# S²V2V: Training-Free Video-to-Video with Sparse Points and Motion Guidance

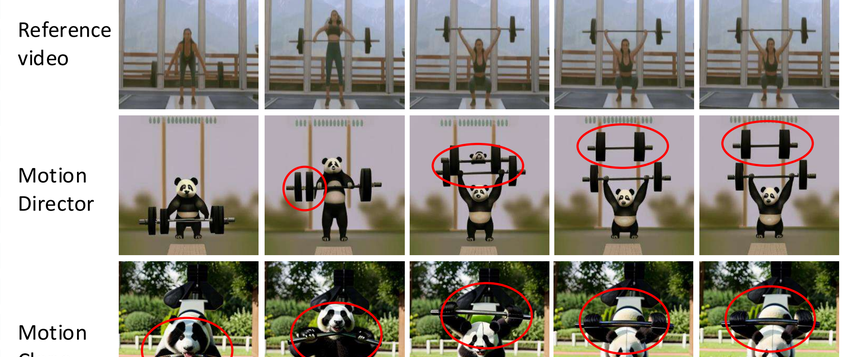
Xinyu Zhang[1], Zicheng Duan[2], Dong Gong[3], Lingqiao Liu[2]
[1] University of Auckland, New Zealand [2] The University of Adelaide, Australia [3] University of New South Wales, Australia

Project website: https://zhangxinyu-xyz.github.io/S2V2V.github.io/

## Introduction

**Motion-based controllable video-to-video generation**

- Content from text descriptions
- Motion from given reference videos




## Motivation

**Related works:**

i) Training-based strategy: various guidance types, such as reference videos, trajectories, bounding box movements, or motion fields, as control signals; LoRA, adapters or modules for fine-tuning.

*Limitations*: inflexibility, needing to re-training to adapt to new foundation models.

ii) Training-free strategy: incorporating motion guidance through test-time operations, such as specific noise schemes or operations on attention maps

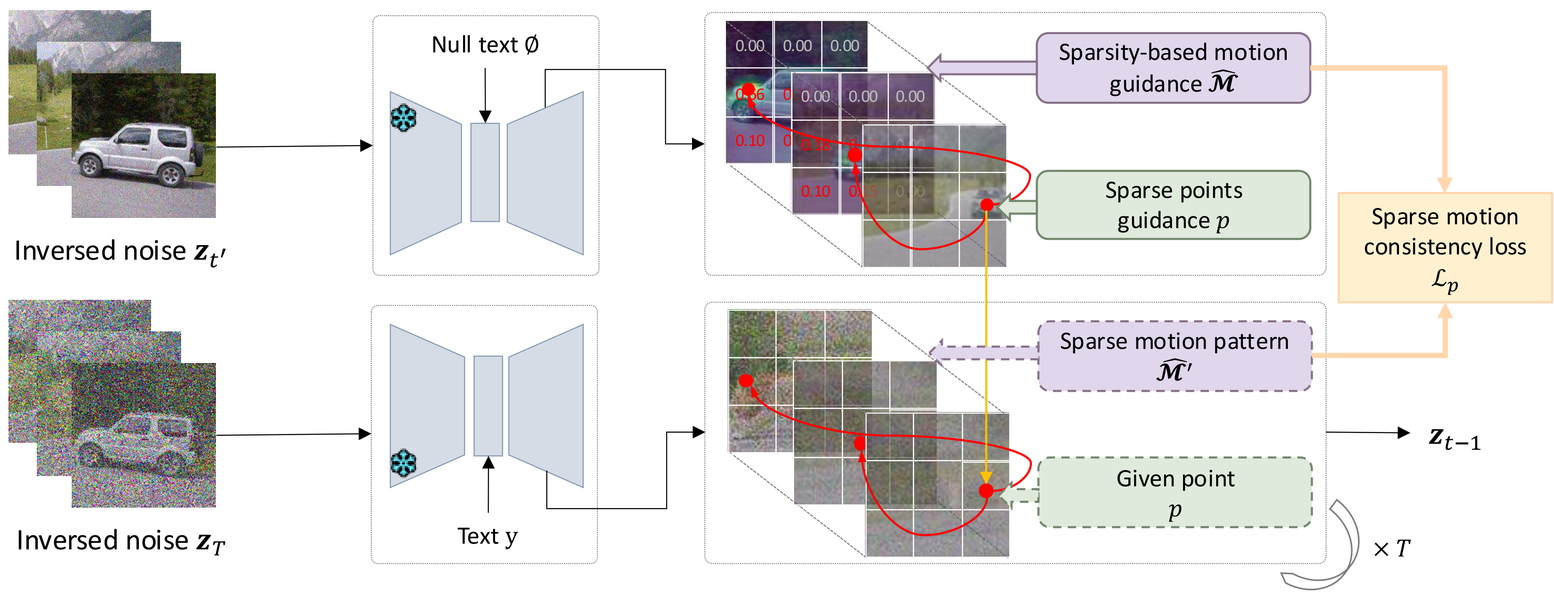
*Limitations*: inconsistent temporal coherence across frames




