Frame In-N-Out: Unbounded Controllable Image-to-Video Generation

Boyang Wang[1] Xuweiyi Chen[1] Matheus Gadelha[2] Zezhou Cheng[1]

University of Virginia[1] Adobe Research[2]

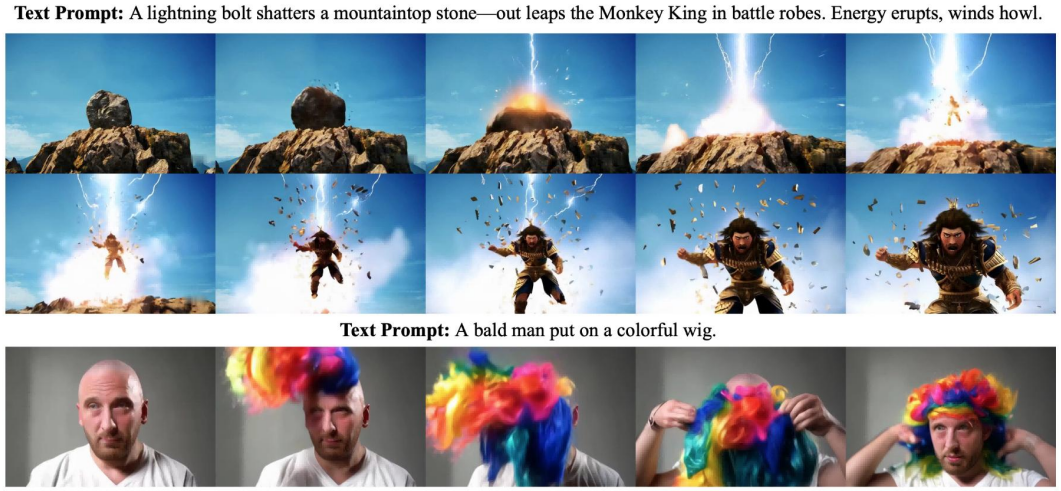
Keyword: **Video Generation**, **Motion Control**, **Identity Reference Control, Unbounded Canvas**

Background

Closed and Open-Source Models like SoRA, CogVideoX, and Wan has achieved marvelous generation quality for Text-to-Video (T2V) and Image-to-Video (I2V) generation tasks.

However, their conditional control is largely depending on **text prompt** or the **first frame** inputs. It is hard to utilize sparse abstractive language information to fully unleash the creativity of the human.

CogVideoX

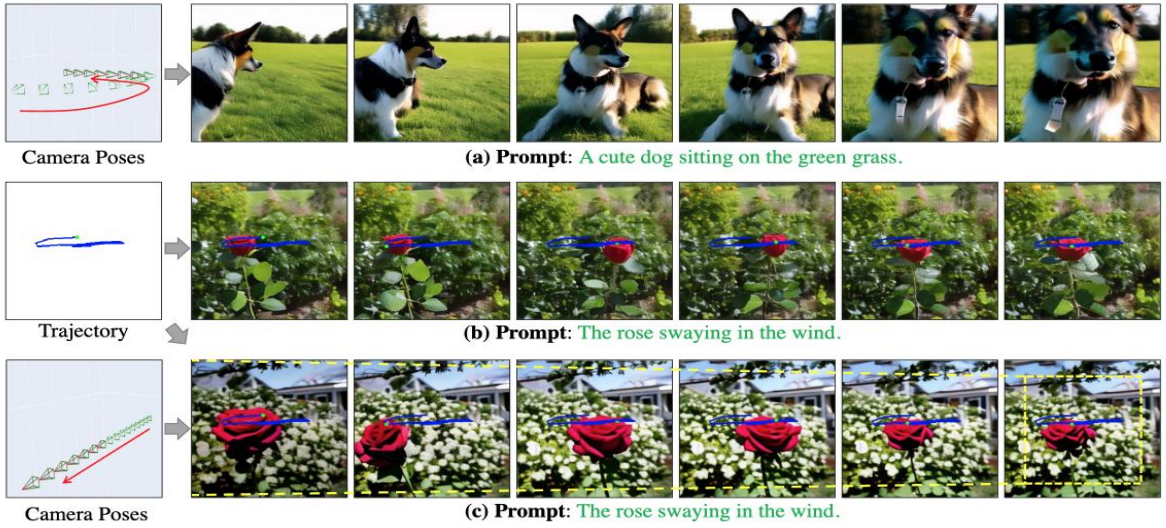


SoRA

Trajectory-Based Control on Objects

Papers like **Motion Control** and **ToRA** apply trajectory lines to control the object to move in the way designated. This method exerts convenient controllability for the user intention. Theis is achieved by Supervised Finetune (SFT) on the base model like SVD and CogVideoX.
However, these motion signal should be **pixel aligned** with the original **first frame condition**.

