

OnlyFlow: Optical Flow based Motion Conditioning for Video Diffusion Models

Supplementary Material

7. Implementation Details

Dataset We use a random horizontal flip for videos with a 50 percent probability and randomly crop a 256 by 384 area out of the spatially downsampled files.

Optimizer. We used the Adam optimizer [25] with a constant learning rate of 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 5e - 8$ for numerical stability and a weight decay of 10^{-2} . For each parameter update, we clip the gradient norm to 0.4.

Optical flow Feature Extraction. The optical flow encoder first unshuffle the input video by a ratio of 8, increasing the number of channels. For each of the output channels resolution (320, 640, 1280, 1280) we want to obtain, the encoder proceed in a cascading way, applying 2 times the following blocks:

- a ResNet block with a downsampling layer
- a temporal attention block containing 8 heads, with a sinusoidal positional embedding on 16 frames

Sampling. We use a PNDM scheduler [29] with a linear beta schedule where $\beta_{start} = 0.00085$, $\beta_{end} = 0.012$, and $T = 1000$. To allow classifier-free guidance, we randomly drop the text condition 10% of the time.

RAFT settings. In both training and evaluation phases we used the RAFT large checkpoint with the defaults number of 12 optical flow refining iteration updates.

8. Usage tips and tricks

Input video frame rate Our model inference contains two opposing constraints on the conditioning frame rate. On one side, optical flow estimation model perform better between frames that are similar, meaning higher frame rate. At the same time, T2V models often generate 16 or 24 frames per forward pass. Training datasets like WebVid often correspond to video downsampled temporally to 8 fps.

If the optical flow given in input to our model is not within the range of motion of what the base T2V model can achieve in its generation, we may observe a deterioration in prompt adherence or realism.

Aspect ratio. As the AnimateDiff model allows it, we can generate non squared video. Nevertheless, because of the non-convolutional nature of our flow encoder, the optical flow dimensions have to be a multiple of 64.

9. User study details

We submitted the following question to the panel of person expressing their choices for each pair of video :

- Which video best respect the input text?
- Which video best replicate the motion from the input video?
- Which video do you prefer overall?