**Motivation:**

Explicitly captures temporal consistency across frames to accurately transfer motion patterns from reference videos to target ones during inference, ensuring both flexibility and coherence.

## Method: S²V2V




**Overview:**

S²V2V is a training-free method to use inter-frame correlation patterns with double sparse guidance as the motion guidance, *i.e.*, sparse points guidance and the sparsity-based motion guidance, to produce more robust and accurate correlation patterns.

**Double Sparse Guidance:**

**Sparse point guidance (SP):**

A sparse set of points of interest on reference videos, coming from user interaction (*i.e.*, manual clicks), centroids of bounding box trajectories, detector-based tracking, or random selection.

Step1: Given one point $p$ and extract its feature $\boldsymbol{f}$

Step2: Calculate the correlation patterns between each point in other frames $\boldsymbol{f}_{(i,j,k)}$ and $\boldsymbol{f}$

Step3: Obtain correction motion pattern

$$\mathcal{M} = \{\mathcal{M}_{f+1}, \mathcal{M}_{f+2}, ..., \mathcal{M}_F\}$$

$$\mathcal{M}_i(j,k) = \frac{\exp(\mathrm{sim}(\boldsymbol{f} \cdot \boldsymbol{f}_{(i,j,k)})/\tau)}{\sum_{h=1}^{H}\sum_{w=1}^{W}\exp(\mathrm{sim}(\boldsymbol{f} \cdot \boldsymbol{f}_{(i,h,w)})/\tau)}$$

*Remark*: SP captures *relative correlations* between a given point and other points in subsequent frames, reflecting the strength of their spatial-temporal relationships.

Higher matching score ↔ Higher likelihood of the position that the point will move to next

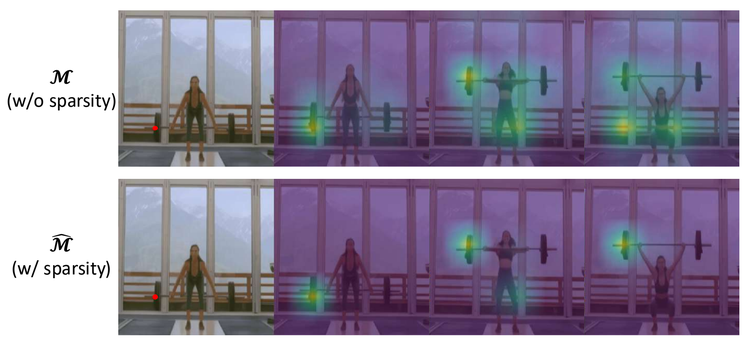
**Sparsity-based motion guidance (SM):**

Address the occasionally inaccurate inter-frame correlations, particularly when the temporal distance between frames is large.

$$\widehat{\mathcal{M}}_i(j,k) = \begin{cases} \mathcal{M}_i(j,k), & \text{if } (j,k) \in \Omega_i, \\ 0, & \text{otherwise.} \end{cases}$$

$\Omega_i$ denotes the local region around the tracked point.

$$\widehat{\mathcal{M}}_i(j,k) = \widehat{\mathcal{M}}_i(j,k) / ||\widehat{\mathcal{M}}_i||_1, \forall j, \forall k$$




**Sparse motion consistency loss:**

$$\mathcal{L}_p = \sum_{f=1}^{F}\sum_{i=f+1}^{F} ||\widehat{\mathcal{M}'}_i - \widehat{\mathcal{M}}_i||_2^2$$

