FlowLoss: Dynamic Flow-Conditioned Loss Strategy for Video Diffusion Models

Kuan-Ting Wu [1] **Kei Ota** [2] **Asako Kanezaki** [1,3]

[1] Science Tokyo, [2] Mitsubishi Electric, [3] RIKEN AIP

Background

Video Diffusion Models (VDM):

- Good at generating visually realistic outputs
- But often lacks physical consistency
- **Optical flow** is widely used as either input condition [1,2] or as auxiliary training objectives [3,4]

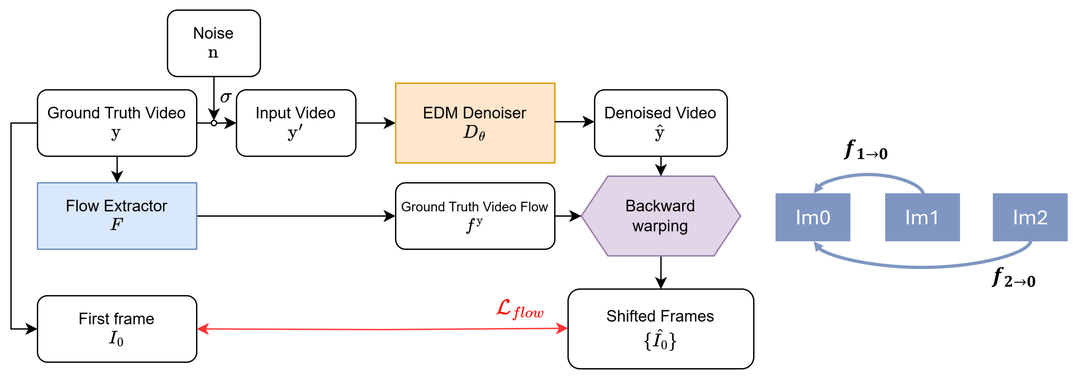
Can we directly provide flow guidance in training?

- Previous researches used optical flow by image warping
- Can we introduce **differentiable optical flow** to our loss function directly?

Related Work

Leverage Optical Flow by Image Warping: [3,4]

- Capture pixel-level information
- Utilize the mapping relationship defined in optical flow



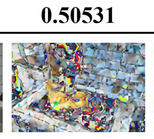
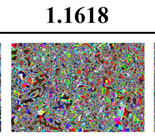
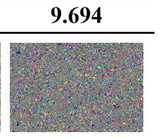
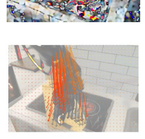
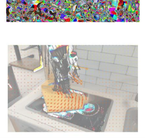
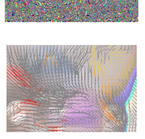
Methodology

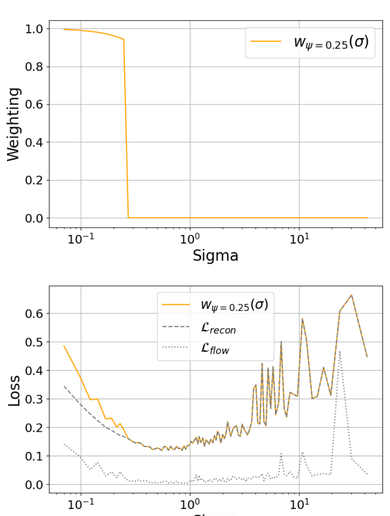
Flow Loss $\mathcal{L}_{\text{flow}}$:

- Use pretrained differentiable optical flow estimator to extract flow
- Compute distance on predicted flow

