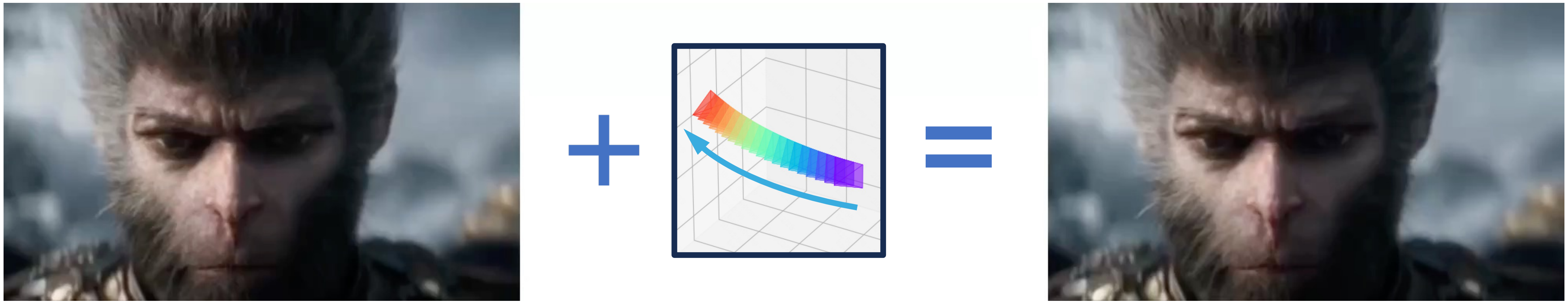
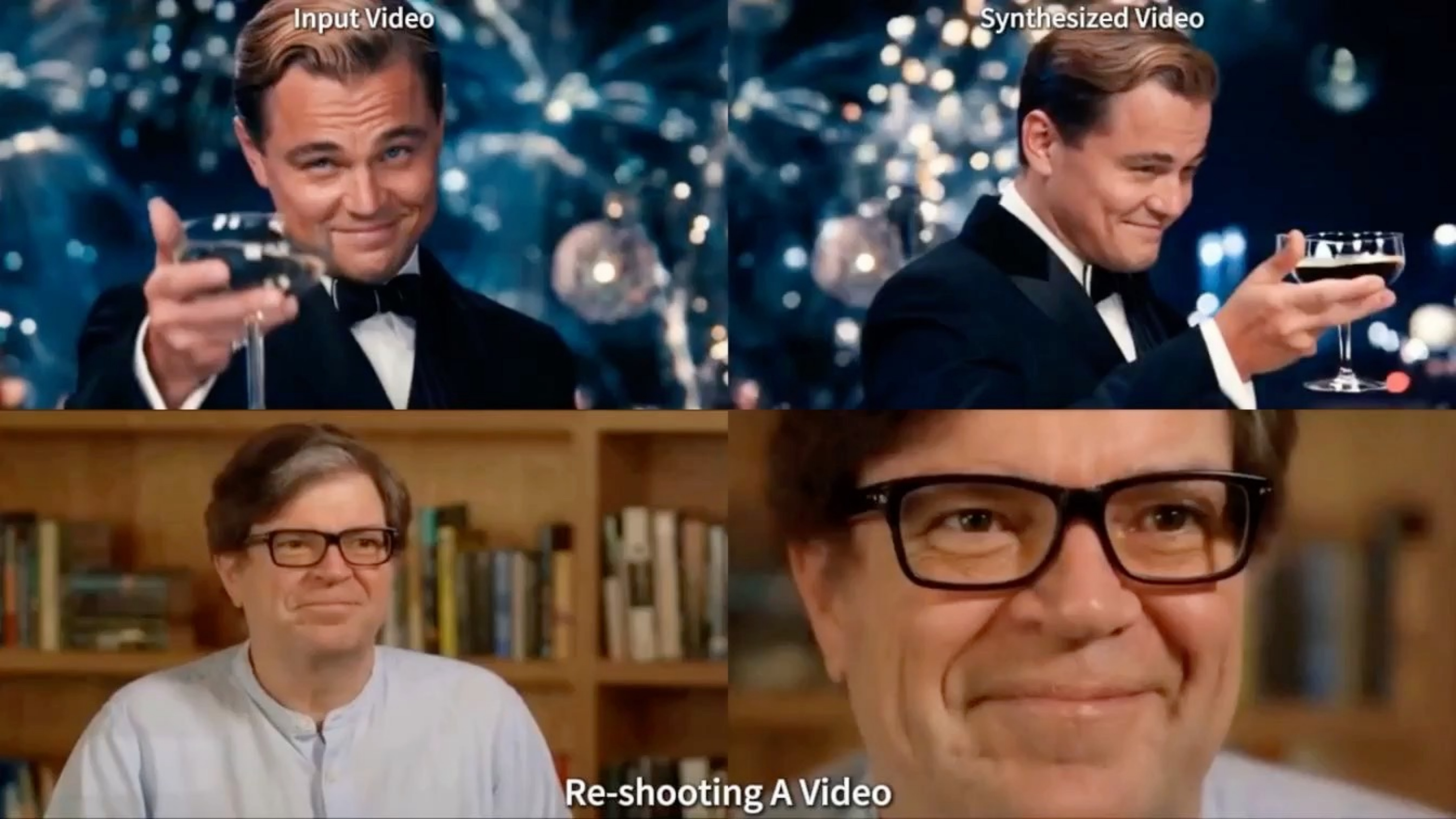
ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

