Iterative Optimization of SDS Loss for Video Generation from Multi-Object Sketches

Donghun Kim **Changho Choi** **Junmo Kim**
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
{actruce, ccho4702, junmo.kim}@kaist.ac.kr

Abstract

In this paper, we present a novel framework for generating videos with realistic movements and natural interactions between multiple objects derived from a single sketch image, utilizing text-to-video diffusion priors. We identify limitations in existing sketch-to-video methods, particularly their challenges in managing multiple objects and accurately representing the independent movements and interactions of separable objects. Our approach aims to facilitate the independent movement of objects while maintaining interaction consistency. To achieve this, we manually separated a vector image into distinct components and trained neural networks for each object using text-to-video diffusion with SDS loss. The optimization process was conducted iteratively: initially, the separated objects were trained individually, followed by joint training of the combined image. As a result, our method effectively produces realistic movements for separable objects while preserving the coherence of their interactions.

1 Introduction

Animated sketches are especially powerful tools for storytelling; however, the traditional process of animating sketches demands substantial manual effort. To alleviate this labor-intensive requirement, several research endeavors have been pursued [1] [2] [3]. Nonetheless, the majority of these approaches are primarily focused on single-object scenarios. We propose a novel framework that extends existing architectures for text-to-video generation in multi-object scenarios. Our approach leverages an image-to-sketch model for sketch generation, followed by iterative optimization using score distillation sampling (SDS) [4] with a pre-trained text-to-video diffusion model. This framework transitions from single-object environments to dynamic scenes where multiple entities interact.

2 Preliminaries

Vector representation Vector images are digital graphics constructed from paths defined by mathematical equations. Unlike raster images, which consist of a fixed grid of pixels, vector images use geometric primitives. This characteristic allows vector images to be resized without degradation in quality, making them particularly advantageous for various applications. Scalable Vector Graphics (SVG) is a prominent example of a vector image format.

The input vector image is composed of strokes on a white background, where each stroke is a two-dimensional cubic Bézier curve with four control points. Each control point is represented by its coordinates: $p = (x, y) \in \mathbb{R}^2$. Also, the set of control points in a single frame is represented as $P = \{p_1, \ldots, p_N\} \in \mathbb{R}^{N \times 2}$, where N is the total number of points in the input sketch. A video with k frames is defined as a sequence of k such sets of control points, denoted by $Z = \{P^j\}_{j=1}^{k} \in \mathbb{R}^{N \cdot k \times 2}$.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

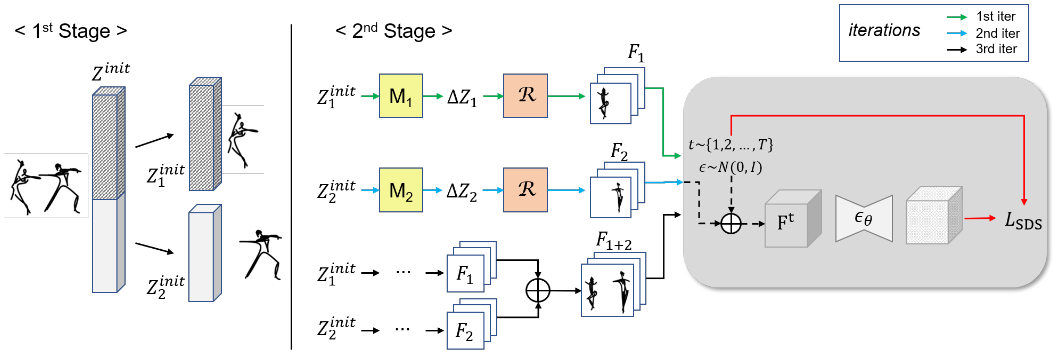


Figure 1: Training scheme. In the 1st stage, we distinguish which stroke belongs to which object. Then, three different iterations with different combinations of objects are performed in the 2nd stage.