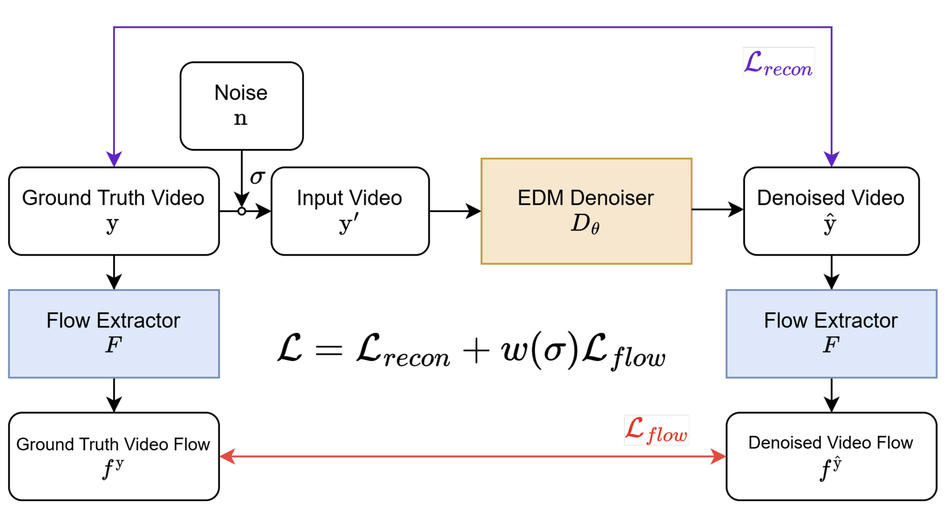
Full Training Objective:

- Combine both reconstruction loss and flow loss
- Use $w(\sigma)$ to balance these two terms
- Apply flow loss only when noise scale is **small**

σ	0.0084365	0.50531	1.1618	9.694
y' $(y+\sigma\cdot n)$				
$\hat{y}$				



$$\mathcal{L}_{\text{recon}} = \mathbb{E}_{\sigma,\text{y},\text{n}}\left[\lambda(\sigma)\cdot\|D_\theta(\text{y}+\text{n};\sigma)-y\|^2\right]$$



Hard Gating Strategy:

- Compute flow loss is time consuming
- Use flow loss only when noise scale is small enough

$$w_\psi(\sigma) = \begin{cases} \frac{1}{\sigma^2+1} & \text{if } \sigma < \psi \\ 0 & \text{otherwise} \end{cases}$$

$$\mathcal{L}_{\text{flow}} = s\cdot\mathbb{E}_{\sigma,y,n}\left[\lambda(\sigma)\sum_t o(\alpha_t^{\text{y}})\cdot\|f_t^y - f_t^{\hat{y}}\|_2^2\right]$$

Experimental Results

Experiment Setups:

- Build **Image-Text-to-Video** VDM on top of Stable Video Diffusion 2.1 (I2V) [5,6] , trained on Bridge V2 dataset

Quantitative Result:

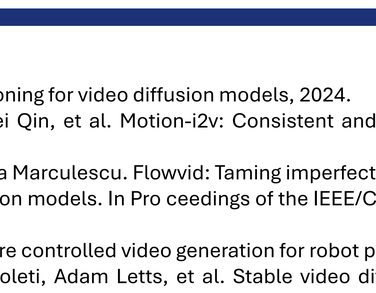
- Slightly outperforms the baseline across some benchmarks
- Hard gating mechanism reduces extra training overhead

Method	FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	FVD ↓	Training time
baseline ($\mathcal{L}=\mathcal{L}_{\text{recon}}$)	**127.6907**	**0.8118**	28.8986	0.0651	165.3577	07h 56m 47s
ours + $w_{\psi=0.0625}(\sigma)$	129.1449	0.8112	**29.0284**	0.0626	191.2058	08h 45m 08s
ours + $w_{\psi=0.125}(\sigma)$	130.1602	0.8097	28.9757	**0.0625**	**161.3095**	10h 31m 28s
ours + $w_{\psi=0.250}(\sigma)$	128.6550	0.8118	28.9896	0.0649	179.0681	14h 29m 26s

*trained with 57,000 steps

Ablation on weighting strategies:

- Validate what is important when noise is large



Weighted Average:

