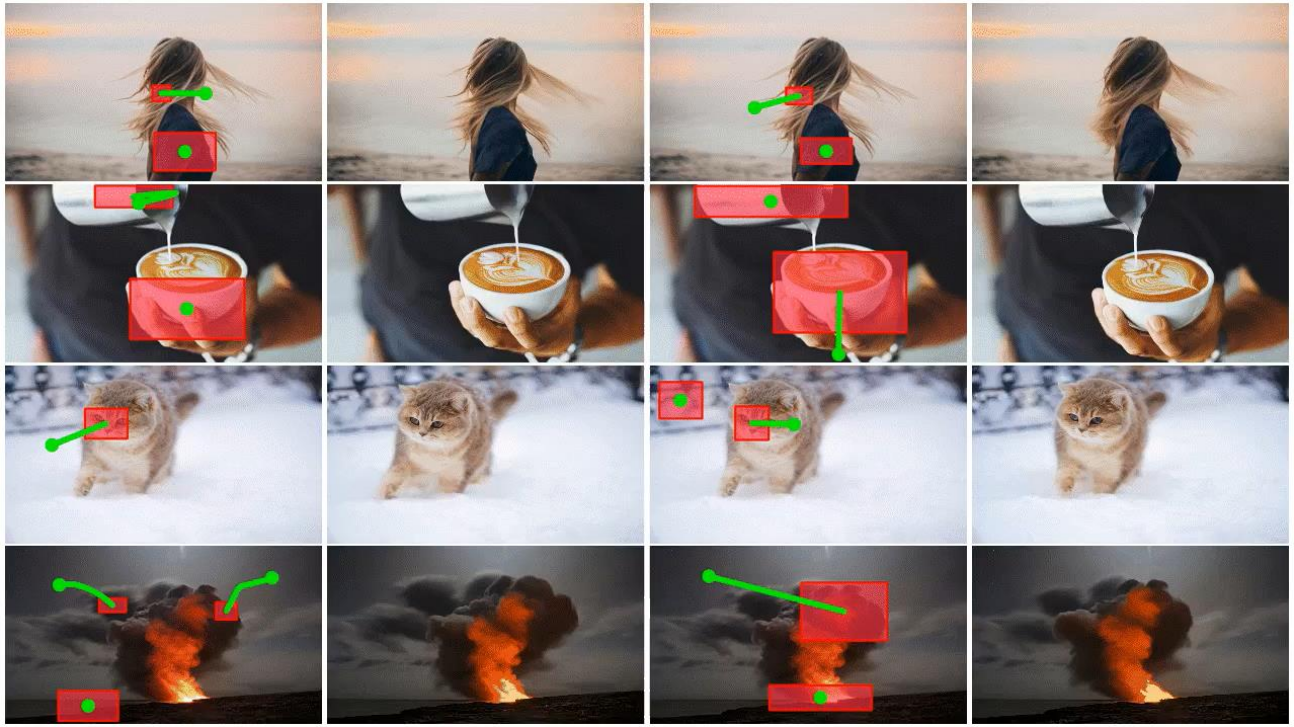
SG-I2V: Self-Guided Trajectory Control in Image-to-Video Generation

Koichi Namekata, Sherwin Bahmani, Ziyi Wu, Yash Kant, Igor Gilitschenski, David B. Lindell

ICLR, 2025



01 Problem Definition

Trajectory-conditioned image-to-video generation

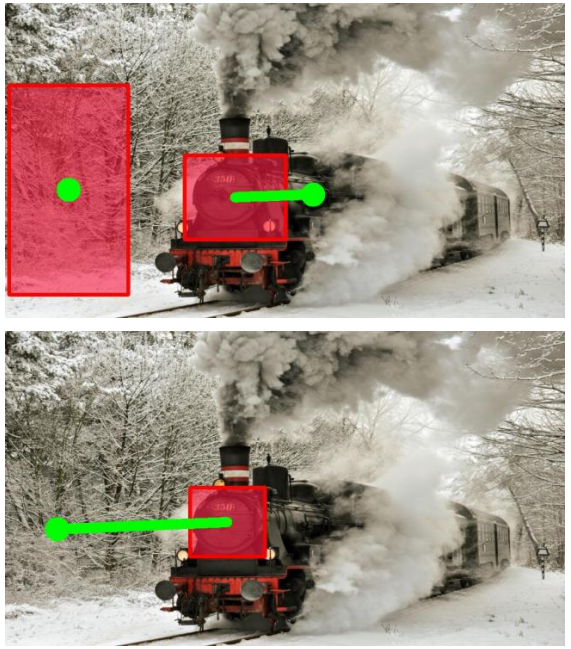
02

Problem Definition

Trajectory-conditioned image-to-video generation

Input

Image,
Bounding boxes,
Trajectories

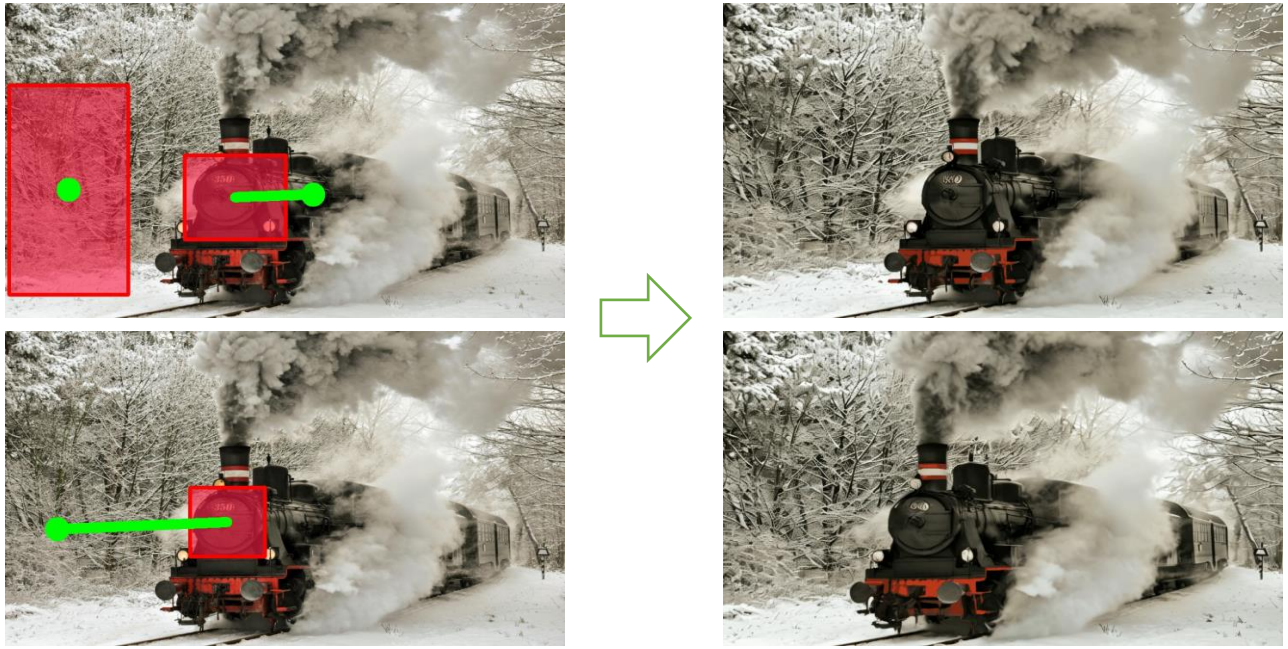


03 Problem Definition

Trajectory-conditioned image-to-video generation

Input

Image,
Bounding boxes,
Trajectories



Output

Videos following
the trajectories

04 Method

Build on pre-trained image-to-video diffusion models (Stable Video Diffusion)

