Editing Motion Graphics Video via Motion Vectorization and Transformation

SHARON ZHANG, Stanford University, USA
JIAJU MA, Stanford University, USA
JIAJUN WU, Stanford University, USA
DANIEL RITCHIE, Brown University, USA
MANEESH AGRAWALA, Stanford University and Roblox, USA

ACM Reference Format:
Sharon Zhang, Jiaju Ma, Jiajun Wu, Daniel Ritchie, and Maneesh Agrawala. 2023. Editing Motion Graphics Video via Motion Vectorization and Transformation. *ACM Trans. Graph.* 42, 6, Article 229 (December 2023), 6 pages. https://doi.org/10.1145/3618316

This document provides more details on results of the motion vectorization pipeline and program transformation API. Section 1 provides more comparison between our method and the sprite-from-sprite method [Zhang et al. 2022] (Section 5 of the main paper) and more examples of vectorization errors. Section 2 includes further implementation details on object transformers (Section 6.3 of the main paper), as well as additional examples of program transformers (Section 7 of the main paper) using these object transformers.

1 RESULTS: MOTION VECTORIZATION

Comparison to Zhang et al. [2022]. We compare our results to sprite-from-sprite [Zhang et al. 2022], which takes a video as input and decomposes it into N sprites. There are several key differences between sprite-from-sprite and our method. Most importantly, the appearance of each sprite in the sprite-from-sprite representation is time-dependent, meaning the per-frame homographies may not fully characterize the sprite motion. On the other hand, our representations contains a single appearance for each object, so the object motion throughout the video is fully explained by our motion parameters. Consequently, the sprite-from-sprite representation may reconstruct the input video more accurately, but editing the video still has to be done at the per-frame level rather than the object level. Moreover, sprite-from-sprite often reconstructs the input video more accurately at the expense of a meaningful sprite decomposition (see Figure 1). Another difference in the sprite-from-sprite method is that it assumes fixed depth ordering. This can impact the quality of the sprite appearances. Figure 2 shows the results of a video with alternating depth ordering after applying our method and sprite-from-sprite.

Out of the 38 test videos, sprite-from-sprite decomposed 30 of them and ran out of memory for the remaining 8 videos. Overall, the average reconstruction error of sprite-from-sprite [Zhang et al. 2022] is 0.018 on the 30 successfully decomposed videos, compared to our average reconstruction error of 0.0079 on the same subset of videos. A number of sprite-from-sprite decompositions also resulted in trivial sprites, *i.e.*every pixel was assigned to a single sprite. We use a Nvidia Titan RTX as opposed to a Nvidia RTX 3080, which was used in the original paper. The Github repository for sprite-from-sprite notes that this may affect the sprite decomposition results.

Non-affine motions and severe occlusions. As mentioned in the main paper, our motion vectorization pipeline assumes an affine motion model. When videos do not follow this assumption well, our pipeline is still able to generate an SVG motion program, but the resulting representation may contain extra objects and the reconstruction may not be as accurate. Objects which never appear fully unoccluded may also appear to go against the affine motion assumption, as we never have a true canonical appearance from the video content. Our pipeline is still able to produce an SVG motion program, but the motion program may not reconstruct the input video as well. Figure 3 illustrates these two cases.

2 MOTION PROGRAM TRANSFORMATION: EXAMPLES

The object transformers in our API transform the timing, spatial motion and appearances of objects. Figure 4 shows code for several object transformers, and Figures 5–7 illustrate applications of different object transformers to create complex variations of an input motion graphic. The following sections describe the implementation of each object transformer in more detail.

Retiming

We have developed several object transformers that change the timing of individual object(s) using the `retime` operator.

Linear time stretch/shrink. To linearly stretch (or shrink) the timing of an object by a factor of k over a source frames `[sFrmA, sFrmB]`, we specify a target frame range `[tFrmA, tFrmB]` such that its duration is k times the duration of the source frame range and we use the identity easing function $f(t) = t$. Code shown in Figure 4c.

Slow in/out easing. To add slow in/out easing to the timing of an object over source frames `[sFrmA, sFrmB]` we specify a target

Authors' addresses: Sharon Zhang, Stanford University, USA, szhang25@stanford.edu; Jiaju Ma, Stanford University, USA, jiajuma@stanford.edu; Jiajun Wu, Stanford University, USA, jiajunwu@cs.stanford.edu; Daniel Ritchie, Brown University, USA, daniel_ritchie@brown.edu; Maneesh Agrawala, Stanford University and Roblox, USA, maneesh@cs.stanford.edu.