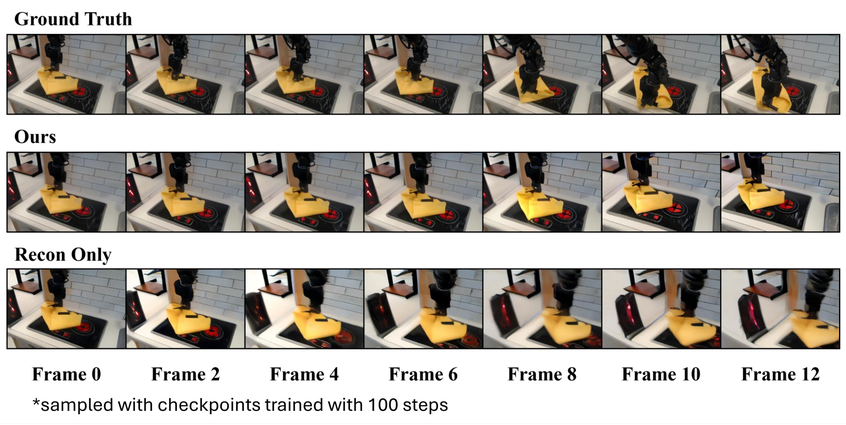
$$\mathcal{L} = (1-w(\sigma))\cdot\mathcal{L}_{\text{recon}} + w(\sigma)\cdot\mathcal{L}_{\text{flow}}$$

$w_{ss}(\sigma)$: Emphasize $\mathcal{L}_{\text{flow}}$ in small-σ
$w_{ls}(\sigma)$: Emphasize $\mathcal{L}_{\text{recon}}$ in small-σ

- Flow supervision is more effective in **low noise stages**

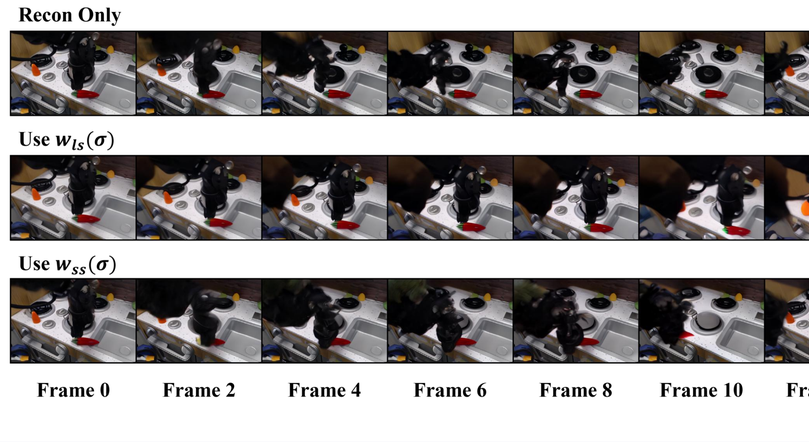
Early-stage Motion Stability:

- Our method can **generate stable outputs in early steps**
- Optical flow help capture **global scene structure**



*sampled with checkpoints trained with 100 steps

Watch the GIFs!



References:

[1] Mathis Koroglu, Hugo Caselles-Dupr ´ e, Guillaume Jean neret Sanmiguel, and Matthieu Cord. Onlyflow: Optical flow based motion conditioning for video diffusion models, 2024.
[2] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In ACM SIGGRAPH 2024 Conference Papers, pages 1–11, 2024.
[3] Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis, 2023.
[4] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In Pro ceedings of the IEEE/CVF conference on computer vi sion and pattern recognition, pages 18444–18455, 2023.
[5] Boyang Wang, Nikhil Sridhar, Chao Feng, Mark Van der Merwe, Adam Fishman, Nima Fazeli, and Jeong Joon Park. This&that: Language-gesture controlled video generation for robot planning. arXiv preprint arXiv:2407.05530, 2024.
[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffu sion models to large datasets. arXiv preprint arXiv:2311.15127, 2023.