Each sample contains the original prompt, objects, and their motion-related descriptions. These samples are fed into GPT4 in JSON format, and then we ask the same question to GPT4 to predict other motion prompts in WebVid-10M. Finally, the re-constructed dataset contains captions and their motion-related phrases, such as "turn the head", "smile", "blink" and "running". We finetune our model on this dataset to obtain a better ability of motion prompt following.

Motion-augmented module. With a trained model via the previous techniques (Guo et al. 2023), to make the network further aware of motion prompts, we design the motion-augmented module to improve the model's responses to motion-related prompts. In detail, we insert a new cross-attention layer in each motion module block. The motion-related phrases are fed into a motion-augmented module for training, and during inference, these phrases are input into both the motion-augmented module and the cross-attention

Method	Automatic Metrics				User Study			
	I_1-MSE↓	Tem-Consis↑	Text-Align↑	FVD ↓	Mask-Corr↓	Motion↓	Appearance↓	Overall ↓
Gen-2 (gen 2023a)	54.72	0.8997	0.6337	496.17	3.12	5.11	2.52	2.91
Genmo (gen 2023b)	91.84	0.8316	0.6158	547.16	6.43	4.57	3.51	3.76
Pika Labs (pik 2023)	**33.27**	**0.9724**	**0.7163**	**337.84**	3.92	**2.86**	**2.17**	**2.88**
Dynamicrafter (Xing et al. 2023)	98.19	0.8341	0.6654	486.37	5.27	6.25	4.91	5.93
I2VGen-XL (i2v 2023)	117.86	0.6479	0.5349	592.13	7.19	7.79	6.98	7.26
SVD (i2v 2023)	43.57	0.9175	0.5007	484.26	4.91	3.74	3.94	4.81
Animate-anything (i2v 2023)	53.72	0.7983	0.6372	477.42	**2.73**	4.73	5.47	5.75
Ours	**21.46**	**0.9613**	**0.7981**	**271.74**	**1.38**	**1.91**	**1.87**	**1.78**

Table 1: **Quantative comparisons between baselines and our approach**. Our method demonstrates the best or comparable performance across multiple metrics. The metrics for the best-performing method are highlighted in red, while those for the second-best method are highlighted in blue.

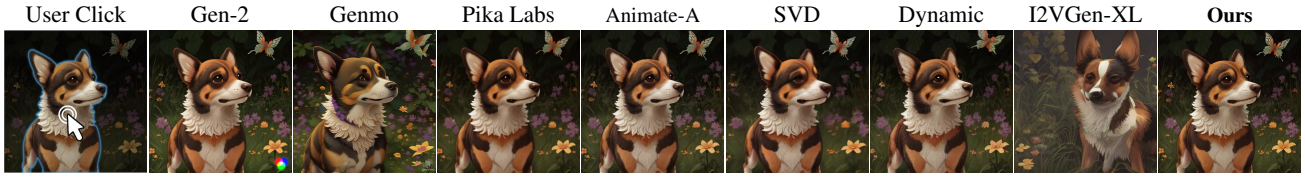

Figure 3: **Qualitative comparisons between baselines and our approach**. We compare with both close-sourced commercial tools including Gen-2 (gen 2023a), Genmo (gen 2023b), and Pika (pik 2023) and research works including Animate-anything (Dai et al. 2023), SVD(Chai et al. 2023), Dynamicrafter(Xing et al. 2023), and I2VGen-XL (i2v 2023). The motion prompt is "shake body". Please click the video to play the animated clips via *Adobe Acrobat Reader*. *Static frames are provided in supplementary materials*.

module in U-Net. Thanks to this module, our model can generate the desired performance during inference with just a motion-related prompt provided by the user, eliminating the need for redundant complete sentences.

Optical flow-based motion strength control. The conventional method for controlling motion strength primarily relies on adjusting frames per second (FPS) and employs the dynamic FPS mechanism during training (Zhou et al. 2022). However, we observe that the relationship between motion strength and FPS is not linear. Due to variations in video shooting styles, there can be a significant disparity between FPS and motion strength. For instance, even in low-FPS videos (where changes occur more *rapidly* than in high-FPS videos), slow-motion videos may exhibit minimal motion. This approach fails to represent the intensity of motion accurately. To address this, we propose using the magnitude of optical flow as a means of controlling the motion strength. As mentioned in 3.2, once we obtain the mask for the area with the most significant motion, we calculate the average magnitude of optical flow within that region. This magnitude is then projected into positional embedding and added to each frame in the residual block, ensuring a consistent application of motion strength across all frames.