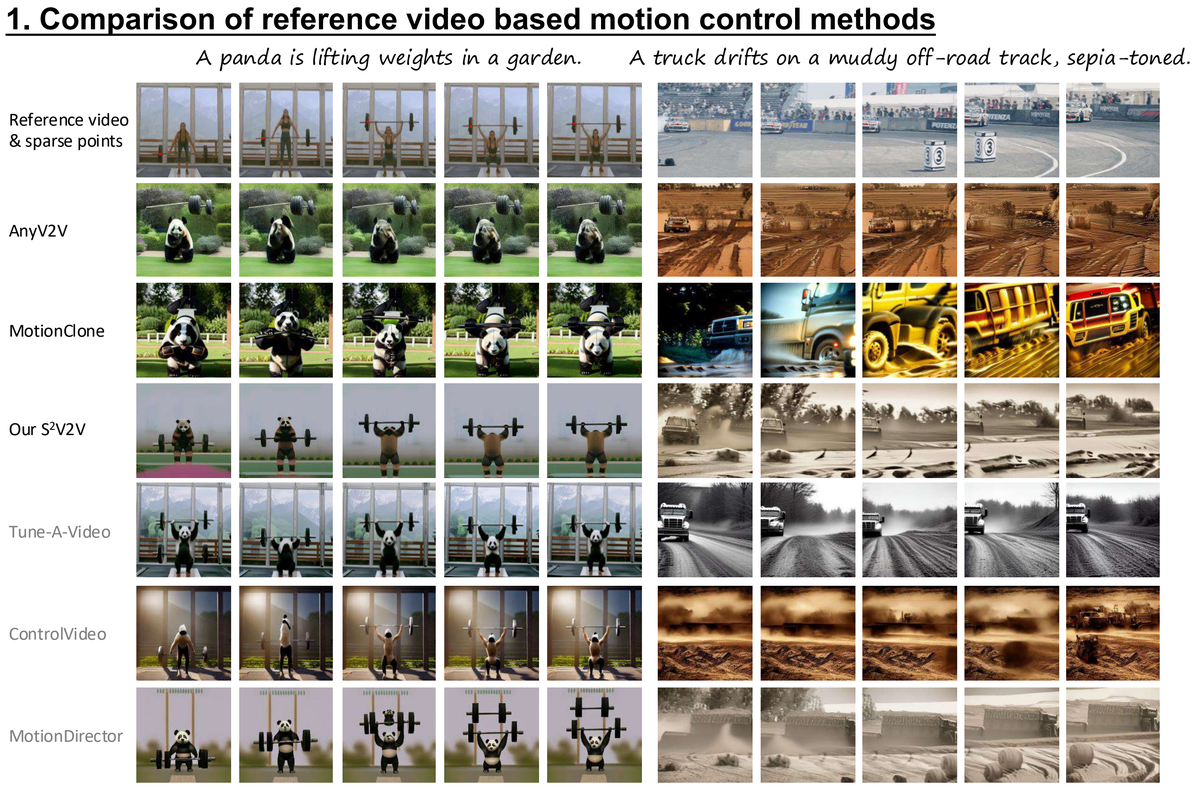
$\widehat{\mathcal{M}'}$ denotes the predicted motion correlation.