0730-0301/2023/12-ART229 $15.00
https://doi.org/10.1145/3618316

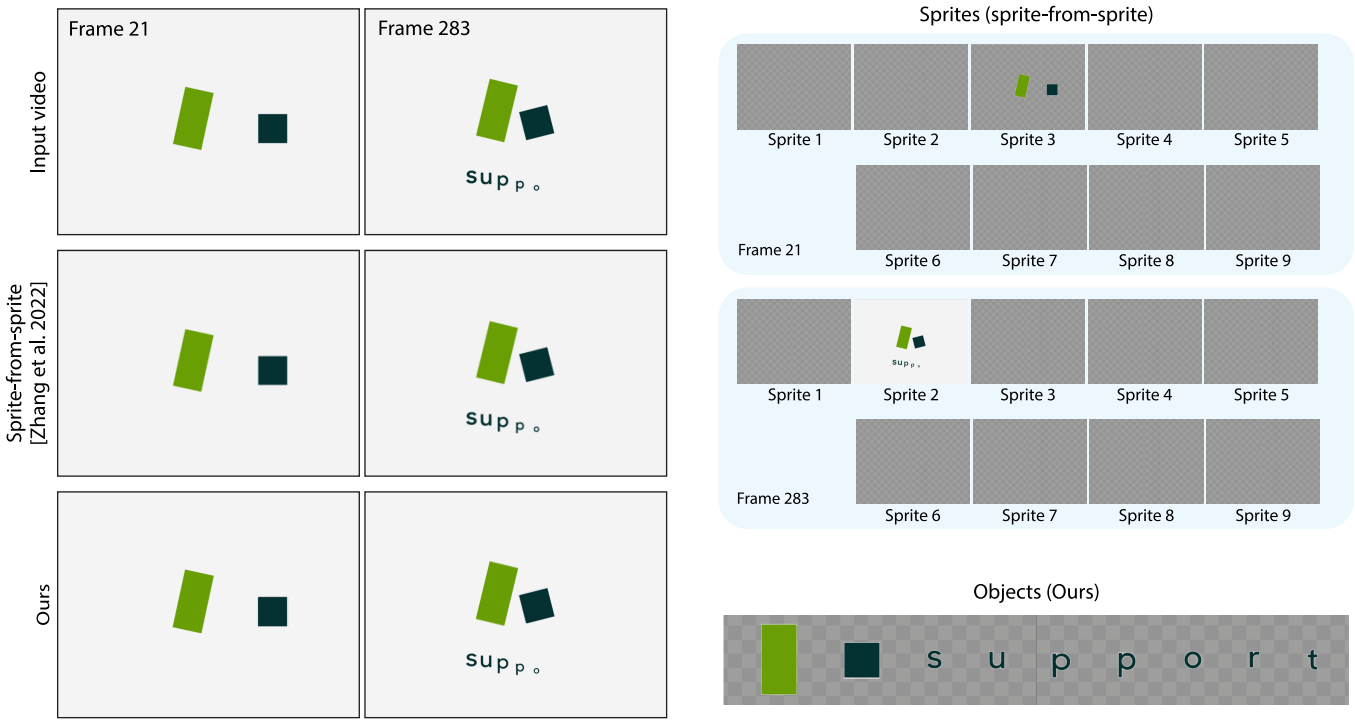

Fig. 1. **Comparison between our method and sprite-from-sprite [Zhang et al. 2022].** While the reconstruction quality of sprite-from-sprite is comparable to ours, the sprite decomposition may not be meaningful. In this video (*logo8*), our method correctly extracts nine objects. Decomposing the same video into nine foreground sprites with sprite-from-sprite results in multiple objects within one sprite, and also inconsistent object assignments across sprites (the two objects in Sprite 3 at frame 21 appear in Sprite 2 at frame 283).