05 Method

Build on pre-trained image-to-video diffusion models (Stable Video Diffusion)

- Previous work:
  - ✗ Computationally expensive finetuning
  - ✗ Require motion-annotated dataset collection

06 Method

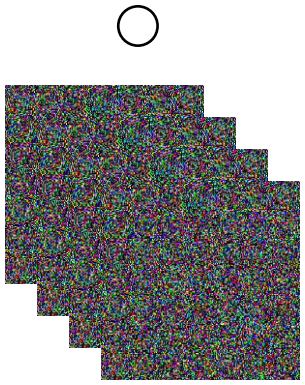
Build on pre-trained image-to-video diffusion models (Stable Video Diffusion)

- Previous work:
  - ❌ Computationally expensive finetuning
  - ❌ Require motion-annotated dataset collection

- This work:
  - ✓ No finetuning
  - ✓ Relies solely on the knowledge present in the pre-trained image-to-video diffusion models.

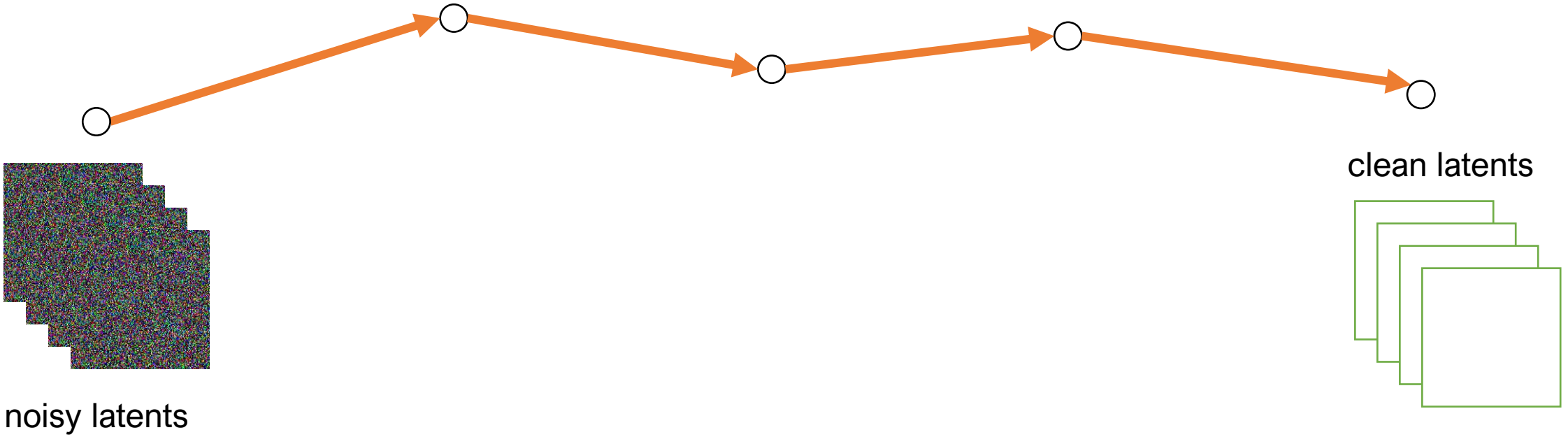
07 Review: image-to-video generation with diffusion models

08 Review: image-to-video generation with diffusion models

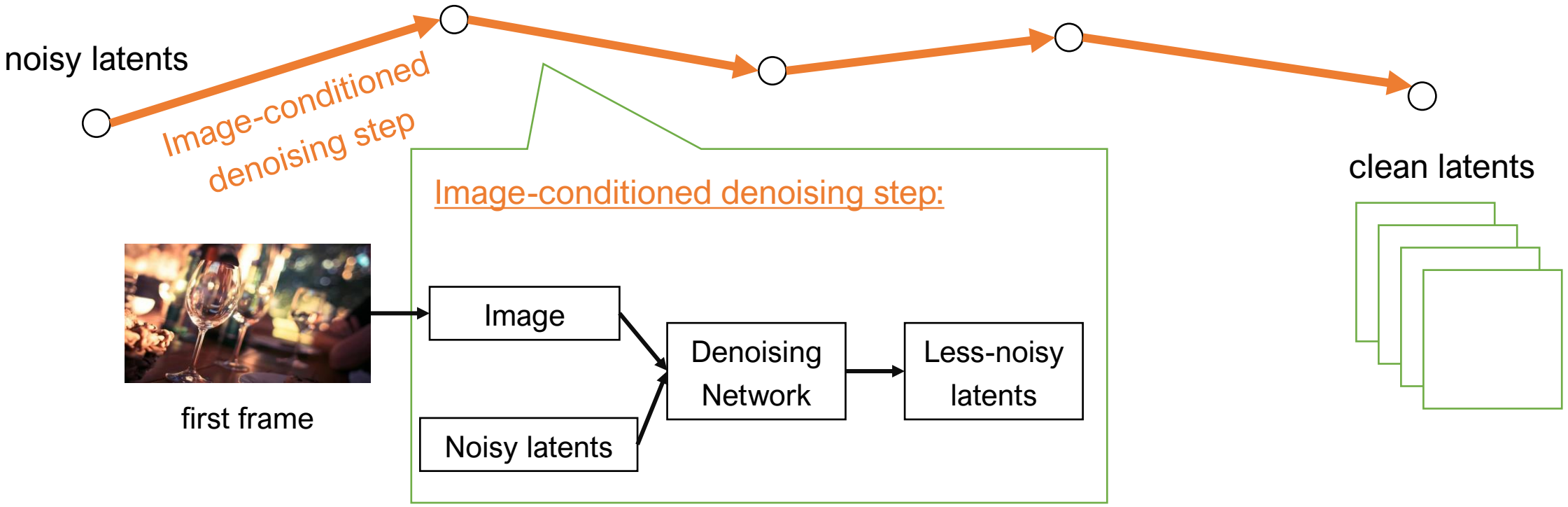


noisy latents

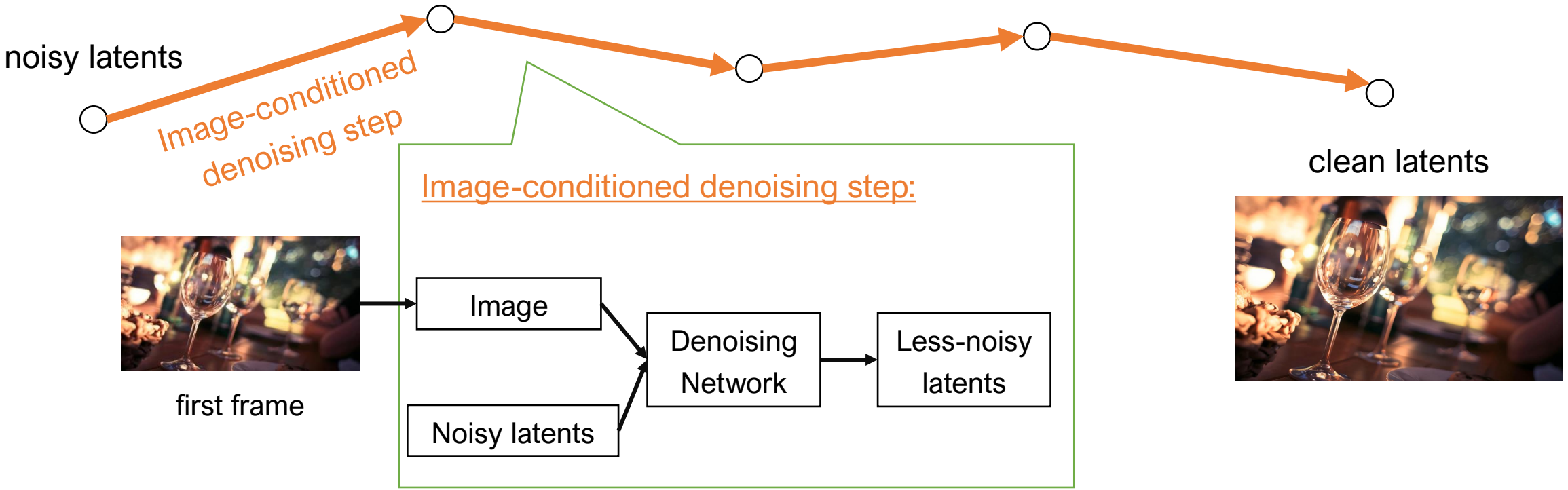
09 Review: image-to-video generation with diffusion models



