CVPR Nashville JUNE 11-15, 2025

KAIST AI Kim Jaechul Graduate School

EverEx

Optical-Flow Guided Prompt Optimization for Coherent Video Generation

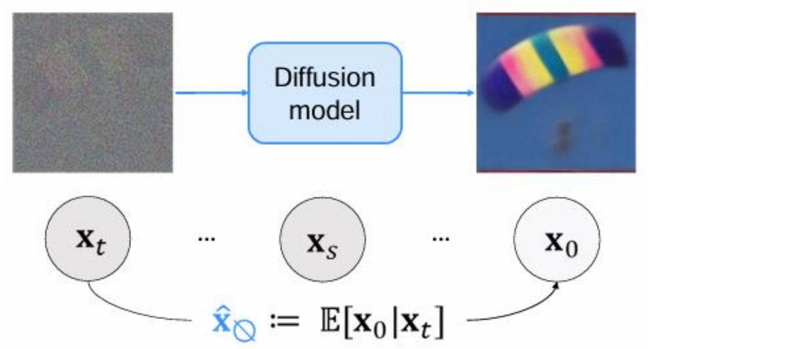
Hyelin Nam†‡* Jaemin Kim‡* Dohun Lee‡ Jong Chul Ye‡

†EverEx, ‡Korea Advanced Institute of Science and Technology (KAIST), *Equal contribution

Paper Project Page

Background

Zero-shot Guidance in Diffusion Models



$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\hat{x}_\oslash - \gamma_t \nabla_{\hat{x}_\oslash} \ell(\hat{x}_\oslash)\right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon}_\oslash$$

1. Modulation for guidance

2. Follow unconditional DDIM sampling

Q. Can we use Zero-shot guidance in Video Diffusion Model (VDM) to improve the motion quality?

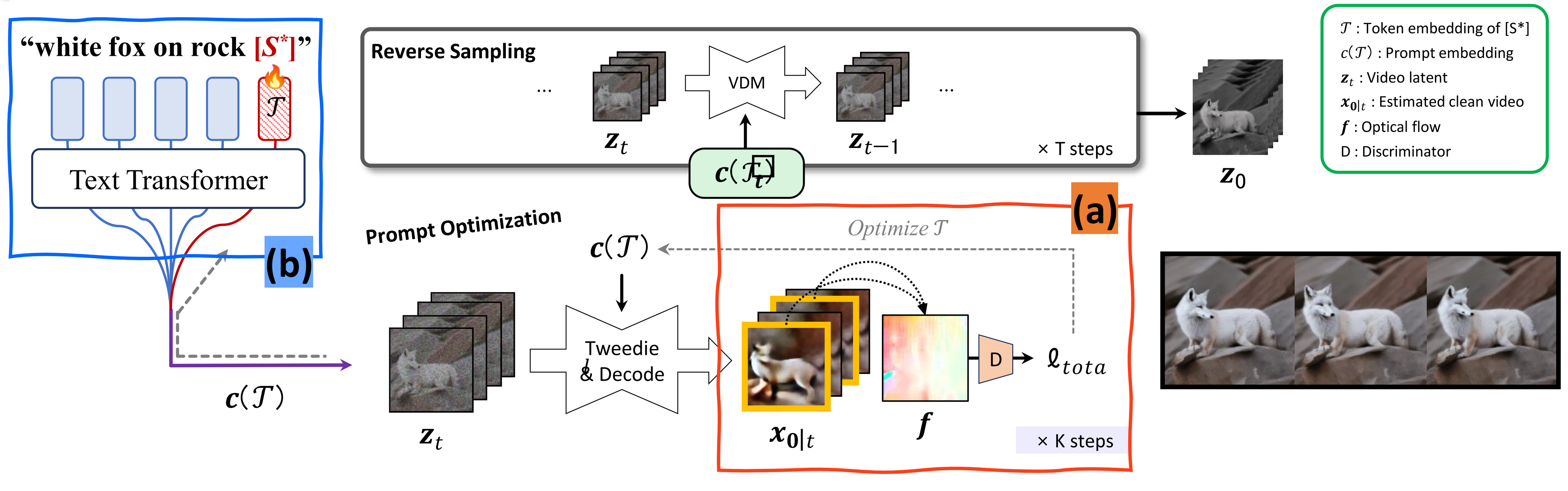
1. Loss Ambiguity

: Motion information must be disentangled to define an effective loss.

2. Computational Cost

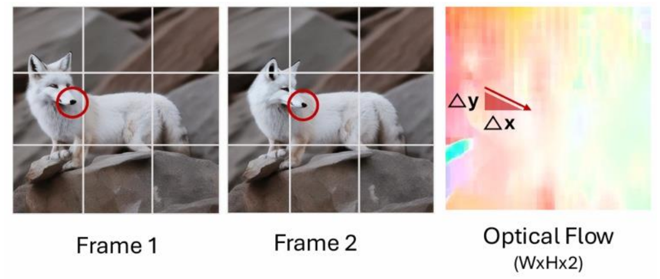
: Consistent updates need all-frame loss computation, which is costly.