ReCamMaster: Camera-Controlled Generative Rendering from A Single Video

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao,
Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, Di Zhang

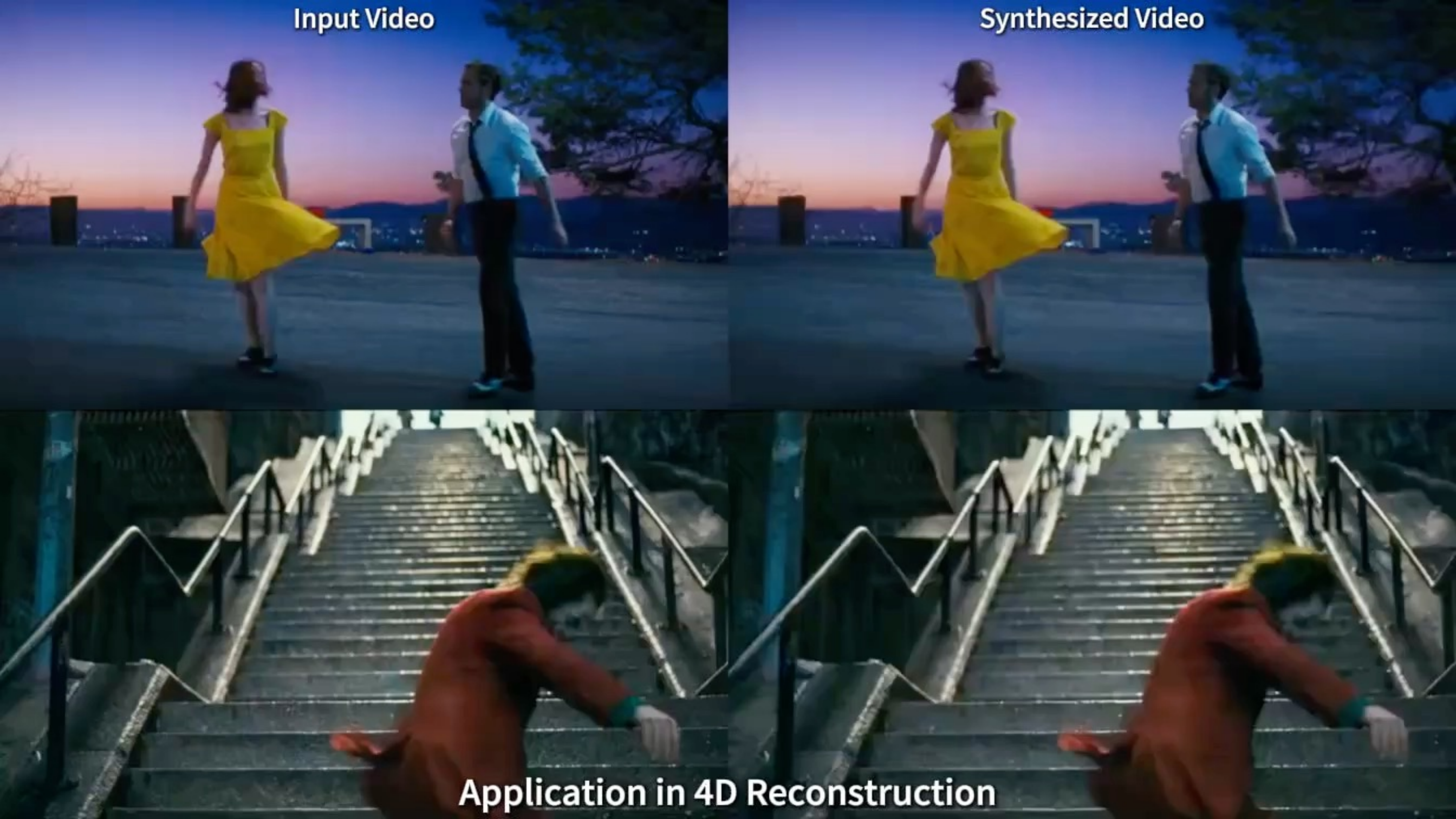


Recapture a video with new camera trajectories



Re-shooting A Video

Re-shooting A Video

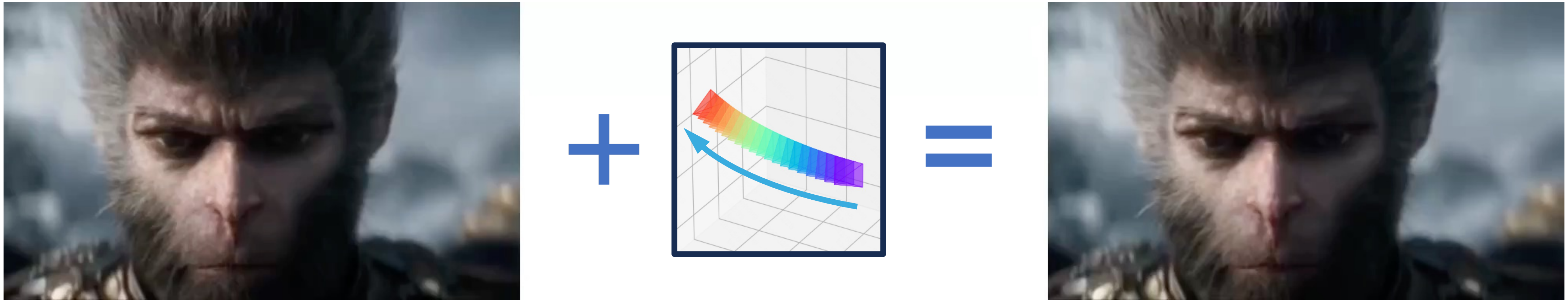