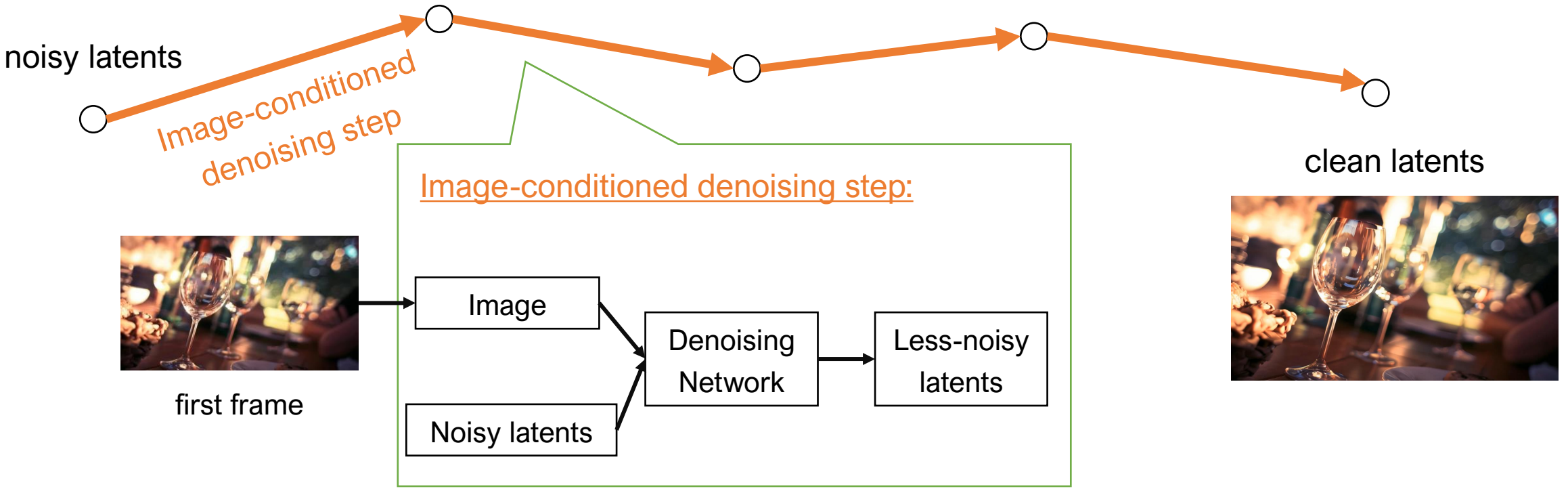
Review: image-to-video generation with diffusion models