Method: MotionPrompt



(a) Optical Flow-based Discriminator

1. Use **optical flow** to disentangle motion from pixel appearance.



2. Pre-train a **discriminator** to classify real and generated flows, and **guide sampling toward realistic motion.**

$$\ell_{\mathrm{disc}}(z_t, c(\mathcal{T})) := \log(1 - \phi_{\theta^*}(\mathbf{f}(\hat{x}_t(c(\mathcal{T})))))$$

$$\ell_{\mathrm{total}}(z_t, \mathcal{T}) := \lambda_1 \ell_{\mathrm{disc}}(z_t, c(\mathcal{T})) + \lambda_2 \ell_{\mathrm{TV}}(z_t, c(\mathcal{T})) + \lambda_3 ||\mathcal{T} - \mathcal{T}_0||_2^2,$$

(b) Prompt Optimization

$$\hat{c}_t := c(\hat{\mathcal{T}}_t), \hat{\mathcal{T}}_t = \arg\min_{\mathcal{T}} \ell(z_t, c(\mathcal{T}))$$

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}(\hat{z}_t(\hat{c}_t)) + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(z_t, t, \hat{c}_t)$$

1. To reduce memory and ensure frame consistency, **attach learnable tokens to the prompt instead of tuning the full embedding.** This preserves semantic meaning and provides consistent guidance across frames.

2. Since it affects low-frequency structure, **optimize only during the first 10–15 timesteps**.

Experiments

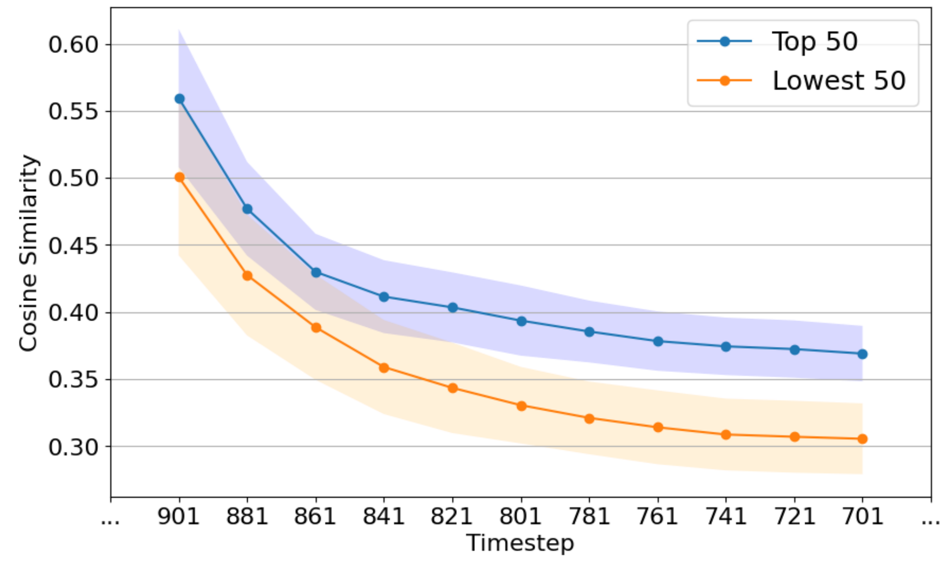
▶ Discriminator Generalization

- We perform **cross-dataset inference** using a discriminator trained on a different dataset.
- It shows improved performance, demonstrating **generalization** ability of the discriminator.

Source Model for Fake Data	AD (Defualt)	Lavie	VC2
Subject Consistency	0.9528	0.9625	0.9535
Background Consistency	0.9763	0.9753	0.9764
Temporal Flickering	0.9258	0.9490	0.9283
Motion Smoothness	0.9599	0.9691	0.9617
Dynamic Degree	0.4125	0.4088	0.4100
Overall Consistency	0.2529	0.2473	0.2509