3 Method

Our core optimization procedure is based on the approach proposed by [1], which adopt local network to estimate object's local motion and global network to predict a per-frame transformation like scaling, rotation, and translation. [1] use text-to-video diffusion model similar to [5]. Although they can transform a single object frame by frame, they still face difficulties in handling multiple objects.

To address this limitation, we propose a novel method in which the SDS loss is designed to enable gradient flows targeted to each individual object within an image. By directing gradients to flow selectively within specific portions of the sketch, independent motion can be achieved in different parts of the overall image. The complete method comprises two main phases, as illustrated in Figure1.

In the first phase, we generate a vector representation of the image and decompose it into multiple sub-images, each containing only a single object from the original composition. In the second phase, we optimize a neural network to predict both local and global displacements for each object, utilizing the SDS loss in conjunction with a text-to-video prior.

Vector Image Generation and Stroke Assignment In the first stage, Adobe Firefly is used to generate raster images with text prompts for multi-object scenarios. Then, CLIPasso [6] is used to convert these raster images into vector images of SVG format. After generating the SVG vector images, the object to which each stroke corresponds is identified. These tasks are done manually and create two independent SVG format files of each objects. Figure5 shows the outputs of the first stage.

Iterative Optimization with SDS loss In the second stage, to generate independent motions for each object, we modified the original optimization method [1] to include three different types of iterative processes. In the initial iteration, both local and global motion predictors, modeled as basic Multi-Layer Perceptrons (MLPs), are optimized with respect to the first object. Subsequently, in the second iteration, an identical optimization is applied to a distinct object. However, following these two stages, the inter-object relationship may exhibit signs of disruption. To mitigate this, a third iteration is introduced where the local motion predictors are further optimized for the image of the merged objects. This approach allows us to account for the relationship between the objects, such as their relative viewpoints, resulting in more natural video output.

Backpropagation As illustrated in Figure2, the network for learning displacement is divided into two components: a global network responsible for capturing displacement between frames, and a local network that models movement within individual frames. The process begins with the injection of a vector sketch image, which is subsequently converted into rasterized video frames. Both global and local displacement parameters are then trained independently through backpropagation, utilizing SDS loss.

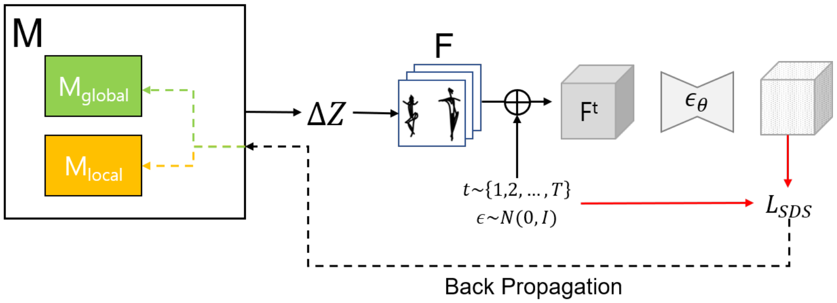


Figure 2: The model optimize the M by using SDS loss. In the 1^{st} and 2^{nd} iteration, M_{global} and M_{local} are both updated. In the 3^{rd} iteration, only M_{local} is updated.

4 Experiments and Results

4.1 Qualtative Results

As illustrated in Figure3., we can observe the independent movements of each object clearly. For example, one fish moves upward while the other moves downward, or one penguin moves left while the other moves right. Additionally, the butterfly and rose shows that it can also generate a video that only one of the objects is moving and the other just remains stationary. These examples demonstrate that our model successfully generates videos with the intended independent movements.