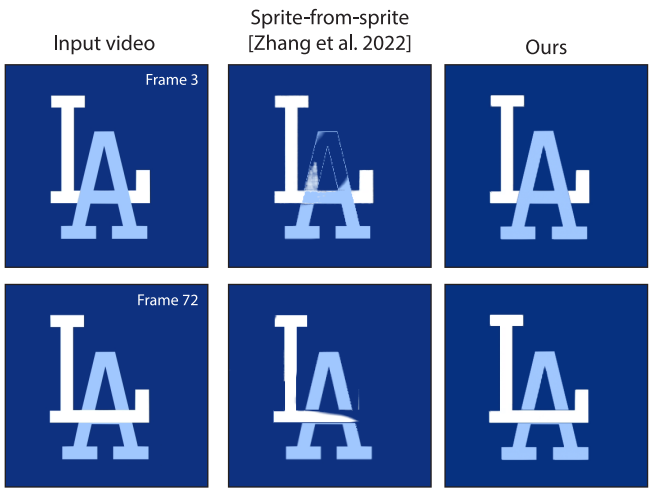

Fig. 2. Sprite-from-sprite [Zhang et al. 2022] decomposes a video into a fixed number of sprites, with depth ordering fixed throughout the video. In this video *LA*, the letters 'L' and 'A' alternate in front and behind positions. Our method is able to model this variation in depth but the sprite-from-sprite reconstruction suffers from flickering artifacts.

frame range of the same duration as the source and use a nonlinear easing function (e.g., $f(t) = t^4$ to generate slow in timing)[1].

Animating on 2s, 3s, Ns. Traditional animators sometimes hold frames of moving objects to produce to stylize the motion or create a stop-motion look. We introduce this effect to the motion of an object, by using a step function as the easing where the size and position of each step is based on how long the each frame should be held and the duration of the target frame range.

Removing held frames for each object. We can also remove held frames from motions of an object to smooth its motion. To do so, we first run an event query to find all the `heldFrames` for an object and linearly shrink each segment of held frames to a single frame. To restore the timing we linearly stretch each reduced frame and next frame back to the duration of the original held frame segment.

Retiming object motion to music beats. Given a piece of music, this object transformer extracts the beats in units of frames using libROSA [McFee et al. 2015] and then breaks the timeline of an object into segments. The segments can either match the length of time between successive beats, or align with events such as `motionCycleFrames` or `collisionFrames`. Next we apply the `retime` operator so that the object motion segments match the beat length and use a non-linear easing function that accentuates the

[1]See `easings.net` for a collection of commonly used easing functions.