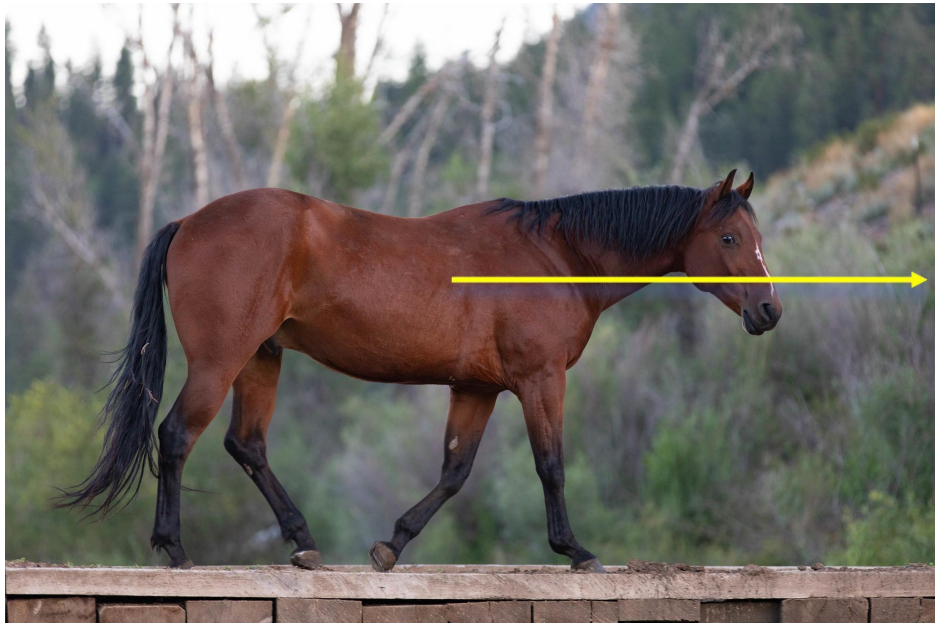
Motion Control



ToRA

Limitation in Previous Work

What if we want the object to move completely out of the scene with the designated trajectroy?
This is termed **Frame-Out** in the cinematic technique.



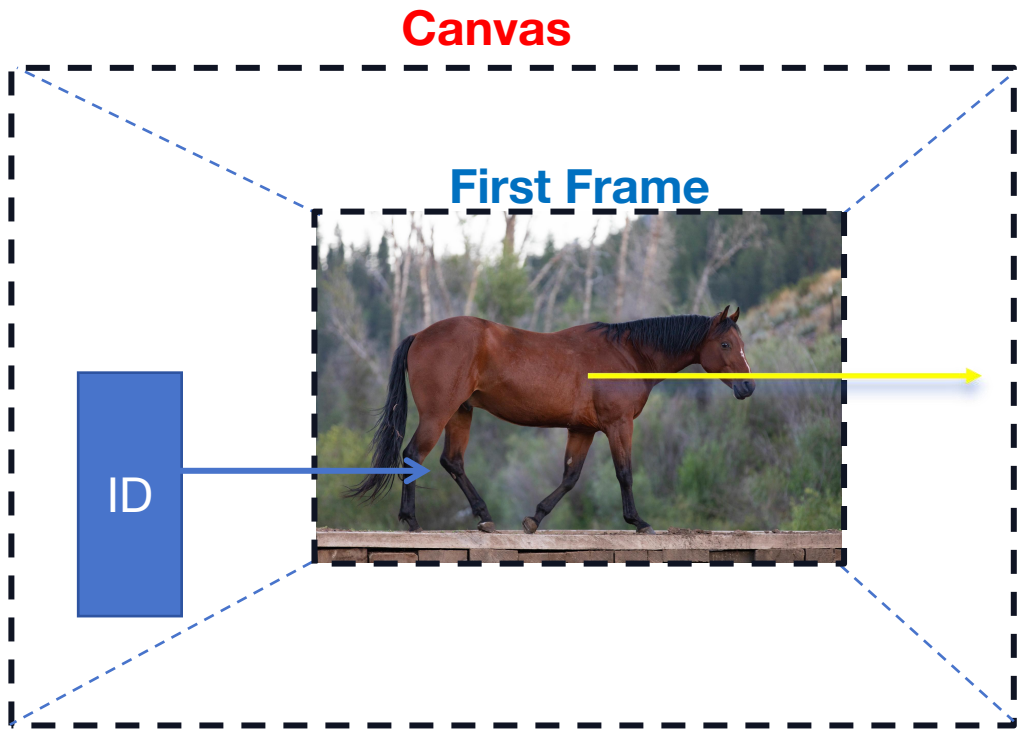
The Motion Trajectory is Completely Bounded Inside the first-frame region.
The User Intension is hard to be spread to **unseen** region.