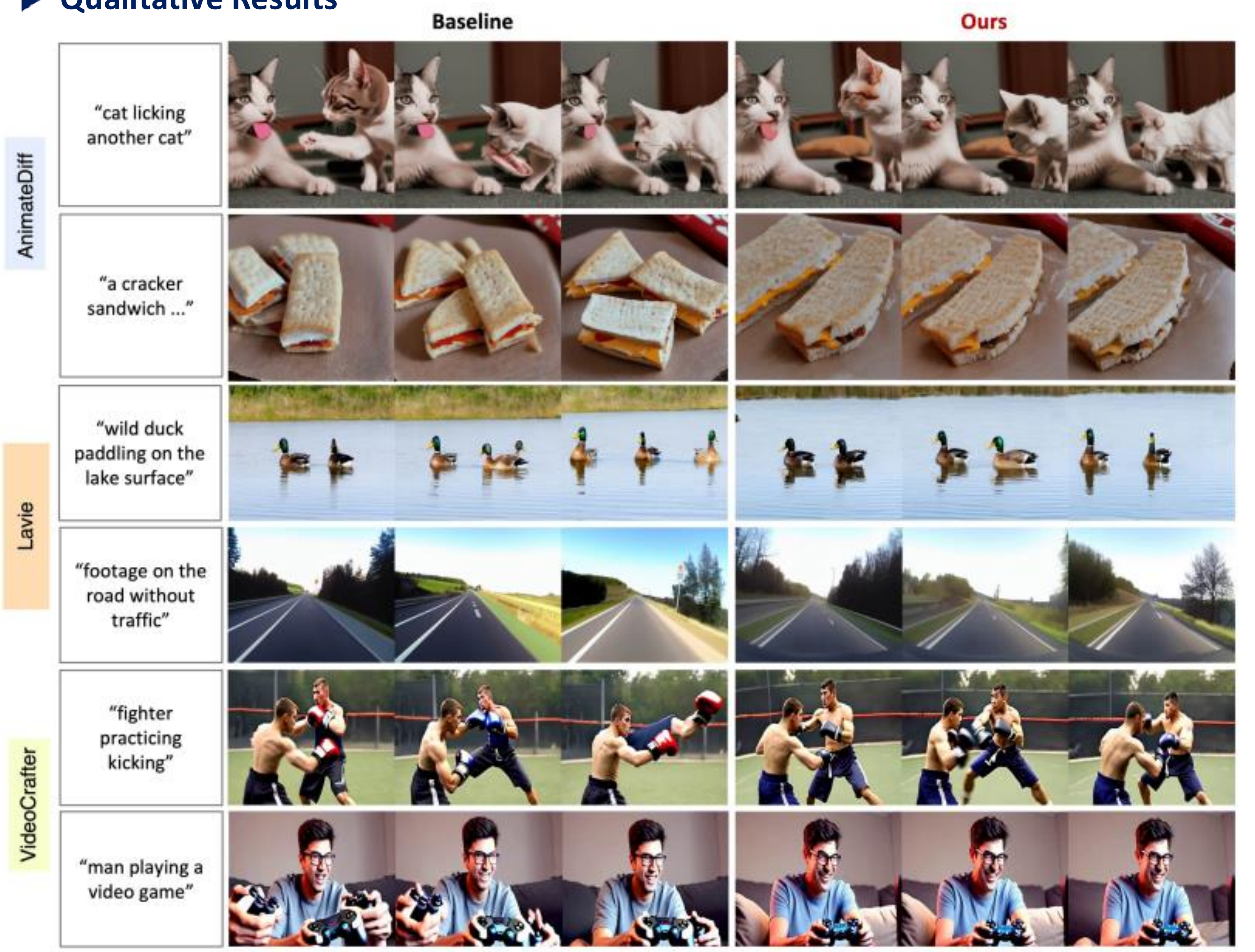
▶ Token Variations

- We measure **cosine similarity** between initial and optimized token embeddings over time.
- It shows **similarity decreases and stabilizes;** high-consistency videos show less token shift, indicating reduced optimization when judged as realistic.



Results

▶ Qualitative Results



▶ Quantitative Results

Method	Temporal Quality					Text Alignment
	Subject Consistency (↑)	Background Consistency (↑)	Temporal Flickering (↑)	Motion Smoothness (↑)	Dynamic Degree (↑)	Overall Consistency (↑)
Lavie [36]	0.9599	0.9739	0.9487	0.9690	**0.5150**	**0.2506**
Lavie + **Ours**	**0.9646**	**0.9781**	**0.9625**	**0.9765**	0.3963	0.2415
AnimateDiff [14]	0.9488	0.9755	0.9228	0.9578	**0.4700**	**0.2532**
AnimateDiff + **Ours**	**0.9528**	**0.9763**	**0.9258**	**0.9599**	0.4125	0.2529
VideoCrafter2 [4]	0.9736	0.9559	0.9559	**0.9750**	**0.4088**	**0.2498**
VideoCrafter2 + **Ours**	**0.9745**	**0.9774**	**0.9588**	0.9759	0.3938	0.2451

▶ User Study

Baseline	Win	Tie	Lose
AnimateDiff	66.5	17.8	15.7
Lavie	55.1	21.1	23.8
VideoCrafter2	53.0	17.7	29.3

▶ Extensions to I2V

Input prompt and image