# SAM-IAM: Segmenting Arbitrary Motion for Image Applications Model

Dominic Chui[†] Brown University
Hancheng Lin[†] Brown University
Gabrielle Shieh[†] Brown University
Brian Xu[†] Brown University

Figure 1. Example output from our method.

## Abstract

*We present a novel modular video synthesis model for transfering the rigid motion from a driving video to a single image. SAM-IAM generates a new video by combining the subject of the input image and the motion of the input video. In constrast with diffusion models that are text, image, and even video conditioned, this offers an alternative generative approach with much creative potential. By first segmenting the driving video, the motion of the selected subject is isolated and approximated using tracking bounding boxes that can then be iteratively applied to the segmented subject of the input image. The stitching together and conversion of the segmentation masks is performed with an off-the-shelf diffusion model to generate temporally and spatially coherent videos. SAM-IAM shows great potential in generating realistic video results in a modular pipeline whose individual components can be iteratively improved upon.*

## 1. Introduction

Creation is not a process that exists in a vacuum but instead relies on and is in conversation with the environment of the artist. Positive inspiration (mimicry and adoption of a similar style) and negative inspiration (contrast with and intentional movement against an established norm) are both integral parts of the process. Although historically the creative process has relied on a high degree of technical knowledge, the development of increasingly sophisticated assistive creative tools has "democratized" the process by offloading the technical and leaving the creative with the user, who increasingly no longer has to be an artist in the traditional sense. Notably, these tools have been focused on positive inspiration, such as with image analogies and style transfers in general [6].

The rise of neural networks has merely accelerated the growth in this domain with the recent breakthrough of image diffusion for high quality image synthesis [7]. The extension into the time domain has resulted in video diffusion models for video generation [8].

We propose SAM-IAM, a continuation in this rich tradition of visual media generation, as video synthesis conditioned on an input image and applying motion transfer from an input video. Fundamentally the difference between a static image and a dynamic video is the change in (typically) the subject over time. Extracting the intrinsics of this change and applying it to a static image is in a sense similar to style transfers and analogies, but differs in that the subject is being switched. The process is similar to video diffusion conditioned on an image and text, but is notably more constrained by the input.

However, the problem is still fundamentally underconstrained and "motion" itself is contextually dependent on both the input and the application. A video of car driving from left to right has the basic left to right motion, but also contains the rotational motion of the wheels. Applying this motion to an image of a motorcycle should preserve the overall left to right motion but also ideally map across the rotation of the wheel. In contrast, applying this motion to an image of a horse should not preserve any rotational

[†]Equal contribution

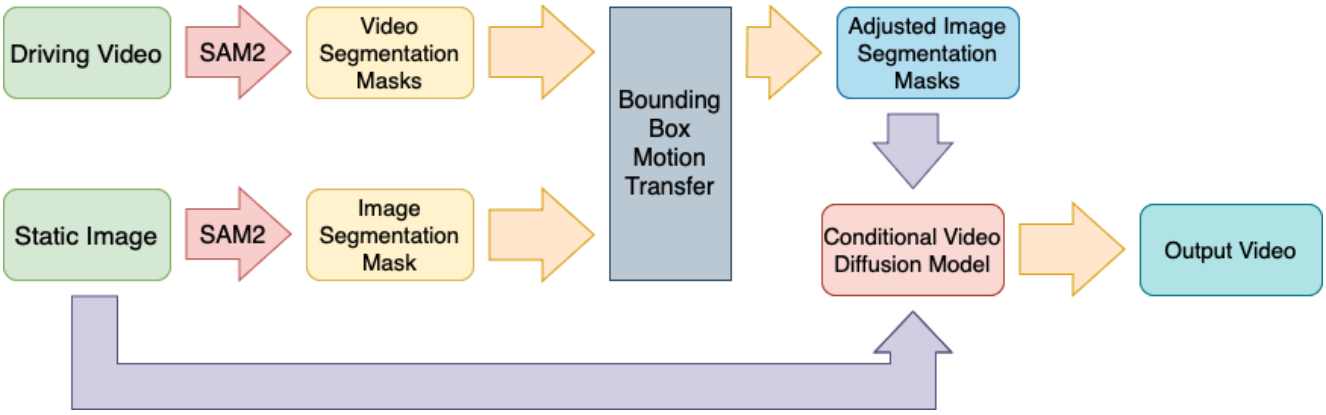


Figure 2. An overview of our pipeline.

motion. Consequently, "reasonable" motion transfer becomes highly semantically dependent.