4 Experiments

In this section, we introduce our detailed implementation in Sec. 4.1. Then we evaluate our approach with various baselines to comprehensively evaluate our performance in Sec. 4.2. We then ablate our key components to show their effectiveness in Sec. 4.3. Finally, we provide three applications to demonstrate the potential of integrating our approach with other tools in Sec. 4.4.

4.1 Implementation Details

In our experiments, the spatial modules are based on Stable Diffusion (SD) V1.5 (Rombach et al. 2022), and motion modules use the corresponding AnimateDiff (Guo et al. 2023) checkpoint V2. We freeze the SD image autoencoder to encode each video frame to latent representation individually. We train our model for 60k steps on the WebVid-10M (Bain et al. 2021) and then finetune it for 30k steps on the reconstructed WebVid-Motion dataset. More details and evaluation metrics can be found in supplementary materials.

4.2 Comparison with baselines

Qualitative results. We qualitatively compare our approach with the most recent open-sourced state-of-the-art animation methods, including Animate-anything (Dai et al. 2023), SVD (Blattmann et al. 2023a), Dynamicrafter (Xing et al. 2023) and I2VGen-XL (i2v 2023). We also compare our approach with commercial tools such as Gen-2 (gen 2023a), Genmo (gen 2023b), and Pika Labs (pik 2023). Note that the results we accessed on Feb.15th, 2024 might differ from the current product version due to rapid version iterations. Dynamic results can be found in Fig. 3. Given the benchmark images, their corresponding prompts, and selected regions, it can be observed that the videos generated by our approach exhibit better responses to motion-related prompts "`Shake body`". Meanwhile, our approach achieves regional animation while also obtaining better preservation of details from the input image content. In contrast, SVD and Dynamicrafter struggle to produce consistent video frames, as

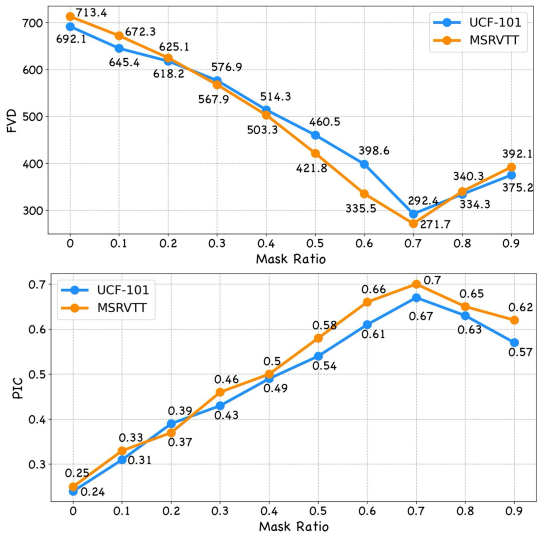

Figure 4: **Ablation study about the masking ratio of the first-frame masking strategy.** Different masking ratios significantly affect the generation quality (FVD) and the perceptual input conformity (PIC) (Xing et al. 2023).

subsequent frames tend to deviate from the initial frame due to inadequate semantic understanding of the input image. I2VGen-XL, on the other hand, generates videos with smooth motion but loses image details. We observe that Genmo is not sensitive to motion prompts and tends to generate videos with small motion. As commercial products, Pika Labs and Gen-2 can produce appealing high-resolution and long-duration videos. However, Gen-2 suffers from the less responsive to the given prompts. Pika Labs tends to generate still videos with less dynamic and exhibits blurriness when attempting to produce larger dynamics. Although both Animate-anything and Gen-2 can achieve regional animation and generate motions as large as those produced by our approach, it suffers from severe distortion and text alignment. These results verify that our approach has superior performance in generating consistent results using motion-related prompts even in the presence of large motion.

Quantitative results. For extensive evaluation, We construct a benchmark for quantitative comparison, which includes 30 prompts, images and corresponding region masks. The images are downloaded from the copyright-free website Pixabay and we use GPT4 to generate prompts for the image content and possible motion. The prompts and images encompass various contents (characters, animals, and landscapes) and styles (*e.g.*, realistic, cartoon style, and Van Gogh style). From Table. 1, It can be observed that our approach achieves the best video-text alignment and temporal consistency against baselines. As for the user study, our approach obtains the best performance in terms of temporal coherence and input conformity compared to commercial products, while exhibiting superior motion quality.

Figure 5: **Visual results of different masking ratios.** Training without masking presents poor movement, temporal consistency and video quality. The prompt is "driving".

Figure 6: **Qualitative results of ablation the constructed motion prompt dataset (D) and motion-augmented module (M)**. The motion prompt is "running".

