

OnlyFlow: Optical Flow based Motion Conditioning for Video Diffusion Models

Supplementary Material

001 1. Implementation Details

002 **Dataset** We use a random horizontal flip for videos with a
003 50 percent probability and randomly crop a 256 by 384 area
004 out of the spatially downsampled files.

005 **Optimizer.** We used the Adam optimizer [1] with a con-
006 stant learning rate of 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon =$
007 $5e - 8$ for numerical stability and a weight decay of 10^{-2} .
008 For each parameter update, we clip the gradient norm to 0.4.

009 **Optical flow Feature Extraction.** The optical flow en-
010 coder first unshuffle the input video by a ratio of 8, increas-
011 ing the number of channels. For each of the output channels
012 resolution (320, 640, 1280, 1280) we want to obtain, the en-
013 coder proceed in a cascading way, applying 2 times the fol-
014 lowing blocks:
015 • a ResNet block with a downsampling layer
016 • a temporal attention block containing 8 heads, with a si-
017 nusoidal positional embedding on 16 frames

018 **Sampling.** We use a PNDM scheduler [2] with a linear
019 beta schedule where $\beta_{start} = 0.00085$, $\beta_{end} = 0.012$, and
020 $T = 1000$. To allow classifier-free guidance, we randomly
021 drop the text condition 10% of the time.

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

025 2. Usage tips and tricks

026 **Input video frame rate** Our model inference contains
027 two opposing constraints on the conditioning frame rate.
028 On one side, optical flow estimation model perform bet-
029 ter between frames that are similar, meaning higher frame
030 rate. At the same time, T2V models often generate 16 or
031 24 frames per forward pass. Training datasets like WebVid
032 often correspond to video downsampled temporally to 8 fps.

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

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

3. 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?

References

- [1] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1
- [2] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 1