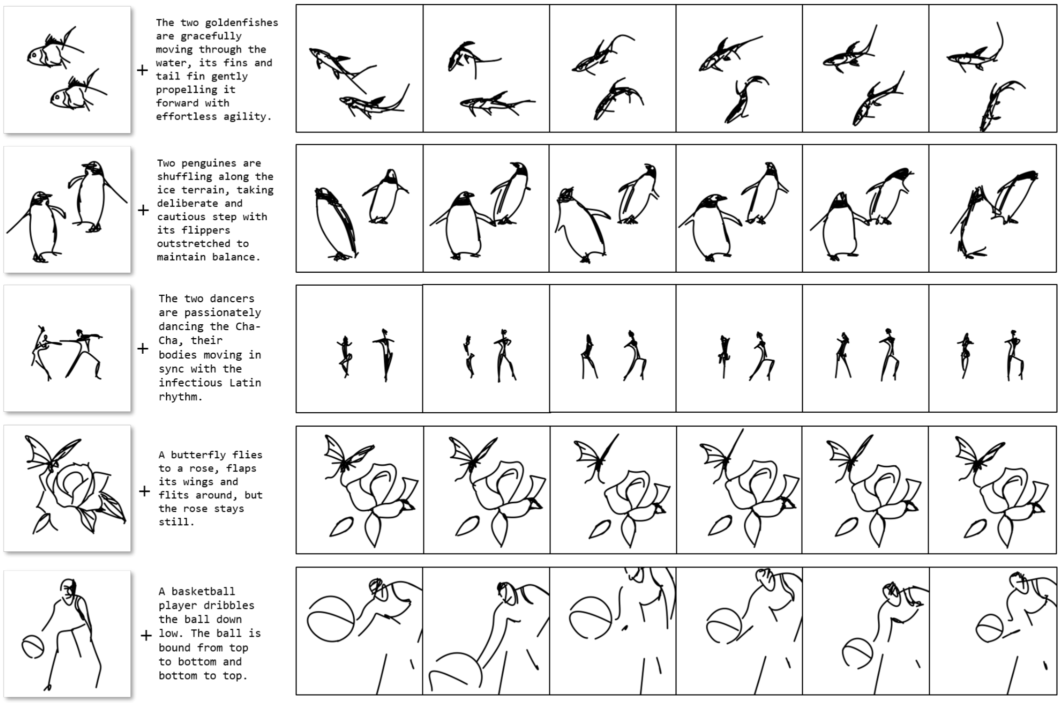

+
The two goldenfishes are gracefully moving through the water, its fins and tail fin gently propelling it forward with effortless agility.
+
Two penguines are shuffling along the ice terrain, taking deliberate and cautious step with its flippers outstretched to maintain balance.
+
The two dancers are passionately dancing the Cha-Cha, their bodies moving in sync with the infectious Latin rhythm.
+
A butterfly flies to a rose, flaps its wings and flits around, but the rose stays still.
+
A basketball player dribbles the ball down low. The ball is bound from top to bottom and bottom to top.

Figure 3: Qualitative Results : Given a multi-object image and a text prompt describing the actions, our model effectively generates videos depicting the multi-object scenario.

4.2 Training Model Selection

We hypothesize that training a single object within a multi-object scenario using a text-to-video prior by optimizing SDS loss is feasible and can yield valuable insights into independent motion dynamics. To identify the optimal combination of learning strategies, we conducted extensive experiments. Our

