MeDM: Mediating Image Diffusion Models for Video-to-Video Translation with Temporal Correspondence Guidance

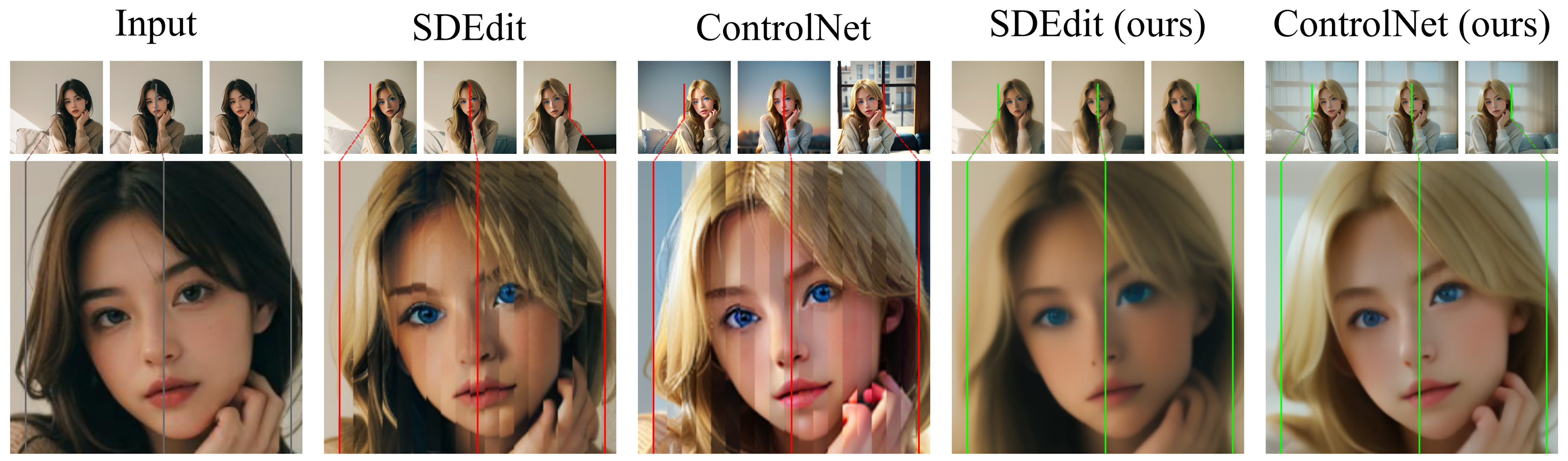
Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, Jun-Cheng Chen

CITI, Academia Sinica, Taipei, Taiwan (R.O.C.)

Summary

Motivation: Many works leverage Stable Diffusion for vid2vid translation. However, the generated videos often lack temporal consistency.