Limitation in Previous Work

What if we want a Breaking New Object (Identity Reference) with the motion trajectory to enter the scene?
This is termed **Frame-In** in the cinematic domain.

Proposed Method

We provide **Controllable Image-to-Video generation** with an Unbounded Border concept.
The control signal can get over the border of the first frame and also a breaking new ID can be introduced.
In this process, we unify **pixel-aligned** (Motion Traj) and **pixel-unaligned** (ID) conditioing in I2V generation.



This paper will achieve both Frame-In and Frame-Out pursuit, and we term it as Frame In-N-Out.

Data Curation

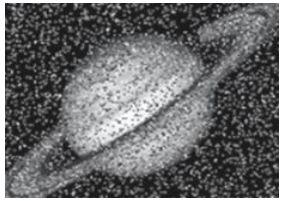
First, we aggregate image and video-level filtering criteria to filter **Million-Scale** datasets from **WebVid**, **OpenVid**, and **VidGen**.

Basic Curation

(a) Metadata filtering

[duration, resolution, fps, *etc*.]

(b) Image-level Filtering



Vision is about solving an amazing
capable of perceiving its surrounding
available in the small number of ele
the observer's sensor (eye or camera),
objects surrounding the observer? Ho
and its three-dimensional (3D) struct
quite likely that if humans did not h
think that solving this problem is imp

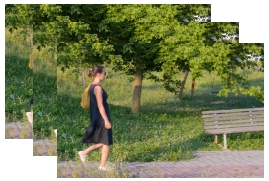
Quality & Complexity & Aesthetic & OCR

(c) Video-level Filtering

Scene Cut

CUT3R → Camera Motion Estimation

(d) Video Captioning



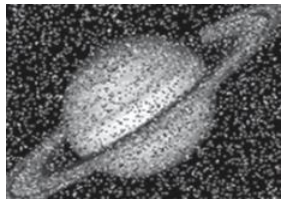
"A woman in a black dress walks through a park on a sunny day, passing an empty bench as green trees sway in the background."

Basic Curation

(a) Meta data filtering

[duration, resolution, fps, *etc*.]

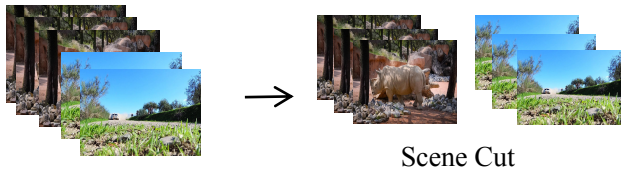
(b) Image-level Filtering



Vision is about solving an amazing
capable of perceiving its surrounding
available in the small number of ele
the observer's sensor (eye or camera),
objects surrounding the observer? Ho
and its three-dimensional (3D) struct
quite likely that if humans did not h
think that solving this problem is imp