Even ignoring the mapping problems, from a single view, different types of rigid motion are often indistinguishable without context. An object scaling up in size is equivalent to it translating towards the camera. In order to reduce the complexity of the decision space, our work focuses on rigid motion of a single constant foreground object. This careful constraining of the problem allows us to offer a new focus into the overall problem of video synthesis and a novel targeting of scope.

## 2. Related works

Much related work in video synthesis have tended to fall into either category of image-to-video methods or video diffusion models.

### 2.1. Image-to-Video Methods

Image-to-video methods exist in two flavors: diffusion and non-diffusion based. Periodic patterns have been manipulated to generate seamlessly animated *endless loops* [5] in a non-machine learning approach, while neural network architecture has been employed in a *conditional invertible neural network (cINN)* architecture [2].

Diffusion based image-to-video methods rely on the remarkable performance of diffusion models in image synthesis by extending the existing text-to-video models with temporal-consistency attention layers for image-to-video synthesis and larger training datasets [1]. This approach has seen applications in character animation [9] and fashion posing [10]. ControlNet [21] style approaches extended to video have also been succesfully built [22].

### 2.2. Video Diffusion

With the success of text-to-image and image-to-video models, diffusion models have been successfully developed for training directly on videos themselves in the video-to-video mode. The major obstacle of temporal consistency and coherence has been tackled in a variety of different ways, typically relying on a pretrained diffusion model supplemented with carefully augmented attention layers and conditioning [12]. Controllable video synthesis that allows for user guidance in textual, spatial, and temporal forms have been explored by using motion vectors and spatio-temporal condition encoders [18], propagating and injecting condition features through condition adapters [17], and explicit motion modeling [16].

### 2.3. Conditioning Object Motion

Conditioning object motion can be viewed as a form of spatio-temporal guidance in video generation. Diffusion-based video generative models supporting motion control can be categorized based on the specificity and type of input they use.

At one end of the spectrum, detailed spatial conditioning sequences are used as frame-by-frame guidance. These spatial constraints, in the form of optical flow [13], motion vector [19], pose [4], depth [4, 19], and sketch outline [19], enable precise control over the generated video. Focusing on specific applications, specialized frameworks have been proposed to customize human videos [4]. When applied to objects of a different kind, these methods often yield results akin to style transfers. The strong spatial constraint sequences used by these methods limits their capability in adaptively transferring motion properties between different object types.

On the opposite end, more abstract representations, such as motion flows represented by arrows [19, 20] and vectors [19] or motion themes like 'inflate,' 'squish,' or 'crumble' [15], are used as motion guidance. These methods offer greater generalization across various object classes while compromising on the control over specific motion details and relative frame positioning.

This paper proposes a modular video generation pipeline with a bounding-box driven motion transfer implementation. The method seeks to explore a midpoint in the spectrum of motion control, offering an alternative perspective

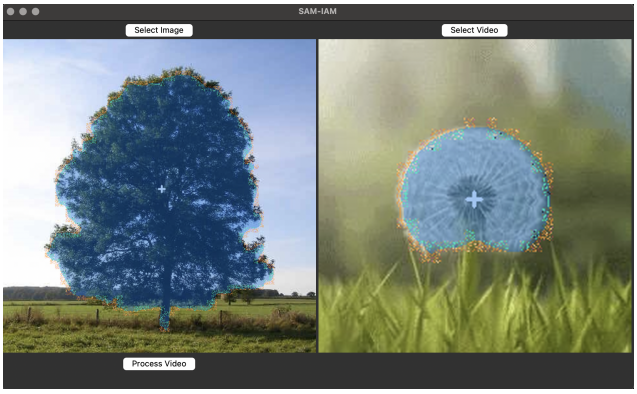


Figure 3. A screenshot of our segmentation UI.

and demonstrating its generative potential. The goal is to retain control over relative frame positions and overall motion from the driving video, rather than requiring detailed, instance-aligned spatial conditioning sequences. This approach aims to enable more flexible and realistic adaptations of motion characteristics in video generation.