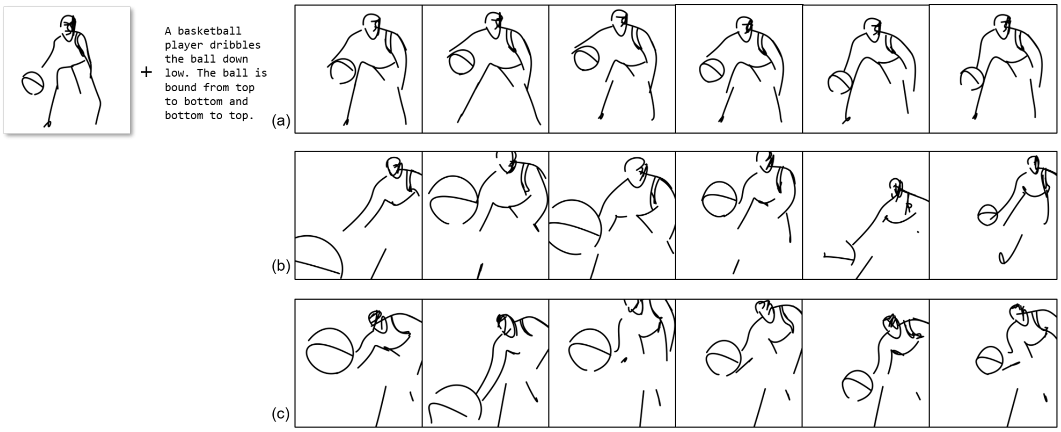

A basketball player dribbles the ball down low. The ball is bound from top to bottom and bottom to top.
+
(a)
(b)
(c)

Figure 4: Training Model Selection: (a) baseline model, (b) both global and local motion predictor for individual objects and combined sketch (c) our final model; solely local motion predictor in the final iteration. In (a) the baseline model the ball does not bound up and down and sticks to the basketball player's hand, while in (b) and (c) the ball bounces up and down and moves more dynamically. Even (c) the movement of basketball player is more natural and continuous.

findings indicate that training both local and global motion predictors for individual objects results in strong performance, leading us to adopt a training scheme that concurrently learns both global and local displacements for each object. While this approach allows objects to move independently and more dynamically, it introduces challenges in maintaining coherence during object interactions. To address this issue, we incorporated combined training, wherein the sketches of each object were merged and trained both globally and locally. But, training only the local displacements in the last iteration can have positive impact even in some cases as shown in Figure4. and can reduce the computational cost. We decide final model in which involves iterative training of both local and global displacements for each object, followed by training solely the local displacements of the combined sketch in the final iteration.

4.3 Comparisons of Methods

Figure6 of the Appendix presents the results of the comparison between the baseline model [1] and ours. In the baseline model, (a), the objects are inseparable, so they move together as if they were a single object. The butterfly and the flower stuck together as they moved up and down. In the basketball scenario, the ball stays attached to the the player's hand without bouncing. In the two penguins, the right penguin's head is tilted to the side. But, in our model, (b), the butterfly flaps and flutters around the flower while the flower remains static. The ball in the basketball scenario starts to bounce and the head of the right penguin in the two-penguin scenario remains balanced and exhibits a natural wobble. On the other hand, the movements of the basketball and the butterfly show that the relationship between multiple objects is adequately considered.

5 Discussions

We propose a new method to generate a video from a sketch for multiple objects. Our method exhibits independent movements on each objects while keeping consistency of their interactions. Our model can be trained iteratively with each objects and combined one by utilizing text-to-video SDS loss. The results indicate that our model outperforms existing models in multi-object scenarios.

Our approach is subject to several limitations. The first stage is labor-intensive, relying on manual effort, and we have not yet developed an end-to-end solution for stroke assignment in sketch images. Furthermore, the method's effectiveness is highly dependent on the text-to-diffusion model, with some prompts yielding outputs that are not conducive to efficient application of SDS loss.

Acknowledgments and Disclosure of Funding

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)

References

[1] Rinon Gal, Yael Vinker, Yuval Alaluf, Amit H. Bermano, Daniel Cohen-Or, Ariel Shamir, and Gal Chechik. Breathing life into sketches using text-to-video priors. *arXiv*, 2023.

[2] Harrison J. Smith, Qingyuan Zheng, Yifei Li, Somya Jain, and essica K. Hodgins. A method for animating children's drawings of the human figure. *ACM Trans. Graph.*, 2023.

[3] Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. Aniclipart: Clipart animation with text-to-video priors. *arXiv*, 2024.

[4] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[5] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv*, 2023.