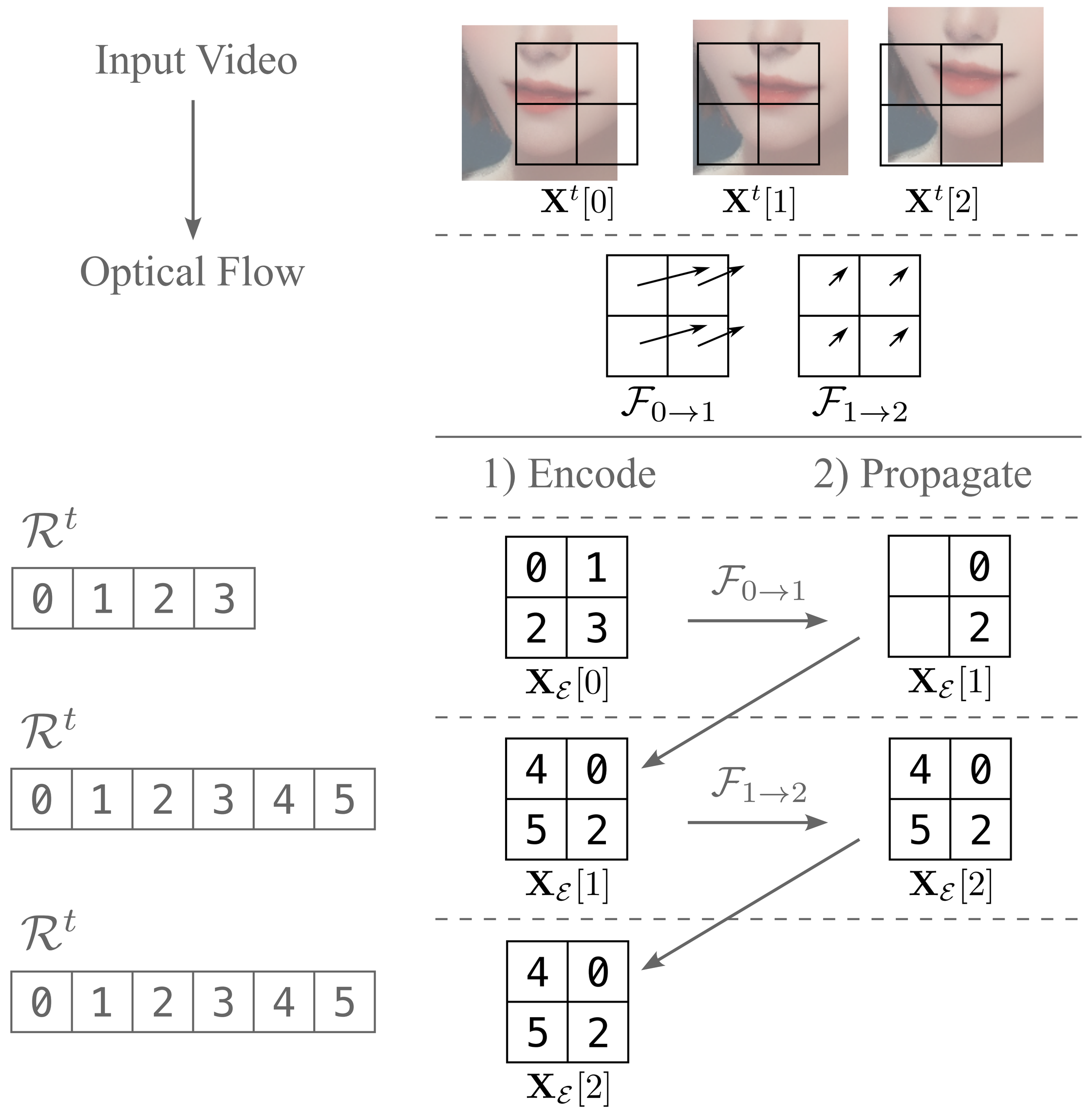
Contribution: We developed a coding algorithm and a pixel repository that enhance image-based denoising cycles for high-quality video-to-video translation, enabling broader LDM applications and effective real-world uses on videos like text-guided editing and anonymization without further finetuning.

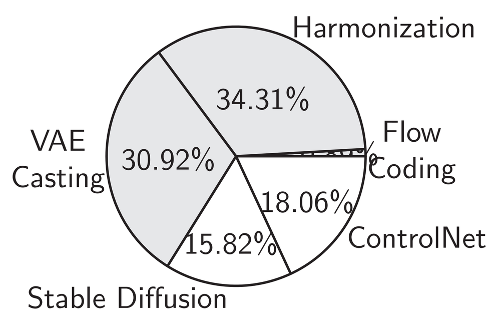


A fluent video should reconstruct a stripe-free image from a horizontal scan.

Proposed Modules

Flow Coding: To guarantee a fluent video, we propose Flow Coding, which leverages optical flows to ensure identical color on each pixel trajectory.





Run-time analysis: our modules only add 2× the amount of time.