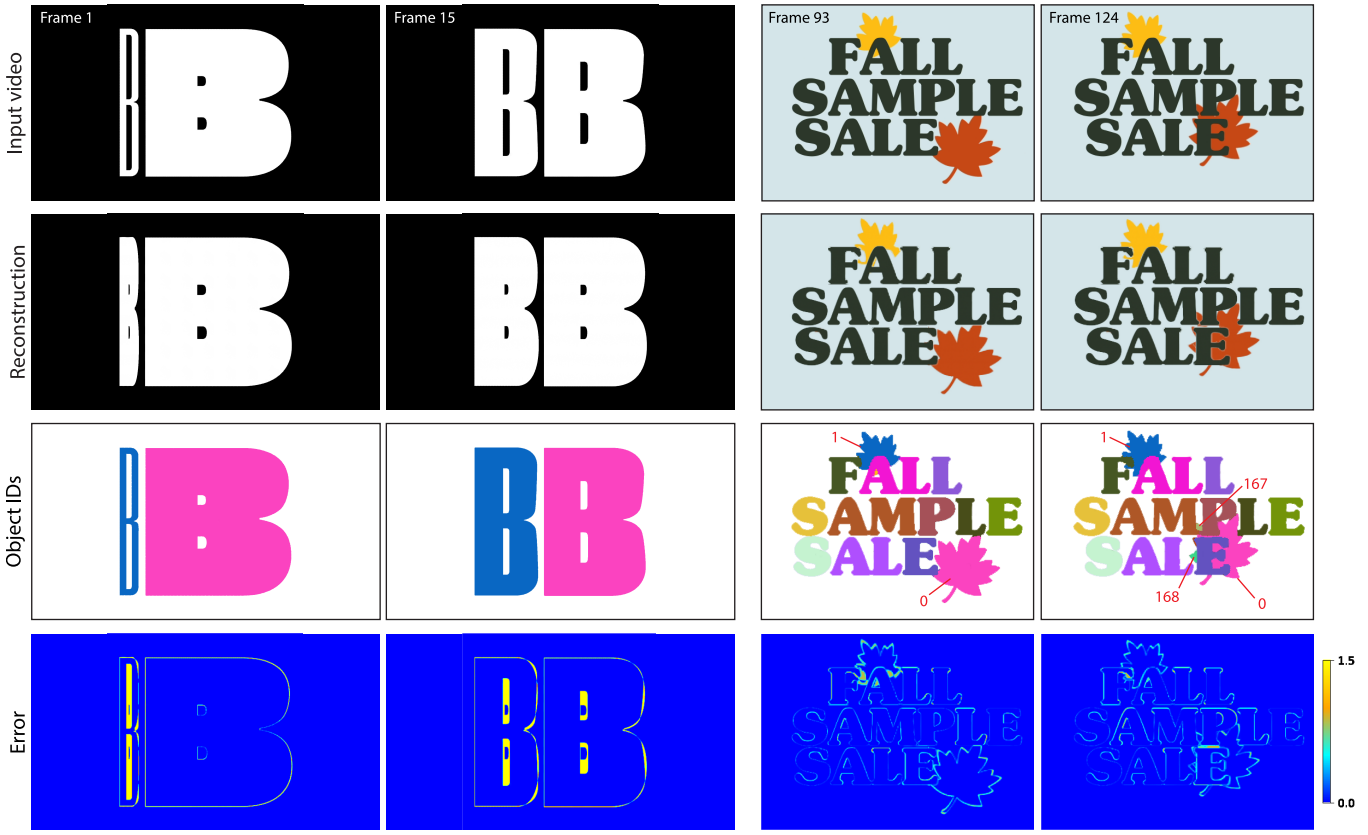

Fig. 3. **Left: Non-affine deformations.** The two 'B's in this video do not deform affinely, so our SVG motion program does not reconstruct it accurately. **Right: Severe occlusions.** The leaves never appear fully unoccluded in this video. This causes the SVG motion program to represent one of them with multiple object IDs (0, 167 and 168) and also results in reconstruction errors along the occlusion boundaries (see misalignments around 'FALL').

motion in and out of each beat point, in the manner of Davis and Agrawala [2018]. Code shown in Figure 4d.

Spatial Motion Adjustment

We can use the `adjLocalMotion` and `adjGlobalMotion` operators to adjust the spatial trajectories of objects.

Motion texture. In the context of motion graphics we define *motion textures* as local spatial perturbations to the motion of an object. We can apply such motion textures to an object by first defining a transform function `xformFn` that specifies a perturbation to be applied in the local coordinate frame (i.e. the canonical image frame) of the object and then passing it into the `adjLocalMotion` operator. For example suppose the input video has and object translating from left to across the screen. We can provide a transform function that translates the object along its local y-axis according to sine function, to produce an oscillating motion up and down as the object moves from left to right in the frame. We can also create multiple copies of an object and add slightly different local perturbations to each one to create a form of Kazi et al.'s kinetic texture [2014].

Anticipation/follow-through. We can apply Wang et al.'s [2006] cartoon animation filter on the motion of an object in the global coordinate frame to add anticipation and follow-through to the motion. As defined by Wang et al., the cartoon animation filter takes a time varying signal $x(t)$, convolves it with an inverted Laplacian of a Gaussian (LoG) filter and adds the convolved result back to the original signal. In our setting, we treat the motion transforms of the object as the signal and define a transform function `xformFn` that performs the convolution. We then pass this transform function to the `adjGlobalMotion` operator to add the convolved result back to the original motion. Code shown in Figure 4e.

Appearance Adjustment

We can use the `changeAppearance` method to replace the canonical image of and object with a completely new visual appearance. However if the new appearance differs significantly in shape from the original it may not preserve collisions with other objects. We have developed a motion program transformer that can can adjust the motions of the colliding objects after such appearance changes to preserve collisions in certain cases.

Collision preserving appearance change. If the new appearance lies within the contour of the original appearance it cannot introduce any new collisions with other objects in the scene. But even with

Table 1. A comparison of L_2 RGB reconstruction errors of sprite-from-sprite [Zhang et al. 2022] against our method. Videos with textures, photographic elements or color gradients in the foreground or background are marked with ‡. Sprite-from-sprite was unable to decompose the eight videos with more than 32 objects due to out-of-memory errors (indicated by —).