[6] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Trans. Graph.*, 2022.

A Supplementary Materials

We provide supplementary materials to support a more comprehensive understanding, which can be accessed at `https://github.com/actruce/MultiSketch/blob/main/README.md`. These materials encompass qualitative animations derived from sketches, comparisons between the baseline method and our approaches, as well as an analysis of failure cases along with their underlying causes.

B Data Generation

In our methodology, raster images were initially generated by inputting a text prompt into Adobe Firefly. Subsequently, these images were converted into vector format (SVG) using the CLIPasso [6]. By default, we used 32 strokes and some images have 16 strokes. For specific images such as dancers, fishes, and penguins, we utilized source files available on `https://github.com/yael-vinker/live_sketch`. These source files were either employed in their original form or modified through object duplication and inflation to create multiple instances. Figure5. shows vector image generation and separation procedure.

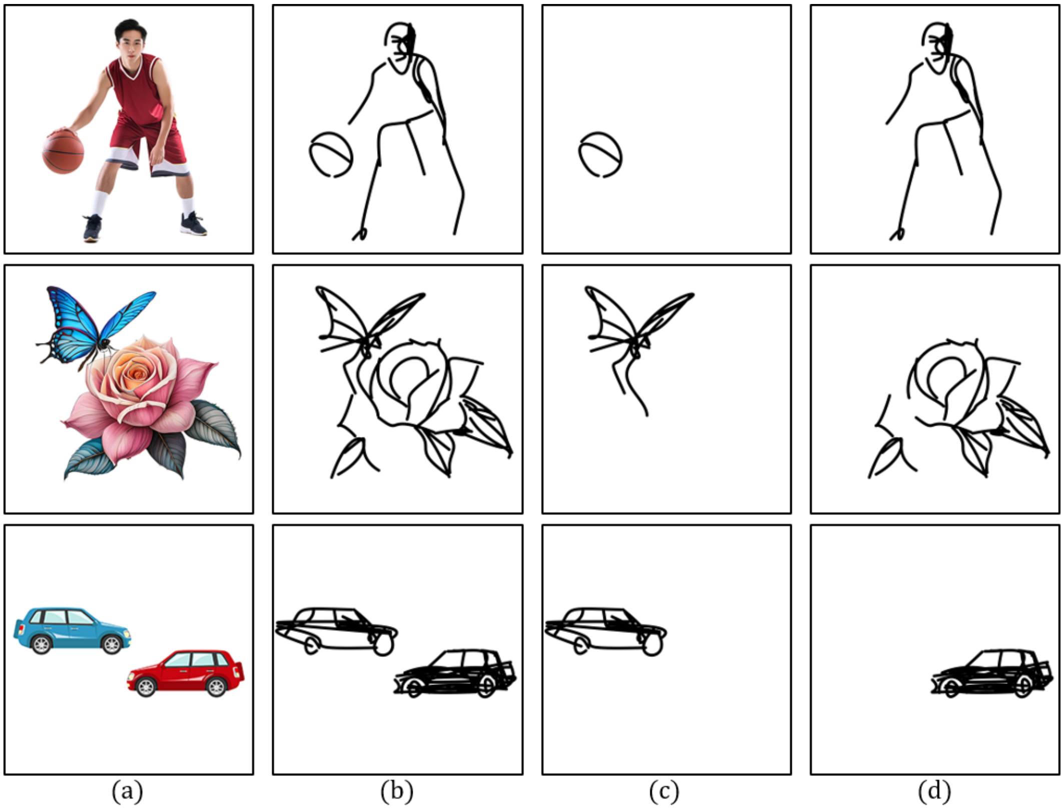


Figure 5: Vector image generation and separation procedure. (a) We generate raster images using Adobe Firefly by injecting text prompt. (b) CLIPasso [6] convert raster images into vector images with SVG file format. (c) and (d) We manually determined which objects each stroke belonged to and separated them into two images.