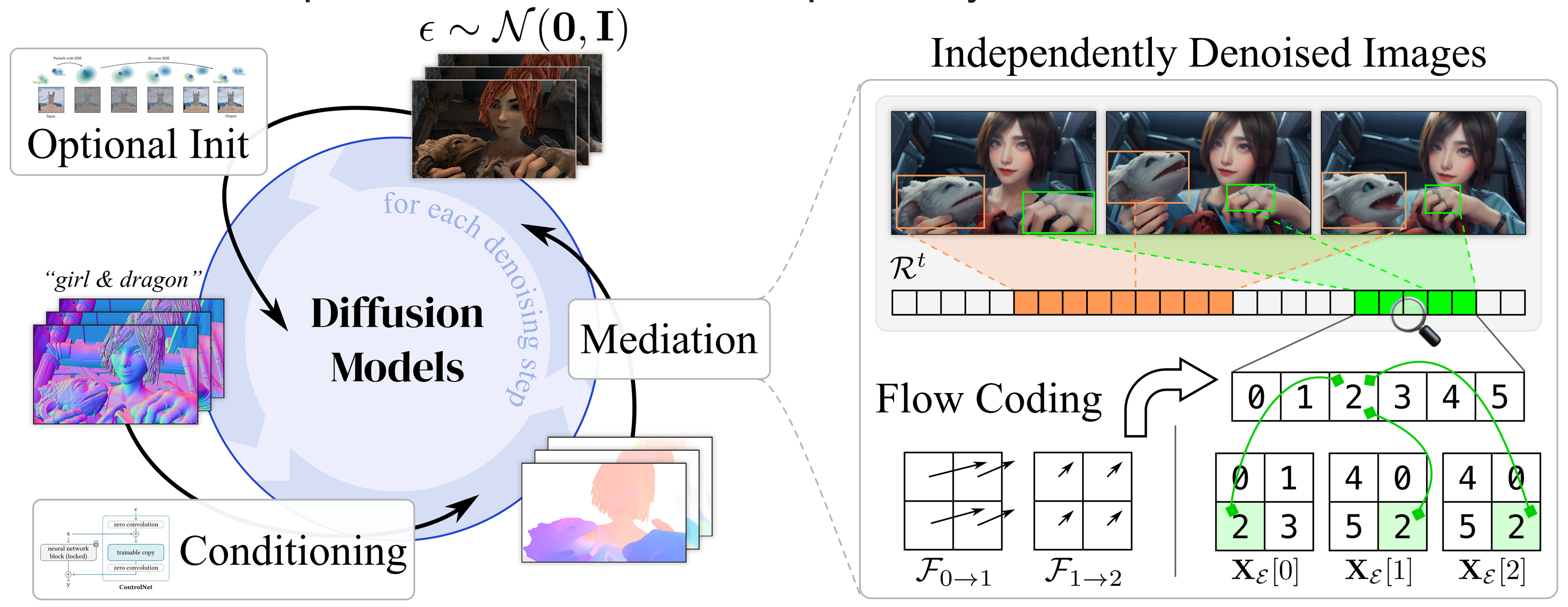
PyTorch Implementation for Fast Mediation:

```python
# assume accumulate=True)
cnt.index_put_((X_E,), 1)
repo.index_put_((X_E,), X_t)
avg = torch.where(cnt>0, repo/cnt, repo)
```

Scan to visit our project page for code & videos!

MeDM: Mediating Image Diffusion Models

We extend the previous works in the image domain to the video domain without any fine-tuning or iterative optimization. Our method mediates independent image score estimations after every denoising step, making them fluent motion pictures when viewed sequentially.