Video	Reconstruction L_2 error	
	Sprite-from-sprite [2022]	Ours
No occlusions and no fast motion		
ball2	0.00019	0.0034
ball3	0.0	0.0024
eyes	0.017	0.0050
format	0.29	0.0036
levers	0.0067	0.0063
support	0.00012	0.0024
Occlusions only		
dog	0.0058	0.017
five	0.0035	0.0024
giftbox1	0.0043	0.0078
giftbox2	0.0038	0.012
hype1	0.0011	0.022
hype2	2.7e-5	0.024
pingpong	0.0095	0.0093
playDesign	1.2e-5	0.0068
sundance	—	0.0071
ball5	0.013	0.0072
sydney (‡)	—	0.0394
morningShow	—	0.011
Fast motion only		
ball4	2.8e-5	0.0026
book2 (‡)	—	0.0095
transforms	0.0	0.0034
seesaw (‡)	0.0095	0.0017
wordAWeek	0.054	0.0036
deconstruct	4.3e-5	0.0010
beautiful	0.013	0.0037
Both occlusions and fast motion		
ball1 (‡)	0.0059	0.0083
face	0.0022	0.0011
filmRadio	—	0.0040
183	3.7e-5	0.010
gsuite (‡)	0.024	0.017
book1 (‡)	0.046	0.0036
kapptivate	5.6e-6	0.0063
avokiddo	0.032	0.0033
dates (‡)	—	0.023
5k (‡)	4.1e-5	0.033
shapeman	0.0	0.0048
confetti	—	0.012
lucy	—	0.013

this restriction, the new appearance may not fill the contours of the original object, and thereby miss collisions that appeared in the original video. However, we can preserve such collisions by locally adjusting the motion of the new object as follows. Our event query method for `collisionFrames` provides the collision point in the local coordinate frame of the original object. We find the closest point on the contour of the new appearance to the collision point and add a local translation to the object, using `adjLocalMotion`, so that this closest point matches the collision point. In practice, we spread the local adjustment so that it occurs gradually over the set of frames from the most recent previous collision of the object. We also allow the motion adjustment to occur on the other object involved in the collision or some combination of both objects. Note that we do not check if the local adjustments will move an object outside the contour of the original object and potentially introduce new collisions. But in practice we have found that new collisions are rare. We also note that this approach only considers pairwise collisions and cannot handle more than one simultaneous collision.

REFERENCES

Abe Davis and Maneesh Agrawala. 2018. Visual Rhythm and Beat. *ACM Transactions on Graphics (TOG)* 37, 4, Article 122 (jul 2018), 11 pages. https://doi.org/10.1145/3197517.3201371

Rubaiat Habib Kazi, Fanny Chevalier, Tovi Grossman, Shengdong Zhao, and George Fitzmaurice. 2014. Draco: Bringing Life to Illustrations with Kinetic Textures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 351–360. https://doi.org/10.1145/2556288.2556987

Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*. 18–24. https://doi.org/10.25080/majora-7b98e3ed-003

Jue Wang, Steven M. Drucker, Maneesh Agrawala, and Michael F. Cohen. 2006. The Cartoon Animation Filter. *ACM Transactions on Graphics* 25, 3 (2006). https://doi.org/10.1145/1141911.1142010

Lvmin Zhang, Tien-Tsin Wong, and Yuxin Liu. 2022. Sprite-from-Sprite: Cartoon Animation Decomposition with Self-Supervised Sprite Estimation. *ACM Trans. Graph.* 41, 6, Article 192 (nov 2022), 12 pages. https://doi.org/10.1145/3550454.3555439

(a) Motion program transformer (skeleton)

```
// Program Transformer structure.
MPTransformer(P, *args, [frmA, frmB]):
  // OBJ SELECTOR: Select objects in P via queries using any criteria
  // specified in the args.
  ...

  // OBJ TRANSFORMER: Apply an object operator to selected objects.
  ...
```

(b) Object selector

```
// Returns a list of object data which match some criteria.
function objSelector(P, queryFn, queryType, criteria, [frmA, frmB]):
  selObjs = {}
  selObjsInfo = {}
  for each obj in selObjs:
    x = queryFn(obj, args.queryType, [frmA, frmB])
    if x matches criteria:
      selObjs.insert(obj)
      selObjsInfo.insert(x)

  return selObjs, selObjsInfo
```

(c) Object transformer: Linear time stretch/shrink

```
// Linear time scale by factor of k in frame range [frmA, frmB].
function linearRetimeObjTransformer(selObjs, k, [frmA, frmB]):
  for each obj in selObjs:
    sourceDur = frmB - frmA + 1
    targetDur = k * sourceDur
    // Retime from source range [frmA, frmB] to target frame range
    // [frmA, frm + targetDur].
    retime(obj, [frmA, frmB], [frmA, frmA + targetDur], f(t)=t)
```