Review: image-to-video generation with diffusion models

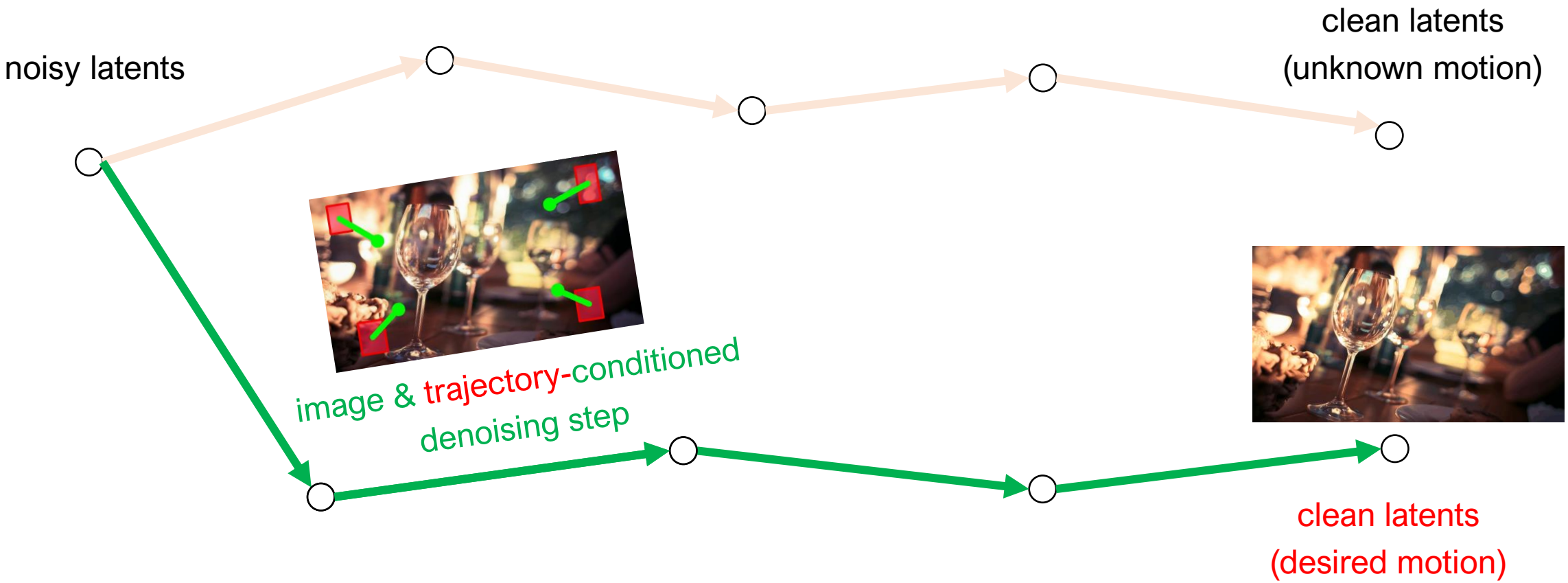


Review: image-to-video generation with diffusion models



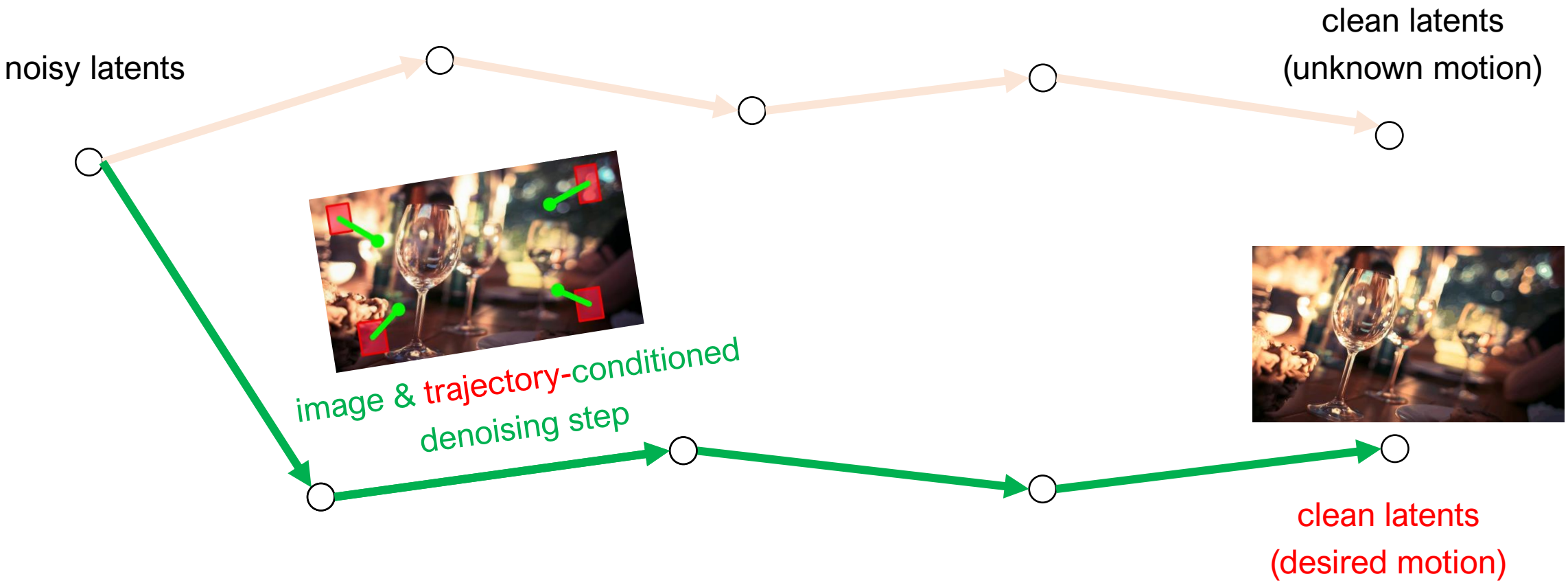
× We do not have control over the motions of generated videos

Previous work

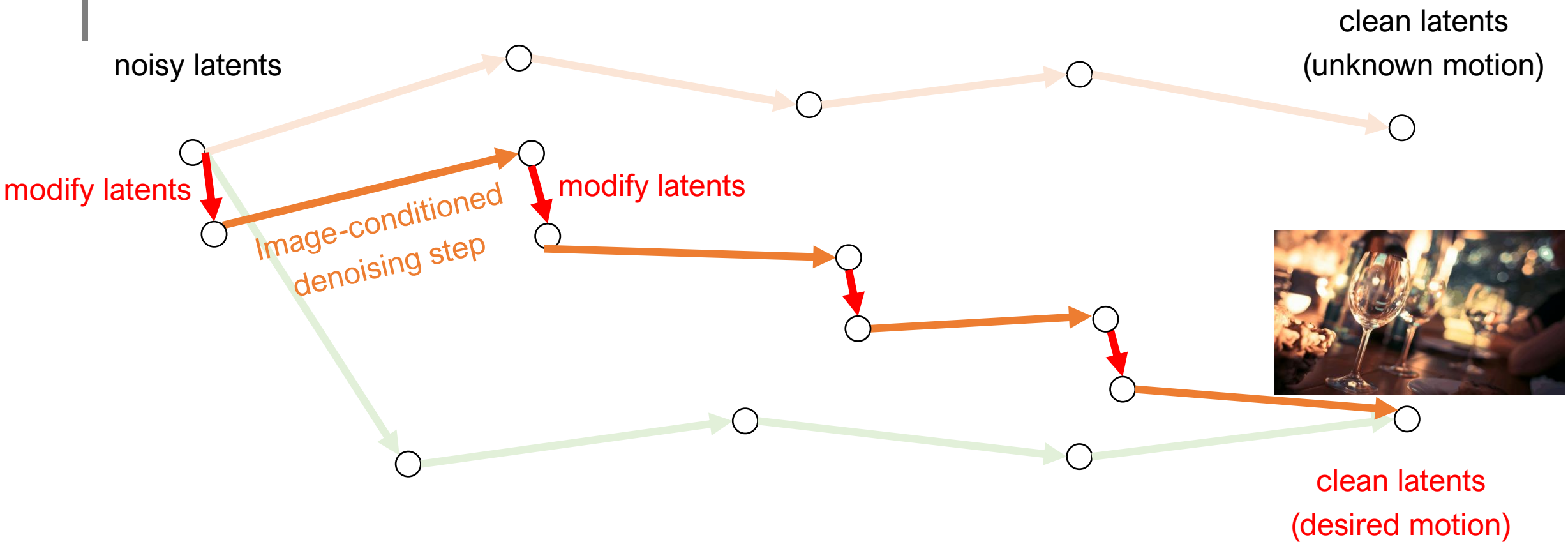


Previous work

× We need to finetune the denoising network



Ours


noisy latents
clean latents
(unknown motion)
modify latents
modify latents
Image-conditioned
denoising step
clean latents
(desired motion)

How do we modify latents?

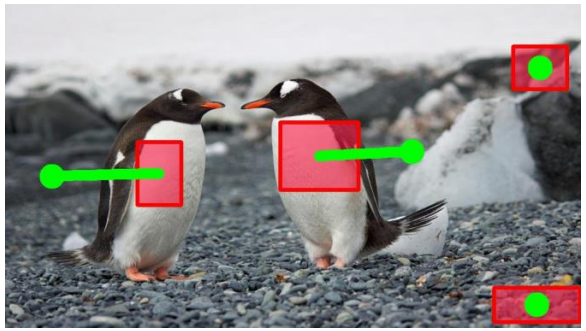
Internal feature maps
(PCA visualization)

Generated videos

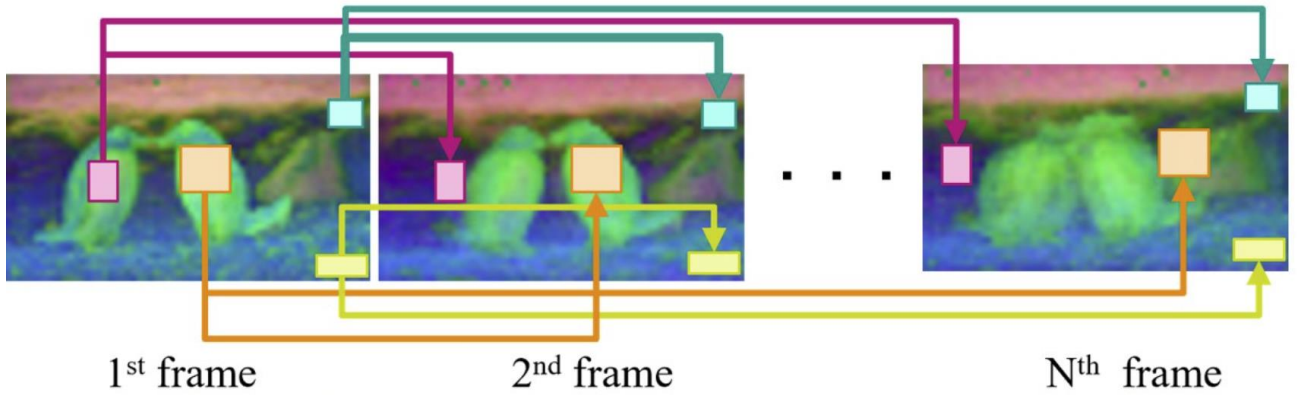
Motion correspondences !