Quality & Complexity & Aesthetic & OCR

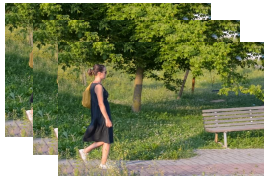
(c) Video-level Filtering



Scene Cut

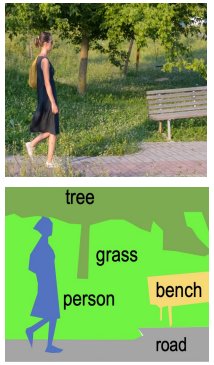
CUT3R → Camera Motion Estimation

(d) Video Captioning

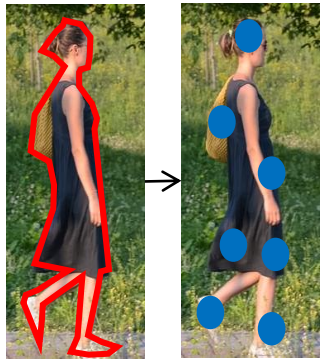


"A woman in a black dress walks through a park on a sunny day, passing an empty bench as green trees sway in the background."

Identity of Interest Filtering



Person Bench



Panoptic Segmentation & Movable Object Selection

Point Sampling via K-means Clustering

Robust Motion Trajectory Generation

- $\|A - B\| <$ threshold $\Longrightarrow$
- $\|A - B\| \geq$ threshold $\Longrightarrow$

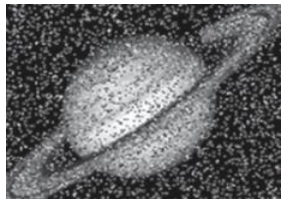
Next, We focus on the Motion-related filtering considering the **Frame In-N-Out** needs.

Basic Curation

(a) Meta data filtering

[duration, resolution, fps, *etc*.]

(b) Image-level Filtering



Vision is about solving an amazing
capable of perceiving its surrounding
available in the small number of ele
the observer's sensor (eye or camera),
objects surrounding the observer? Ho
and its three-dimensional (3D) struct
quite likely that if humans did not h
think that solving this problem is imp

Quality & Complexity & Aesthetic & OCR

(c) Video-level Filtering

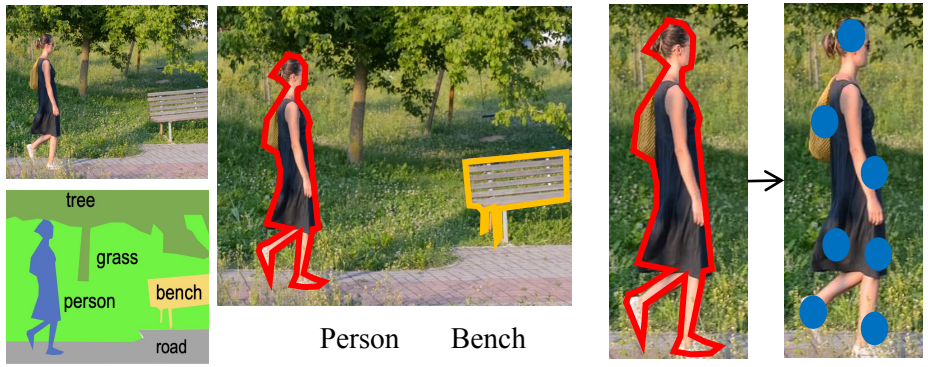
Scene Cut

CUT3R → Camera Motion Estimation

(d) Video Captioning

"A woman in a black dress walks through a park on a sunny day, passing an empty bench as green trees sway in the background."

Identity of Interest Filtering



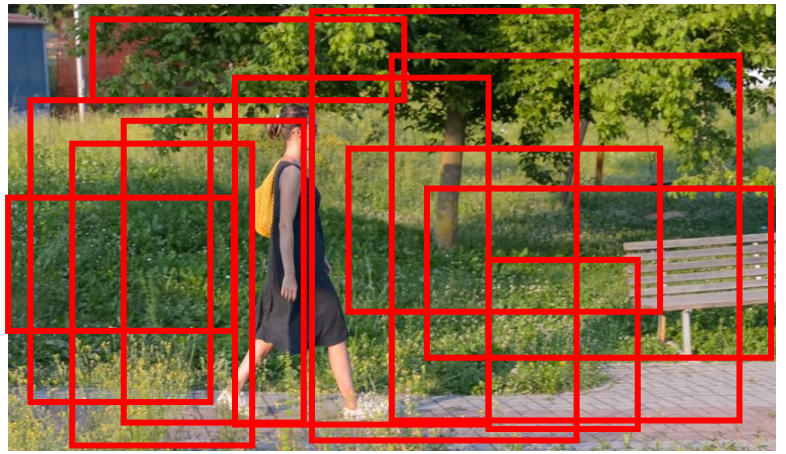
Panoptic Segmentation & Movable Object Selection