Video pixels are essentially views to the underlying objects. We construct an explicit pixel repository $\mathcal{R}^t$ to represent the underlying world. $\mathcal{R}^t$ is derived from the optical flows $\mathcal{F}$ through the proposed Flow Coding and stores all unique pixels of the video. The encoded frames $\mathbf{X}_\mathcal{E}$ and the repository $\mathcal{R}^t$ enable efficient harmonization of the divergent frame-wise score estimations during the generation process of Diffusion Models.

$$\mathcal{L}^t = \|\mathcal{R}^t[\mathbf{X}_\mathcal{E}] - \mathbf{X}^t\|_2 \quad (1)$$

$$\mathcal{R}^t[\mathbf{X}_\mathcal{E}] \leftarrow G(\mathbf{X}^t) \quad (2)$$

$$G_{avg} = \arg\min_G \mathcal{L}^t \quad (3)$$

G is a function that mixes the pixels in $\mathbf{X}^t$ into the ones in $\mathcal{R}^t[\mathbf{X}_\mathcal{E}]$, and $\mathcal{R}^t$ is the pixel repository at time t. Notably, $\mathcal{R}^t[\mathbf{X}_\mathcal{E}]$ contains significant less unique pixels than $\mathbf{X}^t$, and G is required to harmonize the associated pixels in $\mathbf{X}^t$ into a common values before they can be assigned to $\mathcal{R}^t[\mathbf{X}_\mathcal{E}]$.

Temporal Correspondence Guidance

We use a weight w on the harmonized samples that controls the strength of temporal correspondence guidance (Eq. 4-5), so users can trade temporal coherence for better visual quality. However, Eq. 5 does not work with LDM.

$$\epsilon_\theta^t \leftarrow (1-w)\epsilon_\theta^t + w\mathcal{R}_\epsilon^t[\mathbf{X}_\mathcal{E}] \quad (4)$$

$$\boxed{\mathcal{R}_\epsilon^t[\mathbf{X}_\mathcal{E}] \leftarrow G(\epsilon_\theta^t)} \quad (5)$$

$$\mathbf{X}^t = \sqrt{\alpha^t}\mathbf{X}^0 + \sqrt{1-\alpha^t}\epsilon^t \quad (6)$$

$$\mathbf{X}^0 = \frac{1}{\sqrt{\alpha^t}}\left(\mathbf{X}^t - \sqrt{1-\alpha^t}\epsilon^t\right) \quad (7)$$

$$\epsilon^t = \frac{1}{\sqrt{1-\alpha^t}}\left(\mathbf{X}^t - \sqrt{\alpha^t}\mathbf{X}^0\right) \quad (8)$$

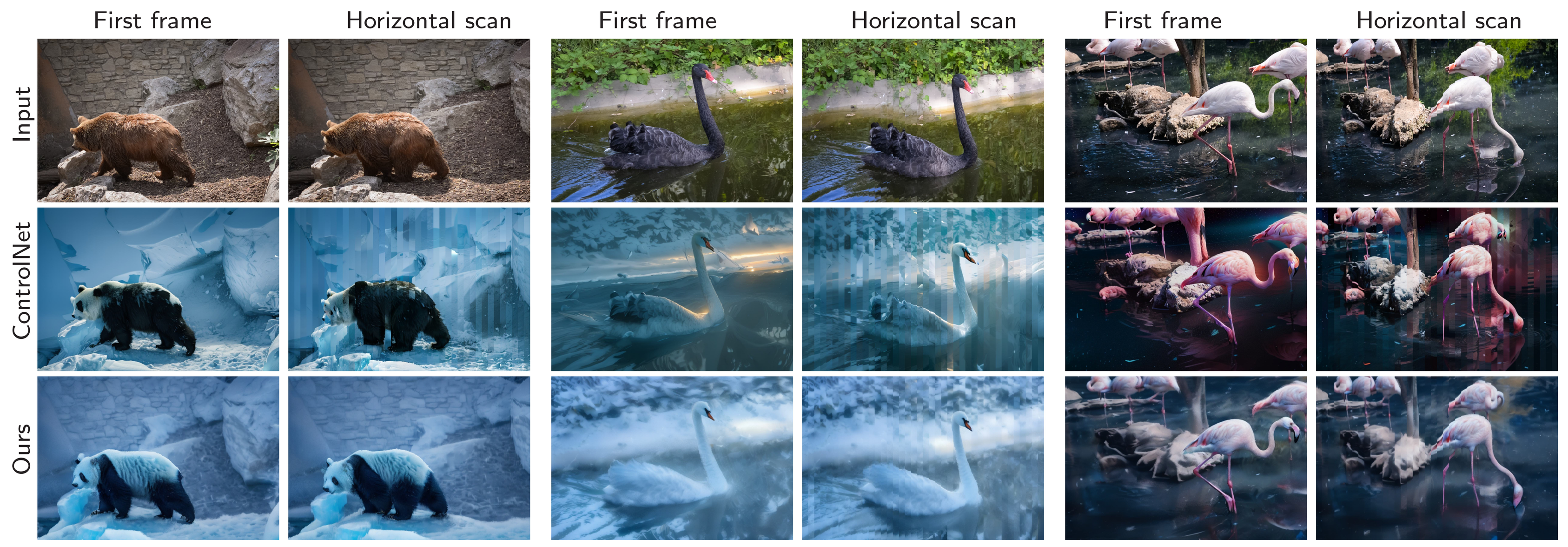
In response, we replace Eq. 5 with Eq. 9 using reparameterization in Eq. 6-8 to provide compatibility of LDMs, where Φ_e and Φ_d are the encoder and the decoder of the Autoencoder in LDM, respectively.