How do we modify latents?

- Design loss function that encourages feature similarity within each bounding box along the trajectory.



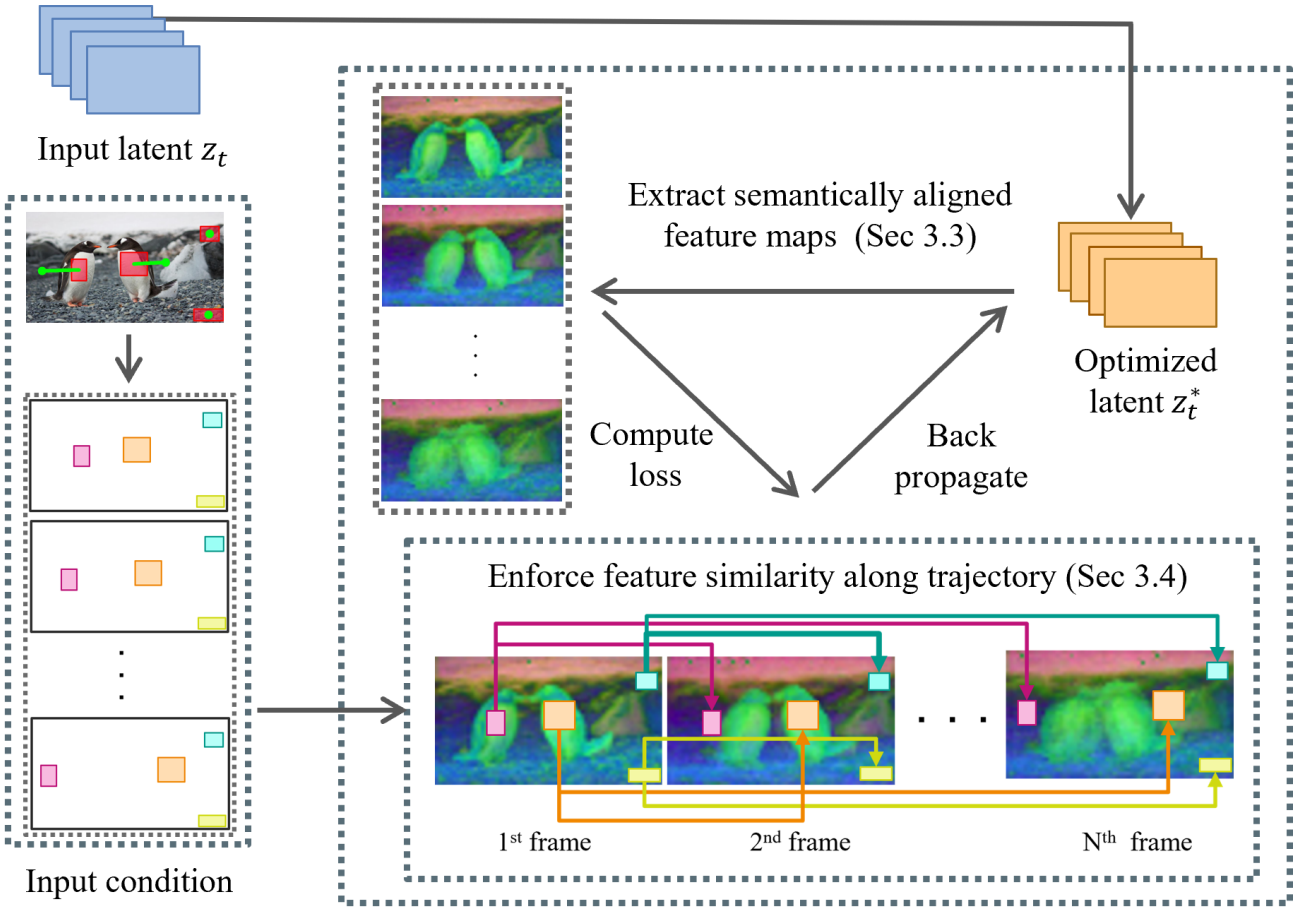
Target trajectory

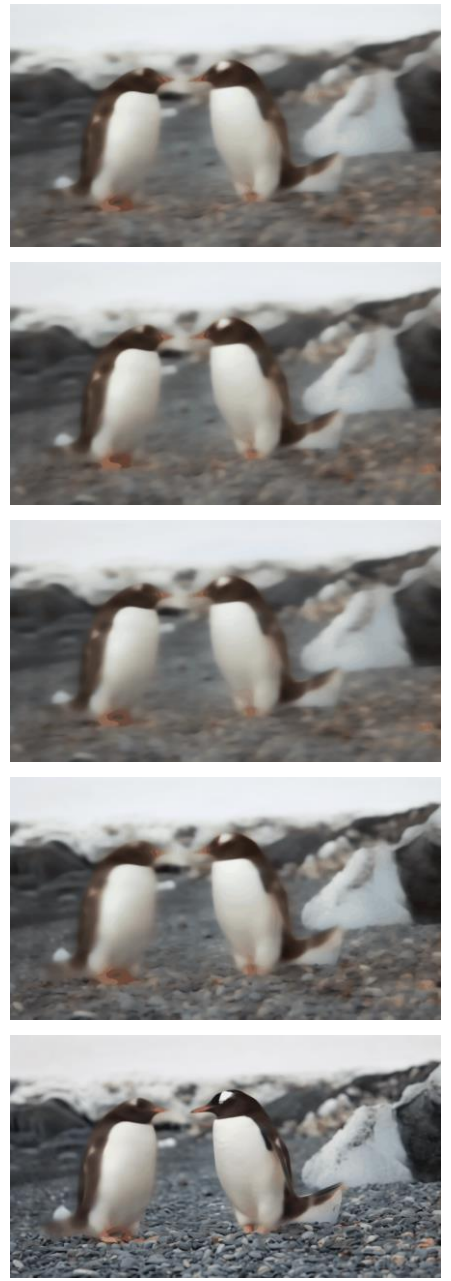


Loss computation

- Loss is backpropagated through the input latents

How do we modify latents?





Optimization process

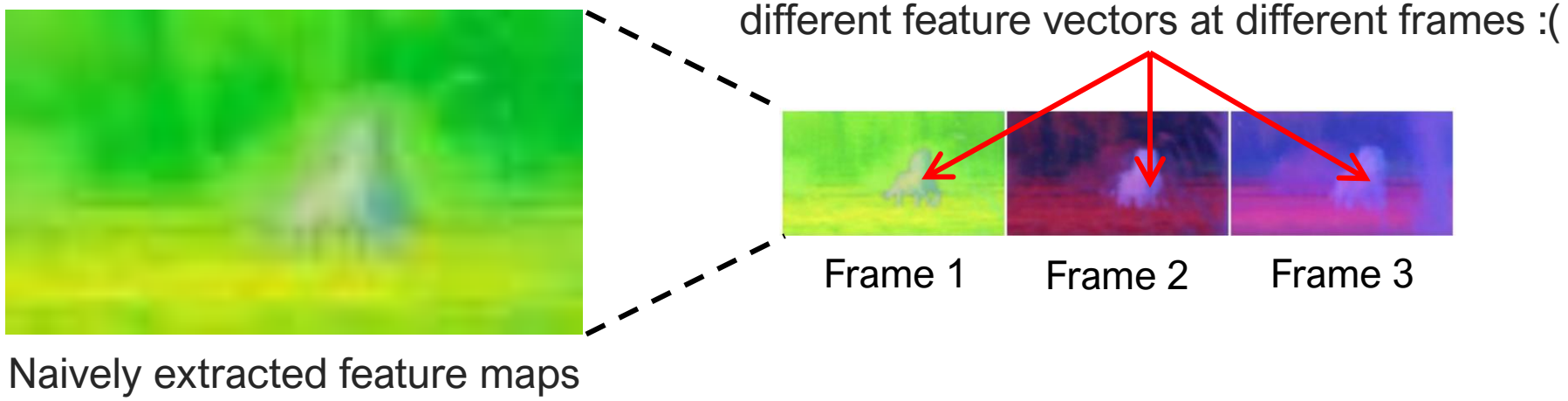
Challenge I:
How to obtain semantically aligned feature maps?