Point Sampling via K-means Clustering

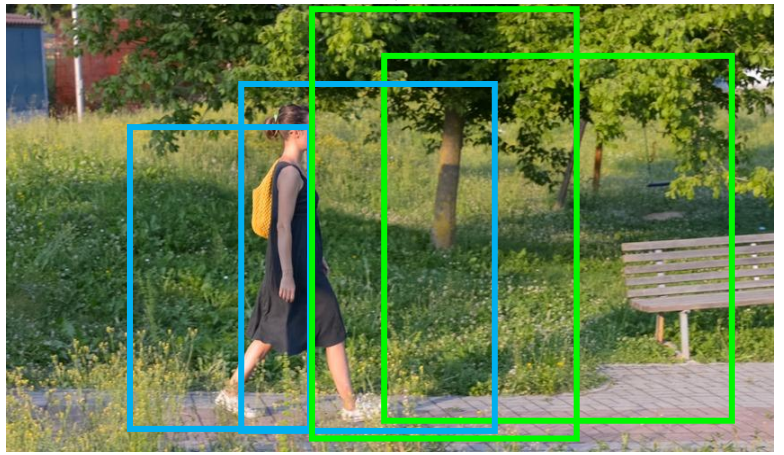
Robust Motion Trajectory Generation

- $\|A - B\| <$ threshold $\Longrightarrow$
- $\|A - B\| \geq$ threshold $\Longrightarrow$

Frame In & Frame Out



Region Proposal Generation



Frame In Frame Out

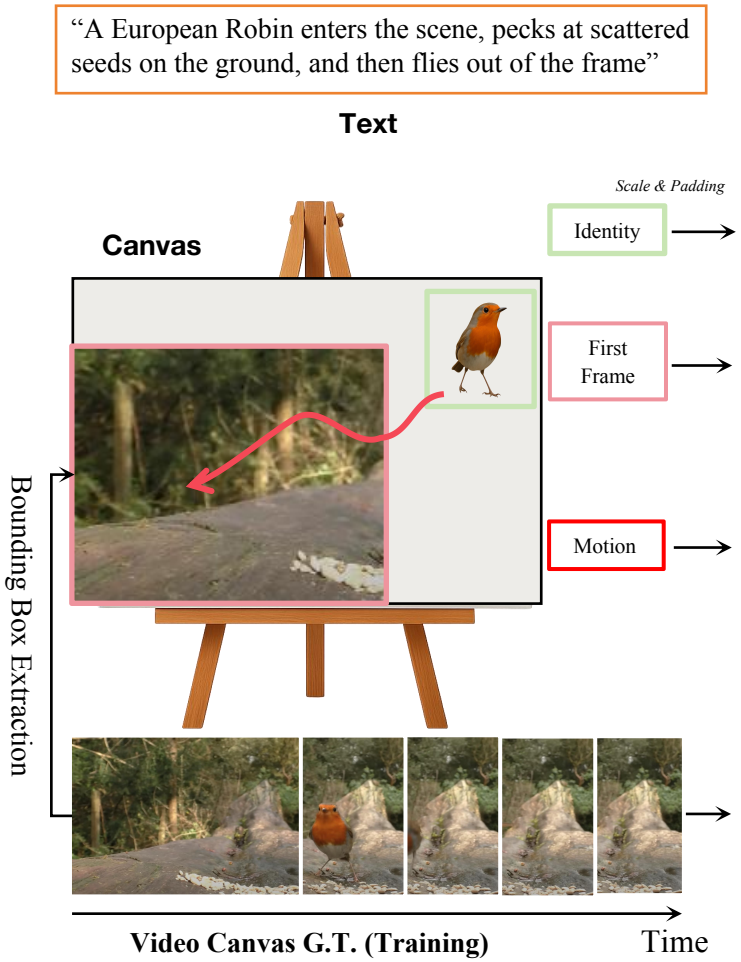
Architecture

Our Model started from a relative light-weight **Diffusion Transformer**, CogVideoX-5B-I2V.