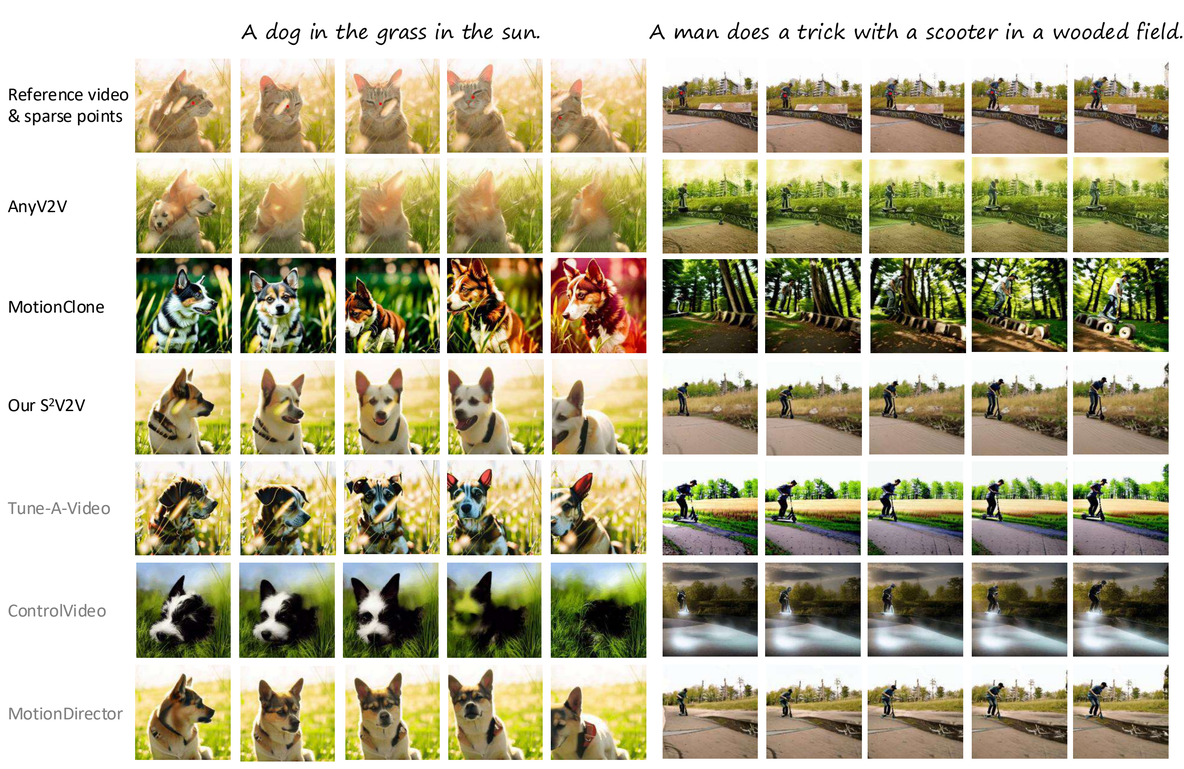
**Optimization:**

$$\hat{\varepsilon}_\theta(\mathbf{z}_t, t, y) := \varepsilon_\theta(\mathbf{z}_t, t, y) + \sigma_t \nabla_{\mathbf{z}_t} \mathcal{L}_p(\mathbf{z}_t)$$

## Experiments

### 1. Comparison of reference video based motion control methods

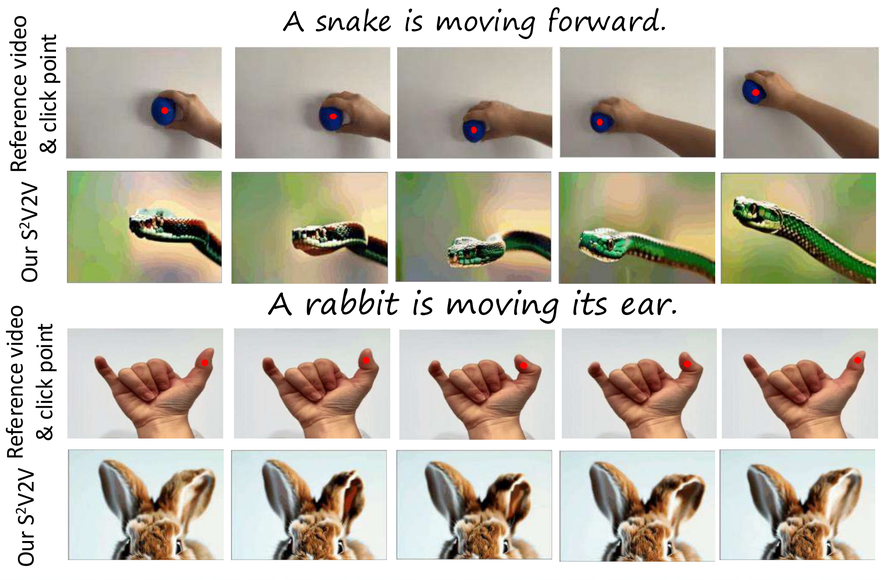



### 2. Comparison of trajectory control methods


| Method | Automatic Evaluations CSIM (↑) | CSIM-GT (↑) | mIoU (↑) | Human Evaluations Text Align | Trajectory Align | Video Quality |
|---|---|---|---|---|---|---|
| Peekaboo [27] | 0.942 | 0.869 | 0.143 | 24.55 *vs.* 75.45 | 10.83 *vs.* 89.17 | 30.13 *vs.* 69.87 |
| FreeTraj [52] | **0.951** | 0.886 | 0.268 | 48.77 *vs.* 51.23 | 46.99 *vs.* 53.01 | 42.97 *vs.* 57.03 |
| Ours | 0.947 | **0.889** | **0.272** | - | - | - |

### 3. Gesture simulation




Code & Model

Contact me