Issue : naively extracted feature maps are not semantically aligned

Naively extracted feature maps

Challenge I:
How to obtain semantically aligned feature maps?

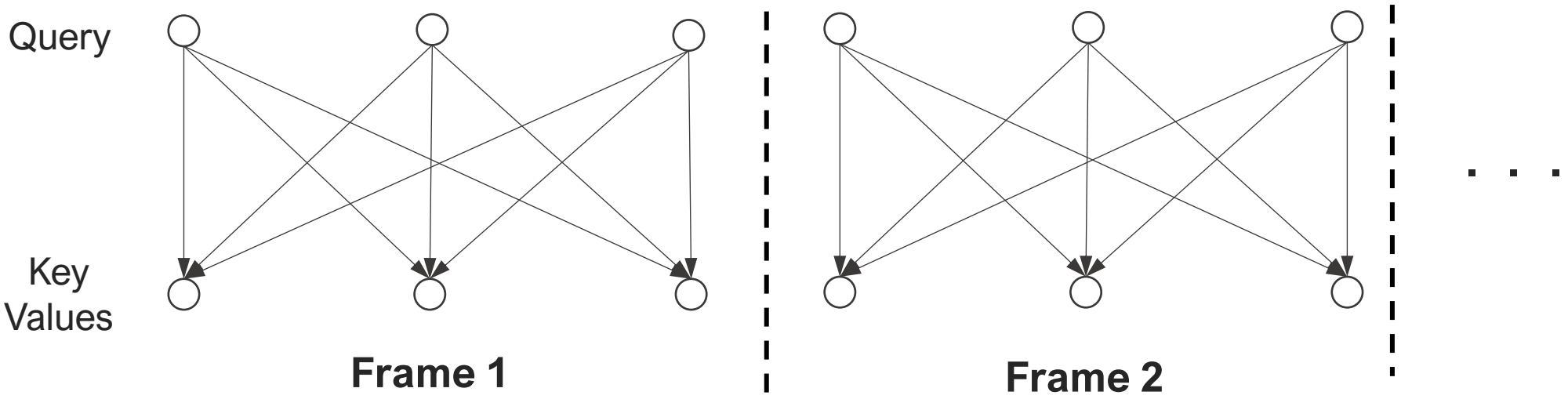
Issue : naively extracted feature maps are not semantically aligned



Challenge I:
How to obtain semantically aligned feature maps?

Key finding: We can produce semantically aligned feature maps by modifying the computations of self-attention layers.

Original spatial self-attention

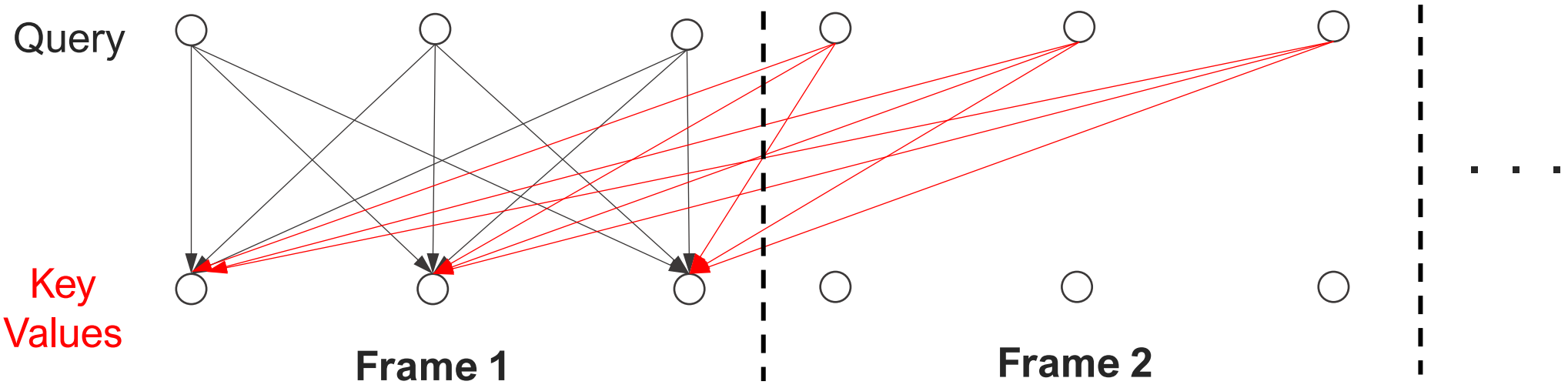


Self-attention is computed independently for each frame

Challenge I:
How to obtain semantically aligned feature maps?

Key finding: We can produce semantically aligned feature maps by modifying the computations of self-attention layers.

Modified spatial self-attention



Replace the key/value tokens with that of the first frame
-> Produced feature maps are weighted sum of the value tokens from the first frame

Challenge I:
How to obtain semantically aligned feature maps?

Key finding: We can produce semantically aligned feature maps by modifying the computations of self-attention layers.