Training is composed of two stage:

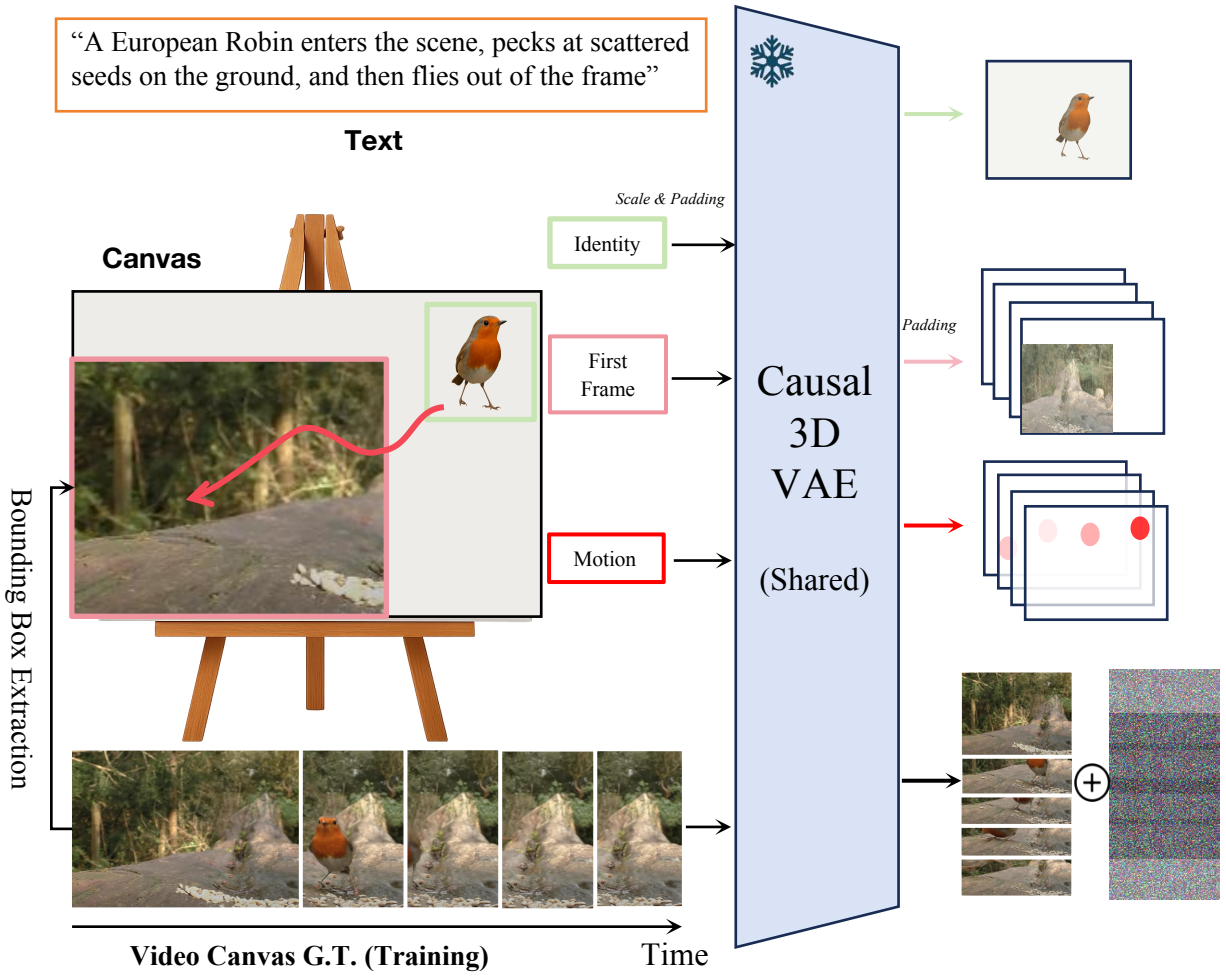
(1) Motion Control Training

(2) **Unbounded Training with ID reference**



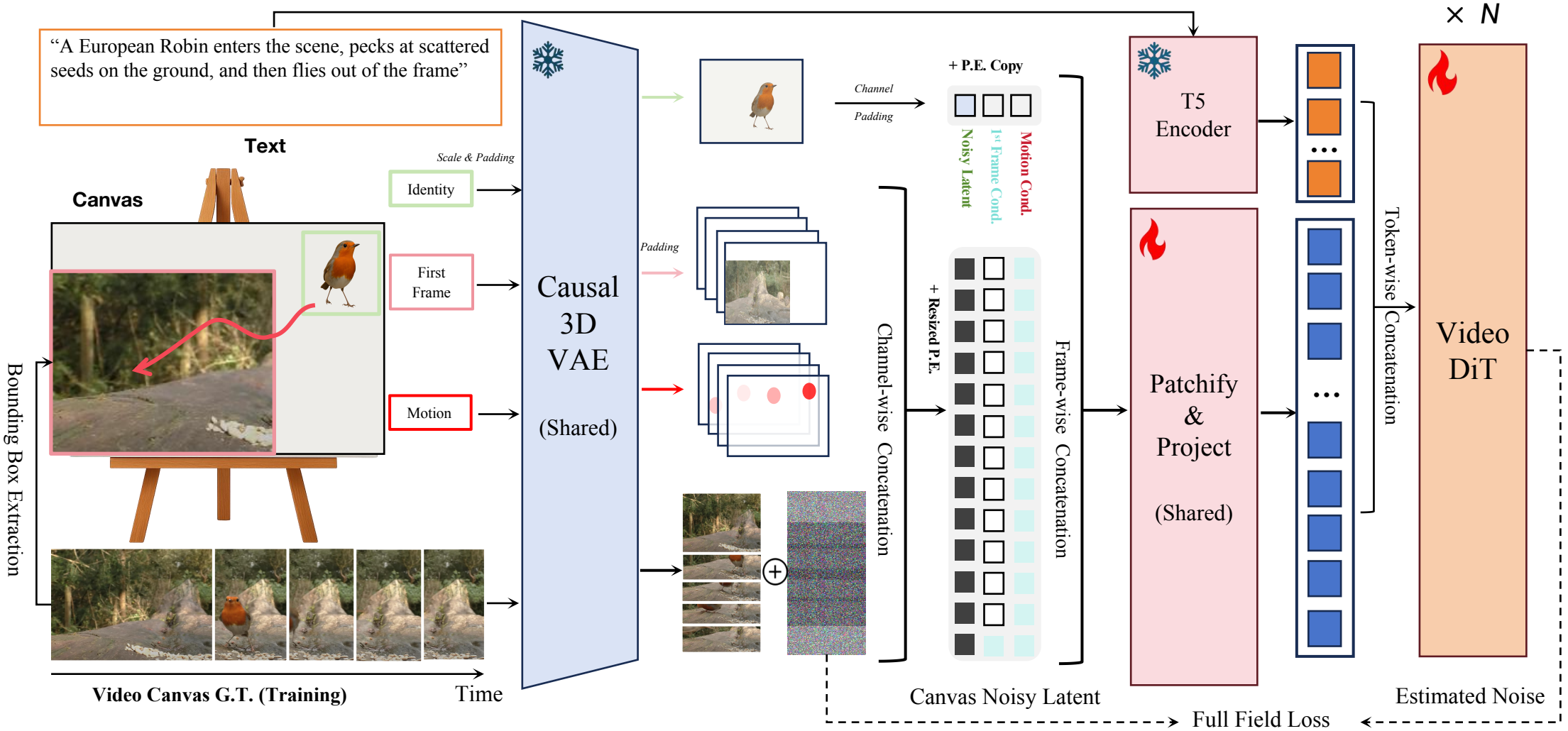
In total, we have 5 conditioning in the model:

Text Prompt
First Frame
Motion Trajectroy
Identity Reference
Canvas Border Setting



Encoded via a shared Causal 3D VAE for both image and temporal condition

Channel-Wise Concatenation for Pixel-aligned Condition.
Frame-Wise Concatenation for Pixel-unaligned Condition.
Loss is applied to the full canvas field.