## 3. Method

We provide an overview of our method in Fig. 2. Our inputs are a video (also referred to as the "driving video") and a static image. We apply an off-the-shelf image and video segmentation model[14] to obtain binary segmentation masks of the foreground object. For the video, we obtain a segmentation mask for each frame, tracking the relative motion of the masked object. We use a bounding box approach to calculate the change in position and scale for each frame in the video mask. Given this per-frame deformation, we can sequentially modify the segmentation mask of the input image to replicate motion analogous to the driving video. Finally, we use the motion-transferred video along with the original image as inputs for a conditional video diffusion model[19].

### 3.1. Image and Video Segmentation

We use an off-the-shelf segmentation model[14] for both the image and the video. Due to the importance of obtaining proper segmentation masks, we elect to involve the user during this step. We present the user a GUI (see Fig. 3), allowing them to interactively select which foreground objects are segmented.

### 3.2. Bounding Box Motion Transfer

Given a binary segmentation mask, it is trivial to draw a bounding box around it. We can obtain a rough estimate for the relative motion of a video by calculating the change in position and scale between each frame. For each frame in the driving video, we obtain a translation $t$ and change in scale $s$ of its corresponding bounding box, relative to the first frame.

$$t_n = (\text{bbox}_{\text{vid}_n}.x - \text{bbox}_{\text{vid}_0}.x, \text{bbox}_{\text{vid}_n}.y - \text{bbox}_{\text{vid}_0}.y)$$
$$s_n = (\frac{\text{bbox}_{\text{vid}_n}.width}{\text{bbox}_{\text{vid}_0}.width}, \frac{\text{bbox}_{\text{vid}_n}.height}{\text{bbox}_{\text{vid}_0}.height}) \tag{1}$$

To preserve the direction of motion while accounting for possible differences in size, we introduce scaling factors $x$ and $y$. These scaling factors are equal to the height and width ratios of the bounding boxes for the input image and first frame of the driving video.

$$x = \frac{\text{bbox}_{\text{img}}.width}{\text{bbox}_{\text{vid}}.width}$$
$$y = \frac{\text{bbox}_{\text{img}}.height}{\text{bbox}_{\text{vid}}.height} \tag{2}$$

We sequentially apply the adjusted set of translations and scales to the mask of the input image to obtain a target video $V$, which we use as input for a conditional diffusion model[19].

$$V_n.size = \text{bbox}_{\text{img}}.size * s_n * (x, y)$$
$$V_n.pos = \text{bbox}_{\text{img}}.pos + t_n * (x, y)$$
$$V_n.pos = V_n.pos - \frac{(V_n.size - \text{bbox}_n.size)}{2} \tag{3}$$

## 4. Experiments

We evaluated our method on several image-video pairs. For a baseline comparison, we used the same inputs on our off-the-shelf model to provide a fair comparison (Tab. 1).

## 5. Discussion

### 5.1. Results

Due to the how novel our approach is, coming up with reasonable baselines and comparisons is non-trivial. We demonstrate the performance of our approach against our baseline, VideoComposer, a controllable video diffusion model. In Table 1, we demonstrate our results through a side-by-side comparison with VideoComposer. As shown, the existing baseline fails to capture the intrinsics of the subject's change in motion, and even in some cases fails to comprehend the motion dynamics of the original driving video. We see in parachute example that SAM-IAM captures the shape of the parachute throughout the length of the projected motion, whereas VideoComposer suffers from shape deformation. Furthermore, the soccer ball example demonstrates the ability of SAM-IAM to comprehend more complex motions as the driving video illustrates a car traveling across the frame from right to left, with the addition of forward motion towards the camera in the beginning frames. However, VideoComposer fails to correctly