Original feature maps

Modified
Self-attention
Computation

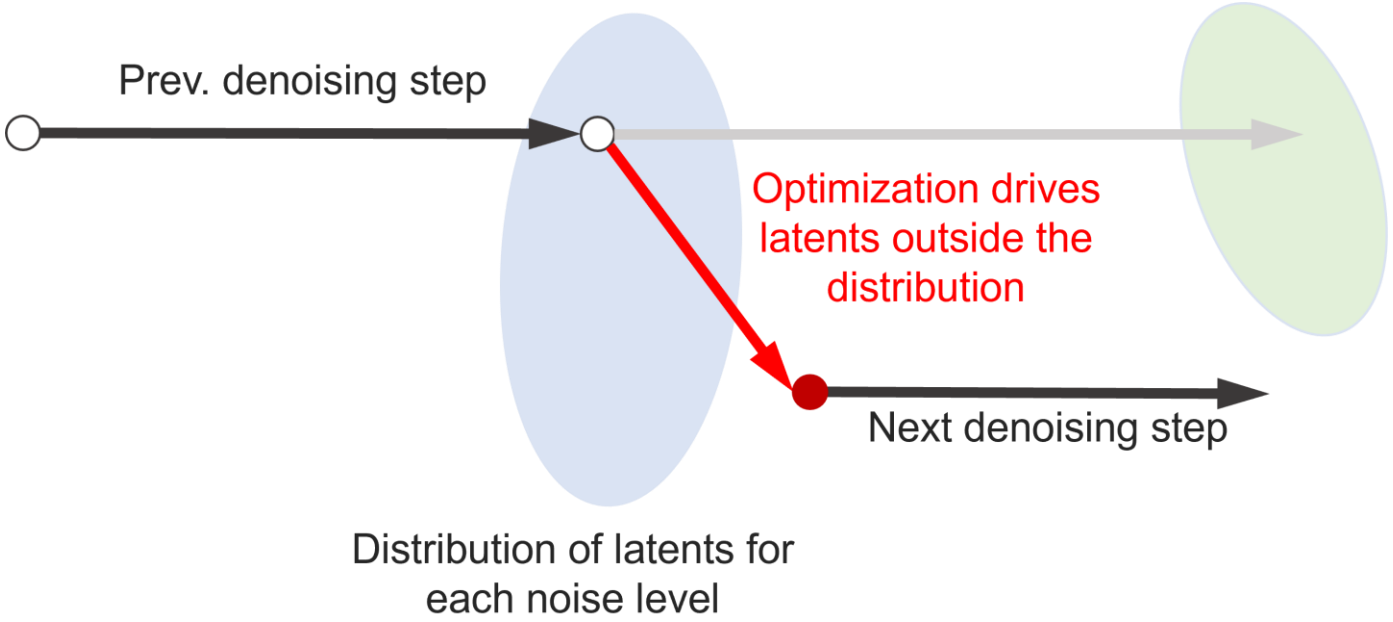
Semantically aligned feature maps

Frame 1 Frame 2 Frame 3

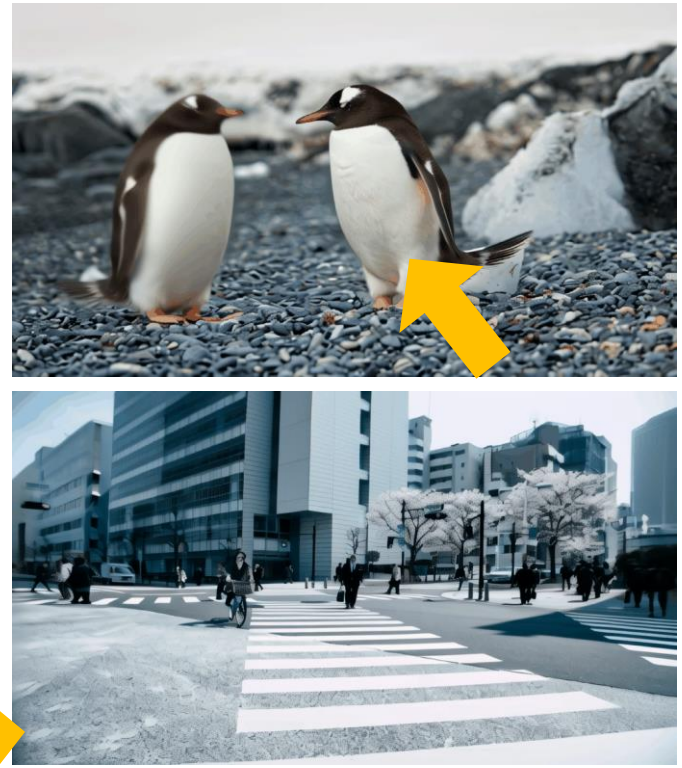
Frame 1 Frame 2 Frame 3

Challenge II: Recovering visual quality

- × Optimized latents may become out-of-distribution



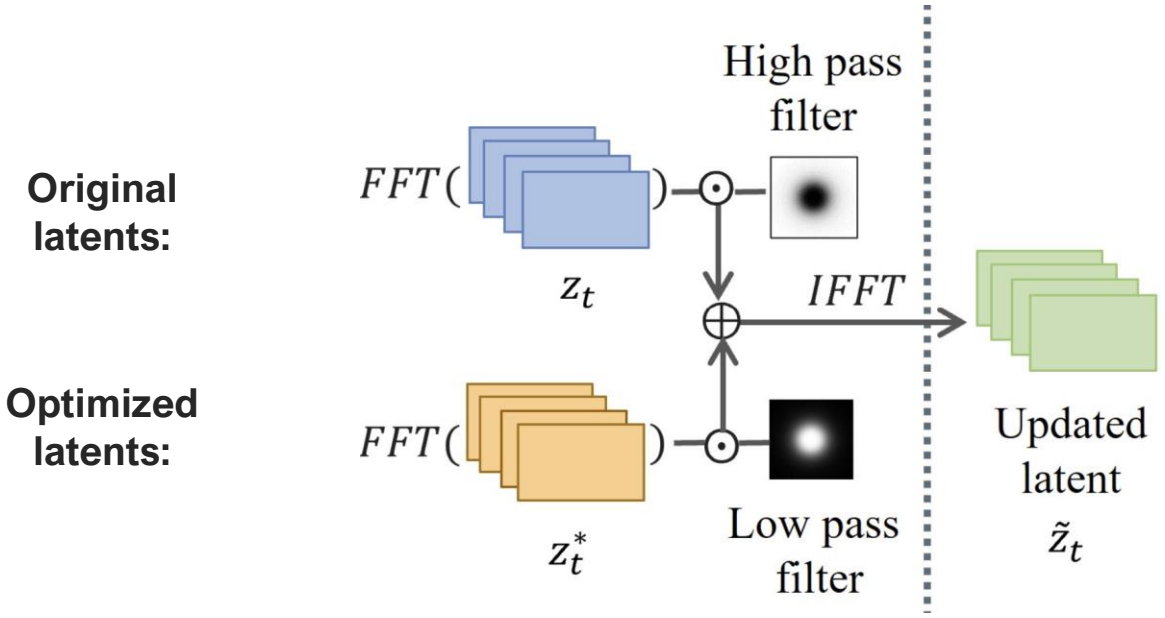
Visual quality degradation



Challenge II: Recovering visual quality

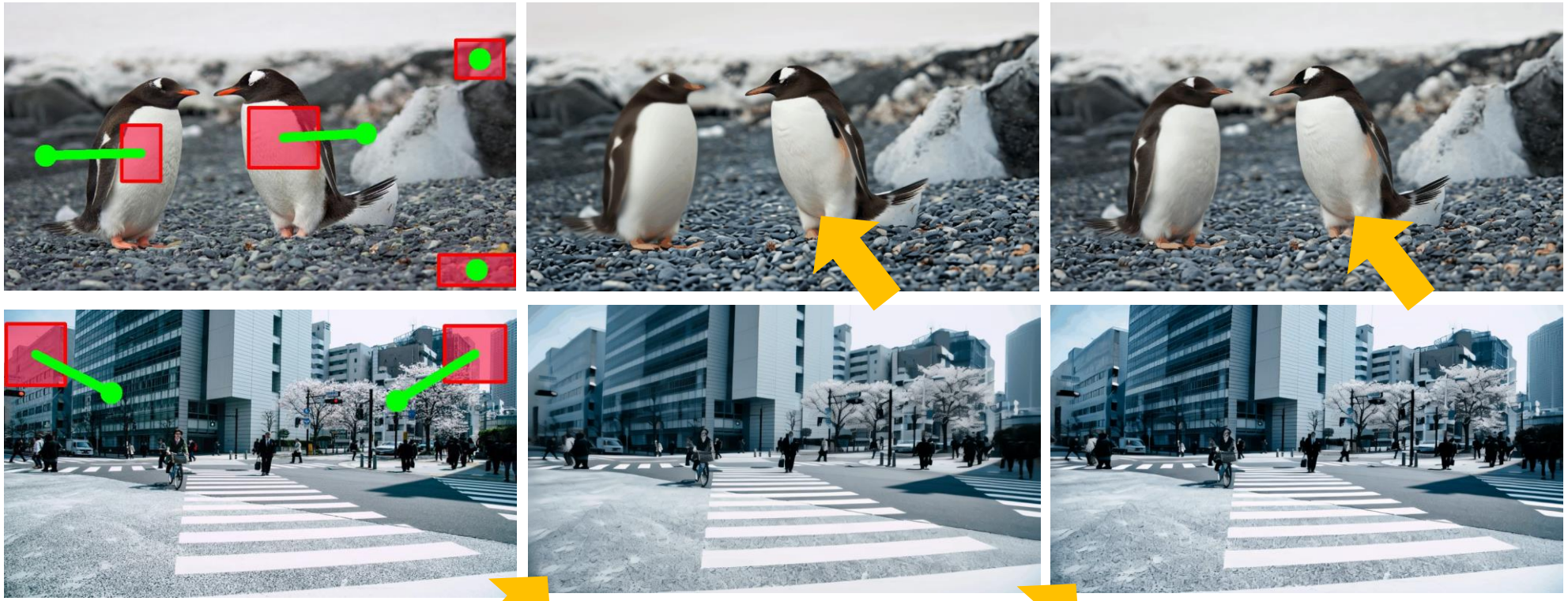
Key observation: only the low-frequency components of optimized latents significantly influence motion.