First Frame

First Frame

Expand

Canvas

First Frame

Expand

Canvas

+

Identity Reference

First Frame

Expand

Canvas

+

Identity Reference

Motion

First Frame

Expand

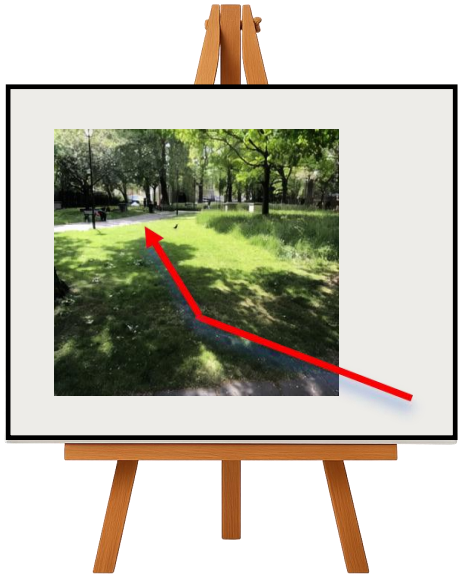
Canvas

Identity Reference

Motion

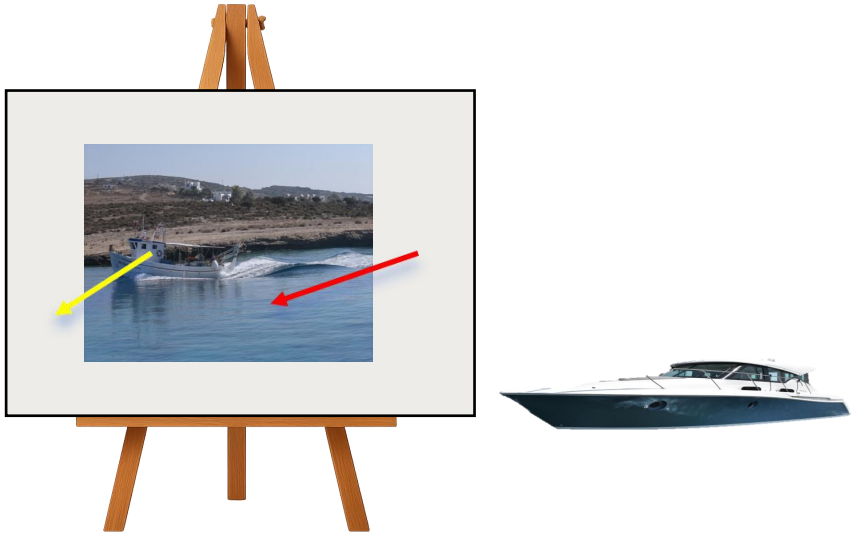
Text Prompt

The lovely puppy is running into the grass happily. It then walks further away.



Text: The lovely puppy is running into the grass happily. It then walks further away.

Text: *The man is walking out of the scene.*

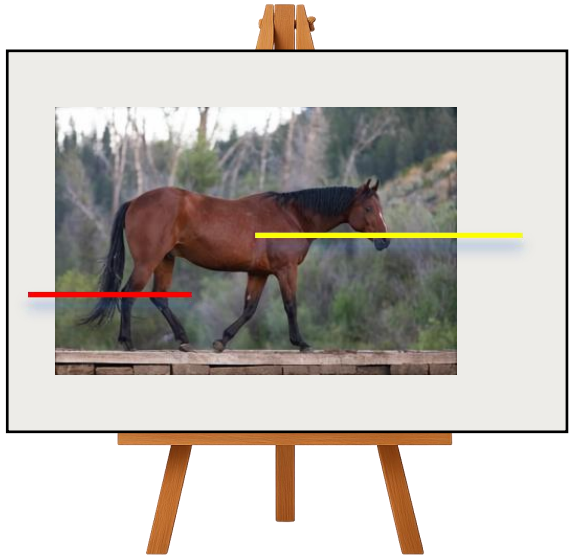


Text: *The fishing boat in the image moving towards the left side of the frame and the leave the frame. Another boat is driving from the right of the scene to the center.*

Text: *The balloon is flying to the left and then adjust back.*

Text: *The yellow ball hit the white ball.*

Text: *A hand grab the cup handle and then take it away.*



Text: *The hose is walking to the right and then a sheep enters from the left.*

Thanks for Listening!