(d) Object transformer: Retiming object motion to music beats

```
// Retime to music beats (assume video has more segments than beats).
function retimeToBeatsObjTransformer(selObjs, music, eventType, [frmA, frmB]):
  // Get music beat points using libROSA in units of frames.
  beatPts = getMusicBeatPts(music)

  for each obj in selObjs:
  // Form video segments for each beat segment between beat points based on
  // eventType. If eventType is null default to beatPts as segment points.
  if eventType == null:
    segPts = beatPts
  else:
    segPts = eventQuery(obj, eventType, [frmA, frmB])

  for index i in segPts:
    // beatPts is in units of frames and includes a beat point at 0.
    retime(obj, [segPts[i], segPts[i + 1]],
           [beatPts[i], beatPts[i + 1]], f(t)=t^4)
```

(e) Object transformer: Anticipation/follow-through

```
// Add anticipation/follow through via Cartoon Animation Filter.
function anticipateFollowThruObjTransformer(selObjs, [frmA, frmB], A, sigma):
  for each obj in selObjs:
    // Define the cartoon animation filter based on Wang et al.
    function cartoonAnimationFilter(t, obj, [frmA, frmB], A, sigma):
      // Copy and pad segment of xForms to be set up for convolution later.
      tmpXForms = copy(obj.xForms[frmA, frmB])
      pad(tmpXForms, 0.5 * sigma)
      // -LoG is the inverse of the Laplacian of Gaussian function.
      newXForms = A * convolve(tmpXForms, -LoG(sigma))
      return newXForms[t]

  adjGlobalMotion(obj, cartoonAnimationFilter, [frmA, frmB]
```

Fig. 4. The general structure of motion program transformer (a) takes an SVG motion program P as input and alternates object selector blocks with object transformer blocks to modify the SVG program. The object selector function `objSelector` (b) selects one or more objects for transformation. It first runs `queryFn` (i.e., either `propQuery` or `eventQuery`) using the specified `queryType` (i.e., `color`, `collisionFrames`) and then filters the objects to only those that match the specified `criteria`. The object transformers adjust the timing (c, d) motion (e) or appearance of a set of selected objects `selObjs`.

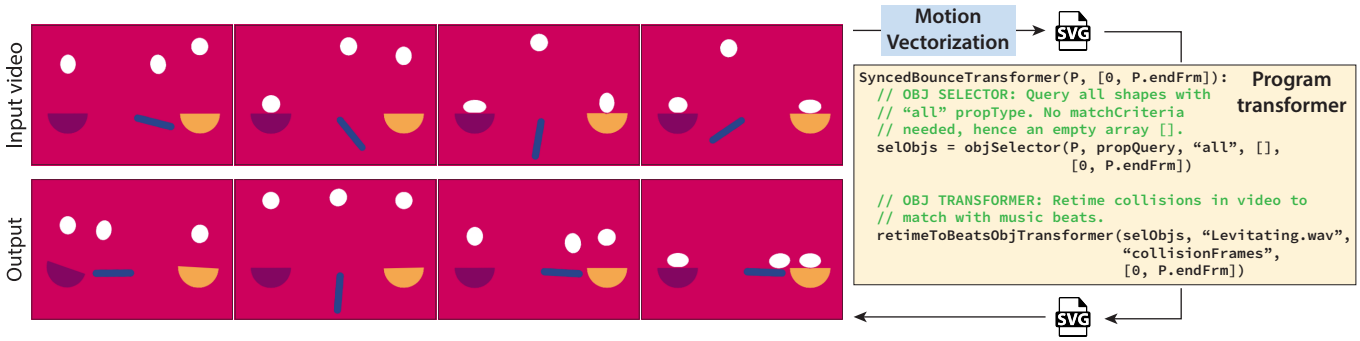

Fig. 5. Retiming collisions to music beat. In the input video the white balls bounce on the platforms underneath at different frequencies. This program transformer retimes the bounces (i.e. collisions) to match the musical beat of a song using the `retimeToBeatsObjTransformer` function (Figure 4d). In the output video all the balls high the platforms at the same time.