Solution: Preserve high-frequency components of the original latents



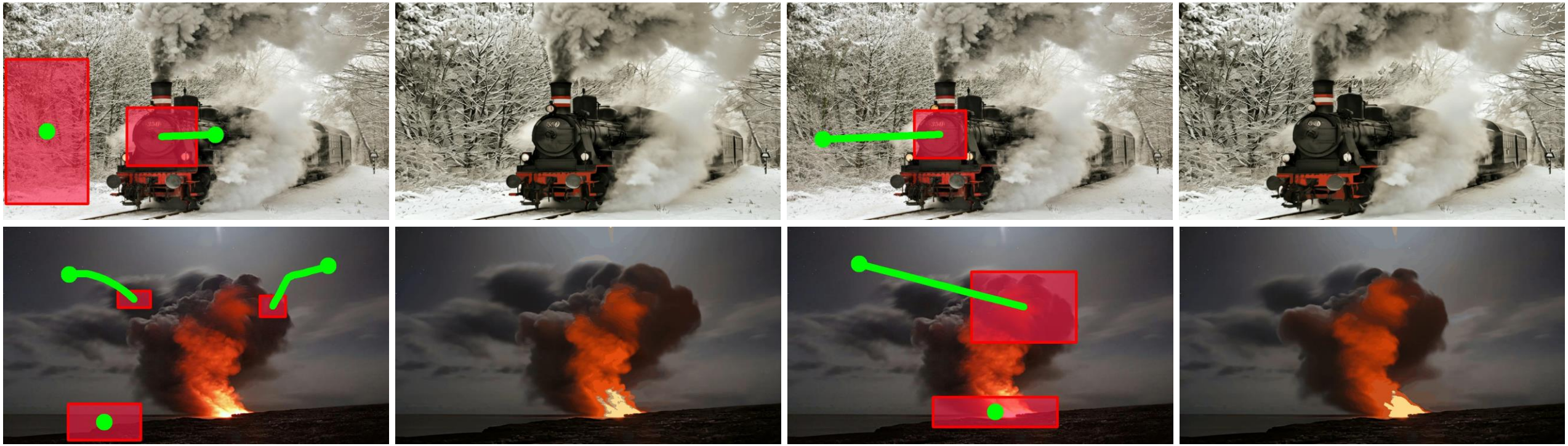
Challenge II: Recovering visual quality

Ablation study: this simple post-processing technique recovers visual quality, while maintaining the motions of the optimized latents.



Results

✓ controls both rigid (e.g., train) and non-rigid (e.g., smokes) motions



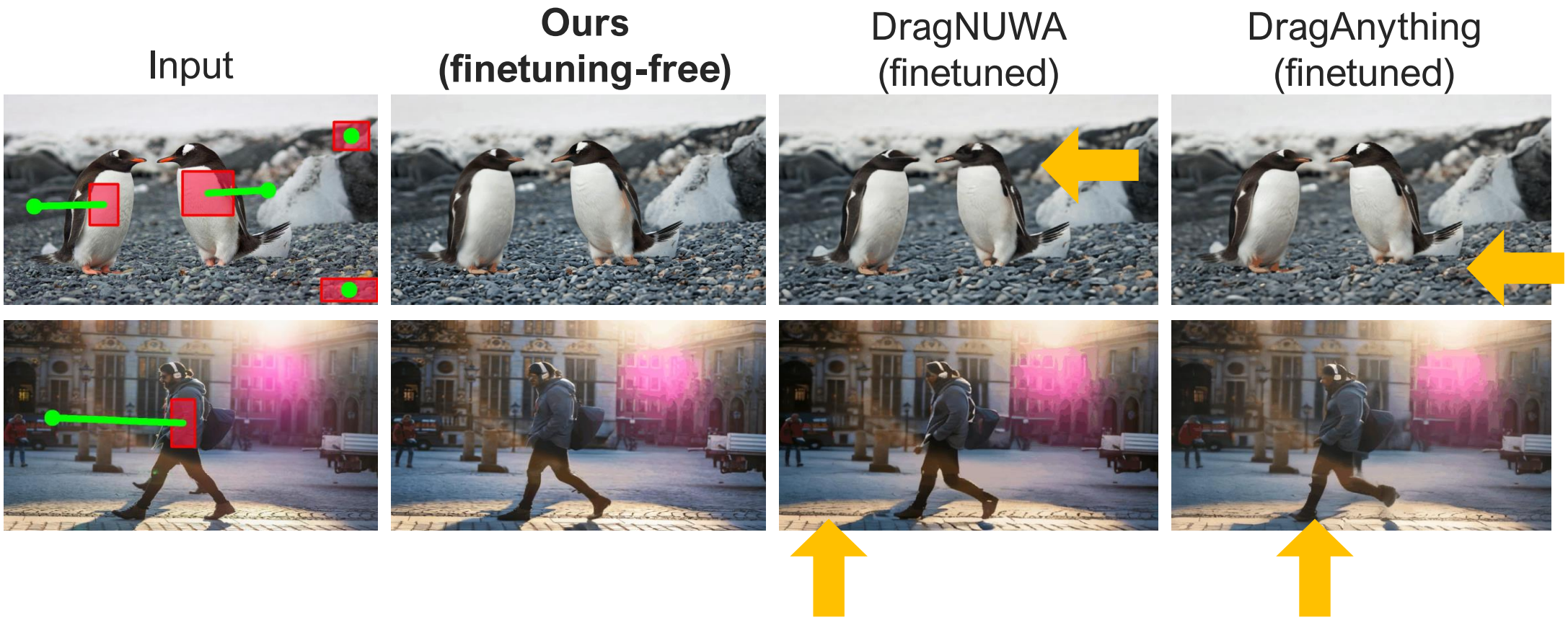
Results

✓ handles camera motions (e.g., zooming in, zooming out)

More Results

✓ These results demonstrate versatile control abilities of our approach!

Qualitative comparison with supervised baselines



Thank you for listening!

- Project website: https://kmcode1.github.io/Projects/SG-I2V/

- Poster session:
  - Title: SG-I2V: Self-Guided Trajectory Control in Image-to-Video Generation
  - Poster Session 2 (Thu 24 Apr 3 p.m. - 5 p.m.)