4.3 Ablation Study

Input image mask ratio. To investigate the influence of the first frame masking strategy and different mask ratios for the input image in training, we conduct quantitative experiments varying the mask ratio from 0 to 0.9. Following (Xing et al. 2023; Blattmann et al. 2023b), we evaluate the generation performance of all the methods on UCF-101 (Soomro, Zamir, and Shah 2012) and MSRVTT (Xu et al. 2016). The Frechet Video Distance (FVD) (Unterthiner et al. 2019) and Perceptual Input Conformity (PIC) (Unterthiner et al. 2019) are reported to further assess the perceptual consistency between the input image and the animation results. The PIC can be calculated by $\frac{1}{L}\sum_{i=0}^{L-1}(1 - D(\mathcal{I}, x_i))$, where $\mathcal{I}, x_i, L$ are input image, video frames, and video length, respectively. $D(\cdot, \cdot)$ denotes perceptual distance metric DreamSim (Fu et al. 2023). We measure these metrics at the resolution of 256 $\times$ 256 with 16 frames. As shown in Fig. 4, the optimal ratio is surprisingly high. The ratio of 70% obtains the best performance in two metrics. An extremely high mask ratio leads to a decrease in the quality of the generated video due to the weak condition of the input image. Also, we compare the visual results of training without first-frame masking and with the optimal masking ratio in Fig. 5. From the results, we can observe that, without the first-frame masking training, the model fails to learn the correct temporal motion and presents incorrect structures.

4.4 Application

Motion-augmented module. To investigate the roles of our dataset and motion-augmented (MA) module, we examine two variants: 1) **Ours w/o D+M**, we apply the basic motion module designed in AnimateDiff (He et al. 2023) and finetune the model on WebVid-10M. 2) **Ours w/o D**, during training stage, we only use public WebVid-10M to optimize the proposed method. The input of MA module is the original prompt from WebVid-10M. 3) **Ours w/o M**, by

Method	Automatic Metrics				User Study			
	I_1-MSE↓	Tem-Consis↑	Text-Align↑	FVD ↓	Mask-Corr↓	Motion↓	Appearance↓	Overall ↓
w/o Data & MA	35.72	0.8465	0.3659	698.21	2.92	3.27	3.34	3.18
w/o MA	**26.46**	**0.9178**	**0.6294**	**391.47**	**1.97**	**2.17**	**2.08**	**2.24**
w/o Data	29.18	0.8824	0.4356	562.33	2.46	2.38	2.35	2.79
Ours	**21.46**	**0.9613**	**0.7981**	**271.74**	**1.43**	**1.59**	**1.17**	**1.31**

Table 2: **Quantitative ablation results of the motion-augmented module (MA) and our constructed motion prompt dataset (Data)**. The best-performing methods are highlighted in red, and the second-best methods are highlighted in blue.

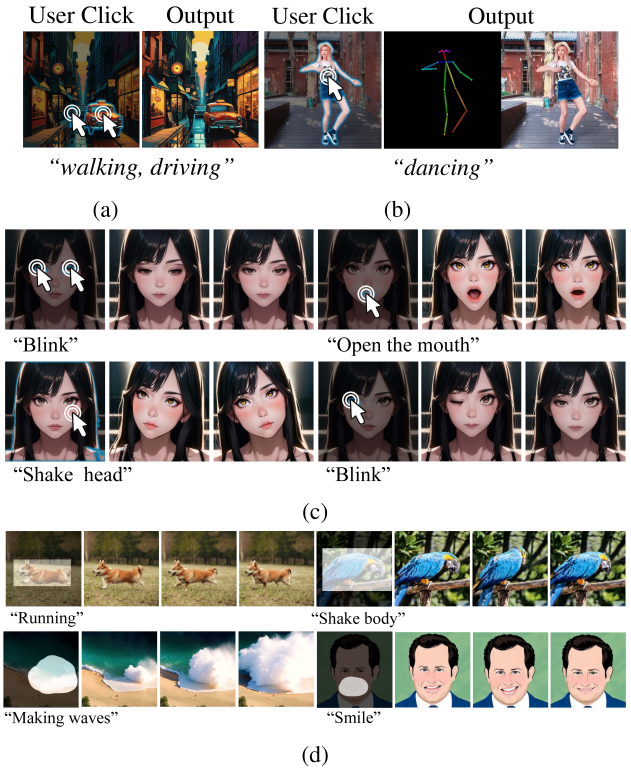

Figure 7: **The Application of our approach**. Our approach can support multiple regions animation in (a) as well as precise motion control in (b) such as human pose. Additionally, our method handles arbitrary regions well, including arbitrary positions in (c) and various mask shapes in (d).

removing the MA module. The motion-related prompts are fed into cross-attention in the spatial module. We also conduct the qualitative comparison in Fig. 6. The performance of "Ours w/o D+M" declines significantly due to its inability to semantically comprehend the input image without a motion prompt, leading to small motion in the generated videos (see the 2nd column). When we remove the MA module, it exhibits limited motion magnitude. We report the quantitative ablation study of the designed module in Table. 2 and the same setting as Sec. 4.2 is applied to evaluate the performance comprehensively. Eliminating Webvid-Motion finetuning leads to a significant decrease in the FVD and text alignment. In contrast, our full method effectively achieves regional image animation with natural motion and coherent frames.