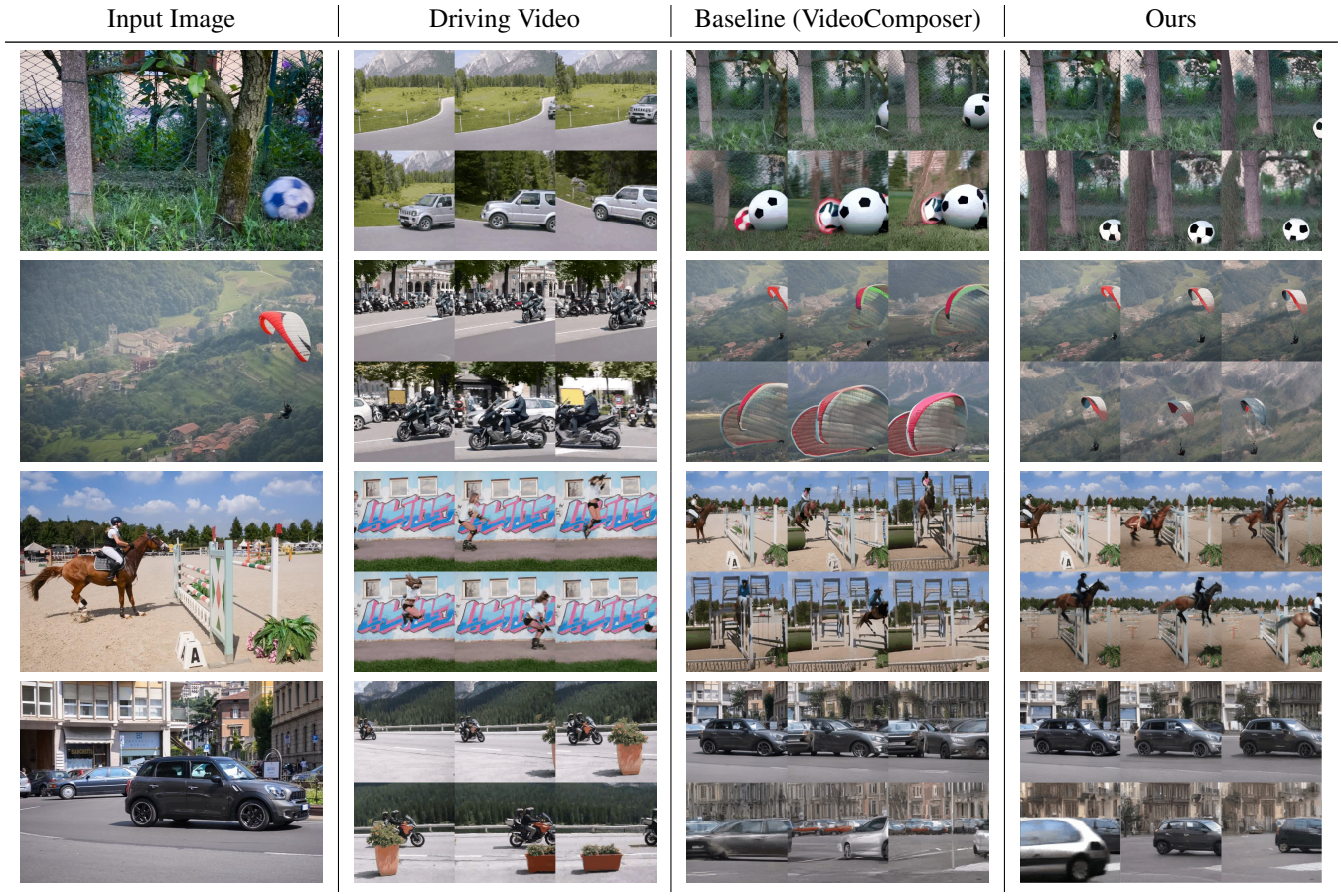


Table 1. A side-by-side comparison of results from our method and the baseline.

capture the scale of the input image while adding additional artifacts to the resulting video.

In the equestrian example, we qualitatively compare SAM-IAM and VideoComposer on a driving video that exhibits human motion in the form of jumping. While VideoComposer adds artifacts to the foreground and background of the frame and incorrectly adds rotation to the target object's movement, SAM-IAM cleanly translates the movement of the jumping rollerblader to the horse in the same arcing motion.

In the case of occlusion, SAM-IAM achieves an accurate representation by adding another moving vehicle to obstruct the view of the car, analogous to the way the plant occludes the motorcycle in the driving video. In contrast, VideoComposer exhibits multiple colliding vehicles to its resulting video, demonstrating a lack of understanding of occlusion.