C Experiment Results

C.1 Implementation Details

Configuration The original configuration [1] was utilized with minimal modifications where necessary. Most configuration settings were retained as provided: *number of iterations* = 1,000 and *number of frames* = 24; however, through experimentation, adjusting the *guidance_scale_global* parameter to 10 yielded the most optimal results. To work with separate image frames, individual sketch objects were introduced as target1 and target2.

C.2 Comparison of Methods

Figure6. presents a comparison between (a) the baseline model by [1] and (b) our proposed model. As illustrated, the movements of individual objects in our model are more natural and independent, while maintaining coherent interactions between objects. In the case of the basketball scenario, the baseline model (a) shows the ball remaining fixed to the player's hand without bouncing, whereas in our model (b), the ball exhibits dynamic bouncing and movement. For the butterfly and rose flower scenario, in (a) the rose flower inconsistently grows and shrinks in size, and the butterfly flaps less frequently, while in (b), the rose flower remains static, allowing the butterfly to flap and flutter around it more naturally. Regarding the movements of two penguins, in (a), the right penguin's head is tilted to one side, whereas in (b), the penguin's head remains balanced and exhibits a natural wobbling motion.

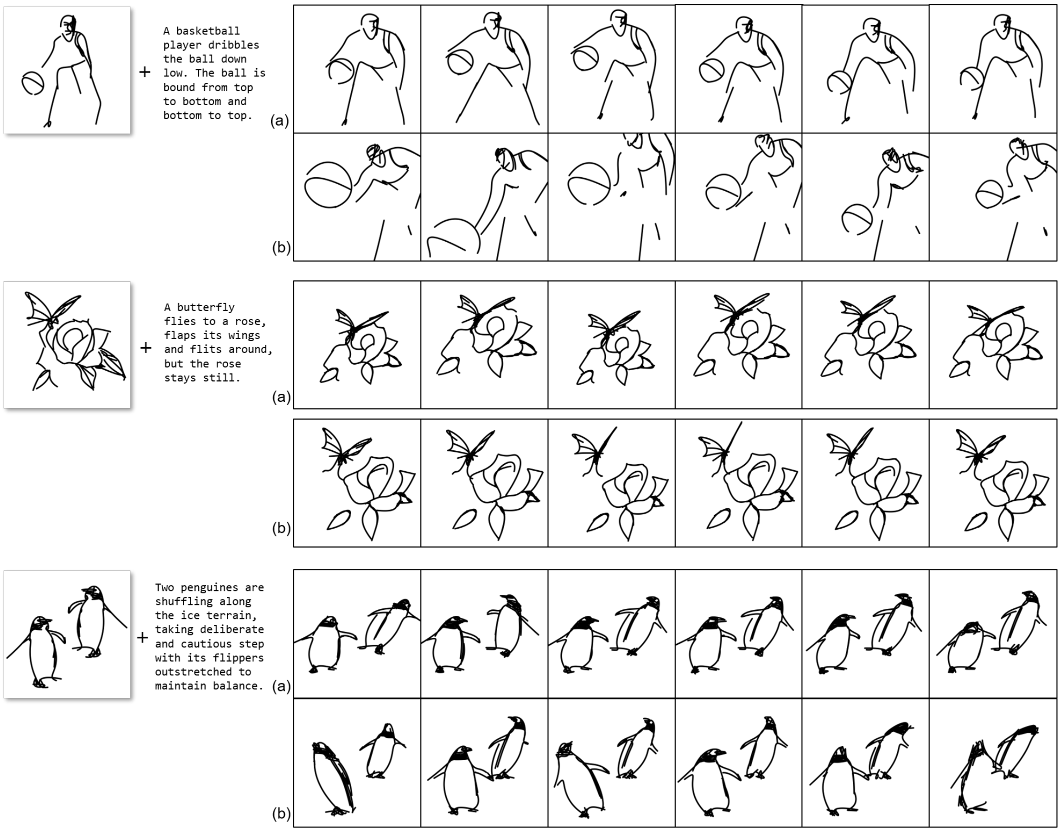
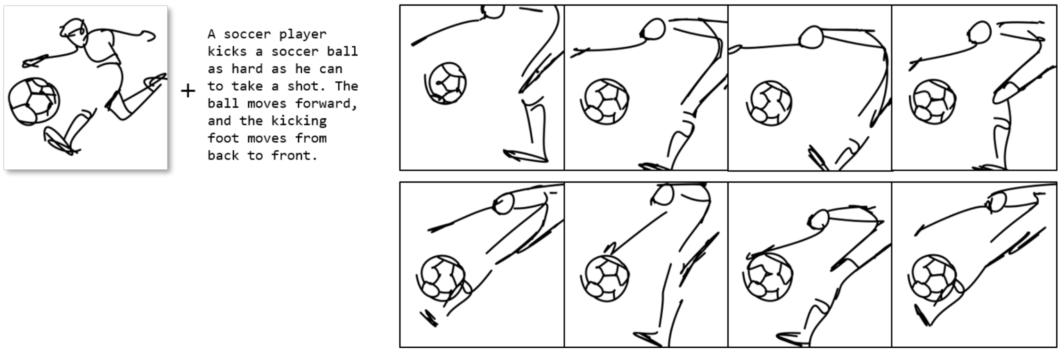