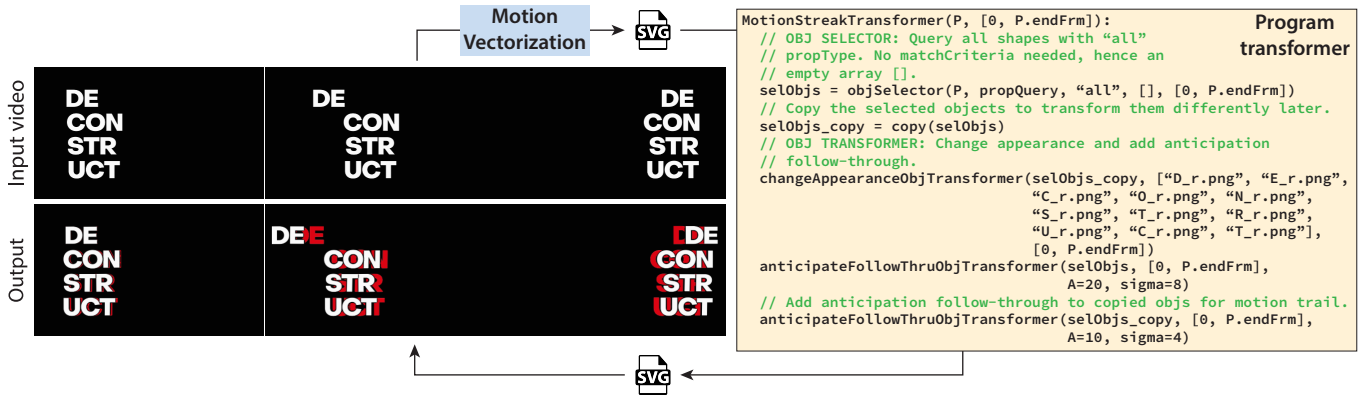

Fig. 6. Adding anticipation/follow-through and motion streak effect. This input video contains white text characters moving over a black background. The program transformer first copies the text objects. It then adds anticipation and follow-through to the original text using the `anticipateFollowThruObjTransformer` (Figure 4e) with a relatively wide $\sigma = 8$. To create the motion streaking effect it it recolors the copied text red using the `changeAppearanceObjTransformer`, and finally adds anticipation and follow-through to the copy using a narrower $\sigma = 4$.

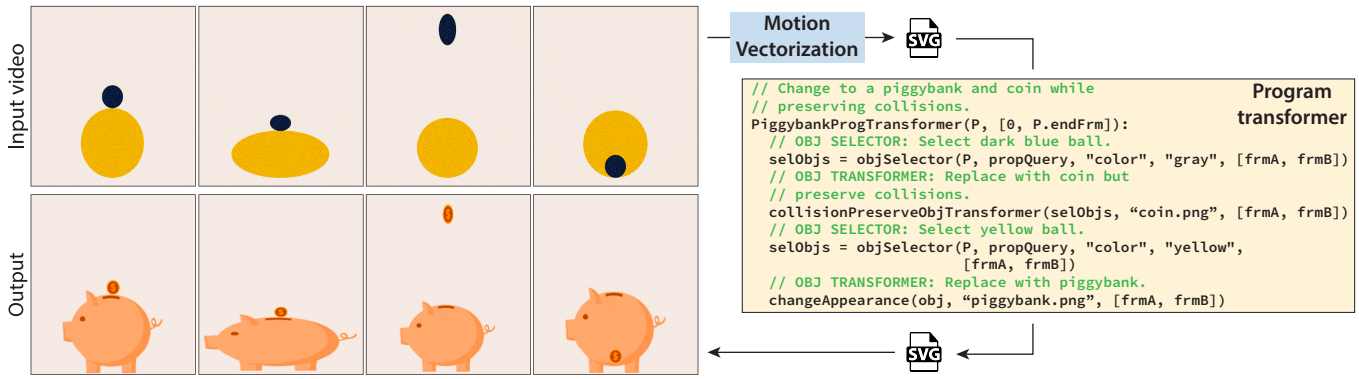

Fig. 7. Changing appearance while preserving collisions. This input video contains two balls that interact with one another with the dark blue ball bouncing around outside and inside the yellow ball. The program transformer changes the blue ball into a coin that is smaller than the blue ball. It then uses the `collisionPreserveObjTransformer` to adjust the motion of the smaller coin so that the collision points are maintained with the yellow ball. Finally it changes the appearance of the yellow ball to a piggy bank with the body of the bank the same size as the yellow ball.