### 5.2. Method Limitations

We initially decided on an autoencoder to understand the motion demonstrated in the driving video. The autoencoder is passed the input image and frames $n$ and $n + 1$ of the driving video, and creates a difference map from the video frames to pass into the encoder. The encoder outputs the latent code, which, with the input frame, would be passed into the decoder to reconstruct the predicted next image frame. However, after training and testing this model on the DAVIS dataset, we found that the autoencoder was greatly overfitting because transferring motion across different classes of object requires a deep understanding of their geometric properties and our frame-by-frame approach led to unstable and inconsistent motions. We theorize these shortcomings could be alleviated with a stronger dataset containing paired data for the type of 'loose' motion transfer we want and a loss function that captures the intricacies of the underlying motion rather than basic change in position across frames. Due to the modal nature of our pipeline, we were easily able to pivot and replace the autoencoder with the bounding box method outlined in our paper to track the motion in the driving video.

## 6. Conclusion

We present SAM-IAM, a novel video synthesis model conditioned on an input image with the application of motion transfer from a given driving video to a single constant foreground object. SAM-IAM achieves a deep understanding of rigid motion transfer as a result of applying motion translation through bounding boxes to segmentation masks, including in cases of occlusion. This approach significantly outperforms the applicable baselines, constructing accurate and clear novel videos analogous to its respective inputs.

Future research could focus on several aspects to enhance the versatility of video-driven motion control in video generation. Refining the granularity of segmentation in motion objects extracted from driving videos could improve the quality of motion guidance used to condition the output video in our pipeline. Additionally, integrating a feature descriptor diffusion model [3, 11] could provide a more semantic mapping of motion details between objects, thereby strengthening the adaptability of motion transfers.

Another direction could involve adopting more sophisticated generative models, such as Variational Autoencoders (VAEs) and diffusion models, for the motion extraction and transfer phases. This approach would require collecting extensive paired data and might be susceptible to inherent biases, such as viewpoint bias, which needs to be considered.

## References


[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 2

[2] Michael Dorkenwald, Timo Milbich, Andreas Blattmann, Robin Rombach, Konstantinos G. Derpanis, and Björn Ommer. Stochastic image-to-video synthesis using cinns, 2021. 2

[3] Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J. Mitra. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4494–4504, 2024. 5

[4] Mengyang Feng, Jinlin Liu, Kai Yu, Yuan Yao, Zheng Hui, Xiefan Guo, Xianhui Lin, Haolan Xue, Chen Shi, Xiaowen Li, Aojie Li, Xiaoyang Kang, Biwen Lei, Miaomiao Cui, Peiran Ren, and Xuansong Xie. Dreamoving: A human video generation framework based on diffusion models. *arXiv*, 2023. 2

[5] Tavi Halperin, Hanit Hakim, Orestis Vantzos, Gershon Hochman, Netai Benaim, Lior Sassy, Michael Kupchik, Ofir Bibi, and Ohad Fried. Endless loops: detecting and animating periodic patterns in still images. *ACM Transactions on Graphics*, 40(4):1–12, 2021. 2

[6] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. page 327–340, 2001. 1

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1

[8] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 1

[9] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation, 2024. 2

[10] Johanna Karras, Aleksander Hołyński, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. 2023. 2

[11] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 5

[12] Andrew Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge Ritter. Video diffusion models: A survey, 2024. 2

[13] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18444–18455, 2023. 2

[14] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[15] Eric Hal Schwartz. This ai video generator can melt, crush, blow up, or turn anything into cake, 2024. 2

[16] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling, 2024. 2

[17] Cong Wang, Jiaxi Gu, Panwen Hu, Haoyu Zhao, Yuanfan Guo, Jianhua Han, Hang Xu, and Xiaodan Liang. Easycontrol: Transfer controlnet to video diffusion for controllable generation and interpolation, 2024. 2

[18] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability, 2023. 2

[19] Wang, Xiang and Yuan, Hangjie and Zhang, Shiwei and Chen, Dayou and Wang, Jiuniu, and Zhang, Yingya, and Shen, Yujun, and Zhao, Deli and Zhou, Jingren. Videocomposer: Compositional video synthesis with motion controllability. *arXiv preprint arXiv:2306.02018*, 2023. 2, 3

[20] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

[21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 2

[22] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation, 2023. 2