+
A basketball player dribbles the ball down low. The ball is bound from top to bottom and bottom to top.
(a)
(b)
+
A butterfly flies to a rose, flaps its wings and flits around, but the rose stays still.
(a)
(b)
+
Two penguines are shuffling along the ice terrain, taking deliberate and cautious step with its flippers outstretched to maintain balance.
(a)
(b)

Figure 6: The comparisons between (a) the baseline model and (b) our proposed model

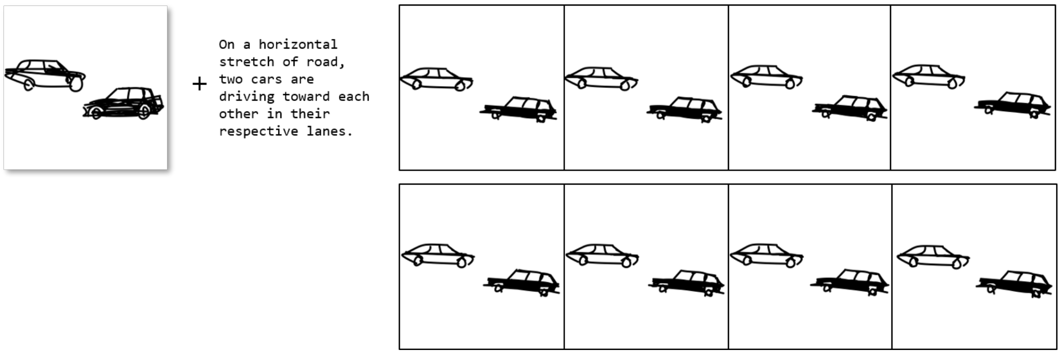
C.3 Failure Cases

As mentioned in the discussion section, our method has the limitation that it relies heavily on pre-trained text-to-diffusion models. Some prompts always produce poor quality videos, making training difficult. Several examples show failure cases Figure7. Figure8. that produce incomplete movements or wired interactions between objects.



A soccer player
kicks a soccer ball
as hard as he can
to take a shot. The
ball moves forward,
and the kicking
foot moves from
back to front.

Figure 7: There is an issue where the movement of the ball is absent, while the supporting foot of the soccer player is in motion.



On a horizontal
stretch of road,
two cars are
driving toward each
other in their
respective lanes.

Figure 8: The vehicles, positioned facing each other, exhibit only minimal movement in place without passing one another.