Multi-regions image animation. Using the technology of regional prompter (reg 2023), we can achieve multi-region image animation by different motion prompts. As shown on the left one in Fig. 7a, we can animate the man and car using "walking, driving", respectively. The background of the video is stable, and only selected objects are animated.

Regional image animation with ControlNet (Zhang, Rao, and Agrawala 2023). In addition, our framework can be combined with ControlNet for conditional regional image animation. In the case on the right side of Fig. 7b, we present the use of pose conditioning for conditional generation. It shows that we generate pose-aligned characters with good temporal consistency while maintaining the stability of the background. The difference between this application with previous works (Xu et al. 2024; Hu 2024) is that our model can handle open-domain contents and backgrounds.

Arbitrary regions image animation. Our model is capable of using arbitrary regions as input prompts. As illustrated in Fig. 7c, user clicks at any position can trigger animations. This enables the animation of various areas, such as the mouth or eye, and even the entire head. In addition to user clicks, various mask shapes, including bounding boxes and brushes, are supported, as shown in Fig. 7d.

5 Discussion

We present Follow-Your-Click to tackle the problem of generating controllable and local animation. To the best of our knowledge, we are the first I2V framework that is capable of regional image animation via a simple *click* and a *motion-related prompt*. To support this, the promptable segmentation tool SAM is firstly incorporated into our framework for a user-friendly interaction. To achieve the motion prompt following abilities, we propose a motion-augmented module and a constructed motion prompt dataset to achieve this goal. To improve the generated temporal motion quality, we propose the first-frame masking strategy which significantly improves the generation performance. To enable accurate learning of motion speed, we leverage the optical flow score to control the magnitude of motion accurately. Our experimental results highlight the effectiveness and superiority of our approach compared to existing baselines.

Acknowledgments

This project was supported by the National Key R&D Program of China under grant number 2022ZD0161501.

References

2023. ChatGPT-4. https://chat.openai.com.

2023a. Gen-2. https://runwayml.com/ai-magic-tools/gen-2.

2023b. Genmo. https://www.genmo.ai.

2023. I2VGen-XL. https://modelscope.cn/models/damo/Image-to-Video/summary.

2023. Pika Labs. https://www.pika.art.

2023. Regional Prompter. https://github.com/hako-mikan/sd-webui-regional-prompter.

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1728–1738.

Bertiche, H.; Mitra, N. J.; Kulkarni, K.; Huang, C.-H. P.; Wang, T. Y.; Madadi, M.; Escalera, S.; and Ceylan, D. 2023. Blowing in the Wind: CycleNet for Human Cinemagraphs from Still Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 459–468.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Chai, W.; Guo, X.; Wang, G.; and Lu, Y. 2023. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23040–23050.

Chen, Q.; Ma, Y.; Wang, H.; Yuan, J.; Zhao, W.; Tian, Q.; Wang, H.; Min, S.; Chen, Q.; and Liu, W. 2024. Follow-Your-Canvas: Higher-Resolution Video Outpainting with Extensive Content Generation. *arXiv preprint arXiv:2409.01055*.

Chen, X.; Liu, Z.; Chen, M.; Feng, Y.; Liu, Y.; Shen, Y.; and Zhao, H. 2023. LivePhoto: Real Image Animation with Text-guided Motion Control. *arXiv preprint arXiv:2312.02928*.

Cheng, C.-C.; Chen, H.-Y.; and Chiu, W.-C. 2020. Time flies: Animating a still image with time-lapse video as reference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5641–5650.

Cheng, Y.; Li, L.; Xu, Y.; Li, X.; Yang, Z.; Wang, W.; and Yang, Y. 2023. Segment and Track Anything. *arXiv preprint arXiv:2305.06558*.

Dai, Z.; Zhang, Z.; Yao, Y.; Qiu, B.; Zhu, S.; Qin, L.; and Wang, W. 2023. AnimateAnything: Fine-Grained Open Domain Image Animation with Motion Guidance. *arXiv e-prints*, arXiv–2311.