$$\boxed{\mathcal{R}_\epsilon^t[\mathbf{X}_\mathcal{E}] \leftarrow \frac{1}{\sqrt{1-\alpha_t}}\left(\mathbf{X}_t - \sqrt{\alpha_t}\Phi_e\left(G\left(\Phi_d\left(\hat{\mathbf{X}}^{0,t}\right)\right)\right)\right)} \quad (9)$$

Experiments

Video Rendering. *Assistive rendering

Text-guided Video Editing.



We estimate the optical flows from the generated videos, compare the flows with the GT flows and use the EPE between the flows to assess the temporal consistency in videos (a). We also conduct user studies (b,c). Finally, we show that MeDM can perform video anonymization out-of-the-box (d).

	Method	Rendering	Assist. Rendering
MPI Sintel	ControlNet	12.757	5.924
	ControlVideo	12.757	N/A
	Video ControlNet	12.878	N/A
	Rerender A Video	7.953	7.775
	Ours (Est. flow)	1.501	1.570
	Ours (GT flow)	**1.456**	**1.202**
	Animation	0.403	
VKITTI 2	ControlNet	12.575	5.070
	Video ControlNet	15.285	N/A
	Ours (Est. flow)	2.857	2.695
	Ours (GT flow)	**2.483**	**2.217**
	Animation	1.737	

(a) EPE for video rendering

	Method	Video quality	Realism
Non-Assistive	ControlNet	1.662	2.718
	ControlVideo	2.437	1.775
	Control-A-Video	3.662	1.606
	Video ControlNet	2.197	2.380
	Rerender A Video	2.183	2.127
	Ours	**4.423**	**4.014**
	Animation	4.423	3.113
Assistive	ControlNet	2.465	2.408
	Rerender A Video	1.887	2.479
	Ours	**4.380**	**3.958**

(b) User study for rendering

Method	Video quality	Text alignment
ControlNet	2.377	3.289
Pix2Video	1.451	1.592
ControlVideo	2.289	2.430
Control-A-Video	2.634	1.859
Rerender A Video	3.042	3.099
Ours (Lineart)	**4.042**	3.810
Ours (Instruct P2P)	3.901	**4.338**

(c) User study for editing

Method	Recognizability	Realism	Faithfulness
DeepPrivacy	63.01%	2.019	4.216
Ours	**20.83%**	**3.507**	**4.258**

(d) Video anonymization

Acknowledgments: This research is supported by National Science and Technology Council, Taiwan (R.O.C), under the grant number of NSTC-112-2634-F-002-006, NSTC-112-2222-E-001-001-MY2, and NSTC-110-2221-E-001-009-MY2, and Academia Sinica under the grant number of AS-CDA-110-M09.