Feichtenhofer, C.; Li, Y.; He, K.; et al. 2022. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35: 35946–35958.

Feng, K.; Ma, Y.; Wang, B.; Qi, C.; Chen, H.; Chen, Q.; and Wang, Z. 2024. Dit4edit: Diffusion transformer for image editing. *arXiv preprint arXiv:2411.03286*.

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. *arXiv preprint arXiv:2306.09344*.

Gao, K.; Bai, J.; Wu, B.; Ya, M.; and Xia, S.-T. 2023. Imperceptible and robust backdoor attack in 3d point cloud. *IEEE Transactions on Information Forensics and Security*, 19: 1267–1282.

Gao, K.; Bai, Y.; Gu, J.; Xia, S.-T.; Torr, P.; Li, Z.; and Liu, W. 2024. Inducing High Energy-Latency of Large Vision-Language Models with Verbose Images. In *ICLR*.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, Y.; Xia, M.; Chen, H.; Cun, X.; Gong, Y.; Xing, J.; Zhang, Y.; Wang, X.; Weng, C.; Shan, Y.; et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Li, Z.; Tucker, R.; Snavely, N.; and Holynski, A. 2023. Generative image dynamics. *arXiv preprint arXiv:2309.07906*.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024a. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Qi, C.; Cai, C.; Li, X.; Li, Z.; Shum, H.-Y.; Liu, W.; et al. 2024b. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. *arXiv preprint arXiv:2403.08268*.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024c. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*.

Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022a. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.

Ma, Y.; Yang, T.; Shan, Y.; and Li, X. 2022b. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*.

Mei, H.; Dong, M.; and Xu, C. 2024. Efficient Image-to-Image Diffusion Classifier for Adversarial Robustness. *arXiv preprint arXiv:2408.08502*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Salimans, T.; and Ho, J. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; and Tang, J. 2024a. Imagdressing-v1: Customizable virtual dressing. *arXiv preprint arXiv:2407.12705*.

Shen, F.; and Tang, J. 2024. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Shen, L.; Hao, T.; Zhao, S.; Zhang, Y.; Liu, P.; Bao, Y.; and Ding, G. 2024b. Tempme: Video temporal token merging for efficient text-video retrieval. *arXiv preprint arXiv:2409.01156*.

Shen, L.; He, T.; Zhao, S.; Shen, Z.; Guo, Y.; Xu, T.; and Ding, G. 2024c. X-reid: Cross-instance transformer for identity-level person re-identification. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Shen, L.; Zhao, S.; Zhang, Y.; Chen, H.; Zhou, J.; Liu, P.; Bao, Y.; and Ding, G. 2024d. Multi-Label Learning with Block Diagonal Labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4832–4840.

Shi, X.; Huang, Z.; Wang, F.-Y.; Bian, W.; Li, D.; Zhang, Y.; Zhang, M.; Cheung, K. C.; See, S.; Qin, H.; et al. 2024. Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling. *arXiv preprint arXiv:2401.15977*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 402–419. Springer.

Unterthiner, T.; van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2019. FVD: A new metric for video generation.

Wang, F.-Y.; Chen, W.; Song, G.; Ye, H.-J.; Liu, Y.; and Li, H. 2023a. Gen-L-Video: Multi-Text to Long Video Generation via Temporal Co-Denoising. *arXiv preprint arXiv:2305.18264*.

Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024. COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. *arXiv preprint arXiv:2406.08850*.

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023b. VideoComposer: Compositional Video Synthesis with Motion Controllability. *arXiv preprint arXiv:2306.02018*.

Wang, Y.; Yang, D.; Bremond, F.; and Dantcheva, A. 2022. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*.

Weng, C.-Y.; Curless, B.; and Kemelmacher-Shlizerman, I. 2019. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5908–5917.

Xiao, W.; Liu, W.; Wang, Y.; Ghanem, B.; and Li, B. 2023. Automatic animation of hair blowing in still portrait photos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22963–22975.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Wang, X.; Wong, T.-T.; and Shan, Y. 2023. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1481–1490.

Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; and Guo, B. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5036–5045.

Yu, W.; Feng, C.; Tang, J.; Jia, X.; Yuan, L.; and Tian, Y. 2024. EvaGaussians: Event Stream Assisted Gaussian Splatting from Blurry Images. *arXiv preprint arXiv:2405.20224*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077*.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.

Zhu, C.; Li, K.; Ma, Y.; Tang, L.; Fang, C.; Chen, C.; Chen, Q.; and Li, X. 2024. InstantSwap: Fast Customized Concept Swapping across Sharp Shape Differences. *arXiv preprint arXiv:2412.01197*.