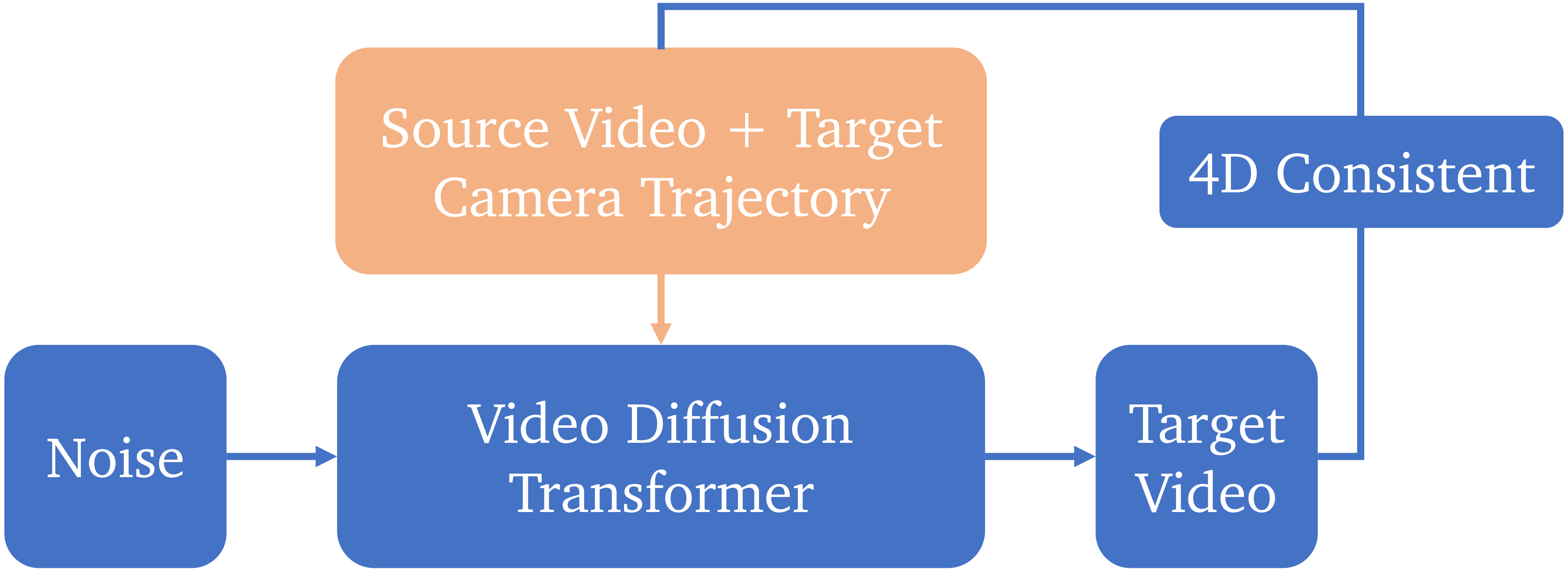
Application in 4D Reconstruction

Application in Video Stabilization

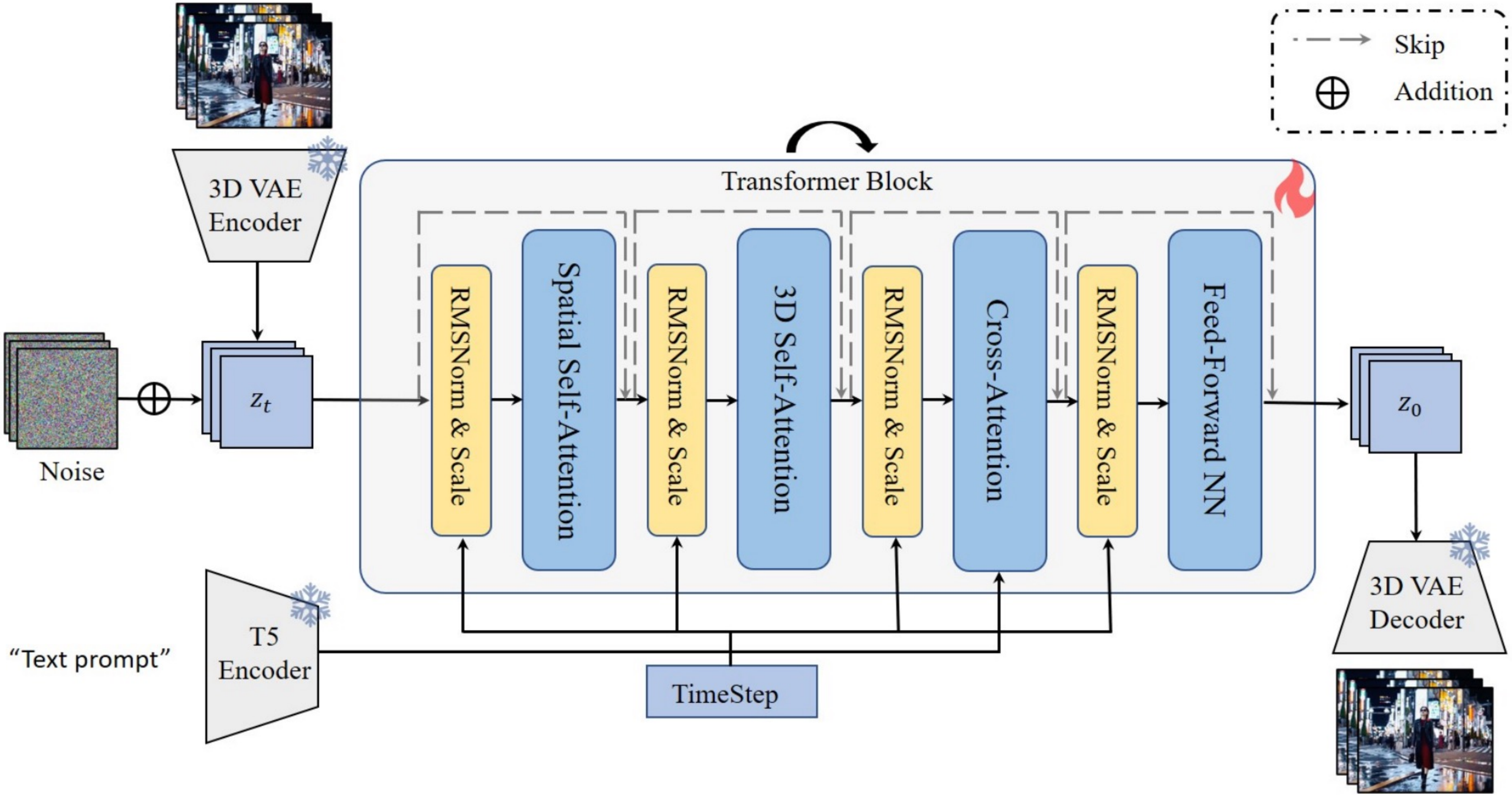
Supreme
Supreme

Given a source video and a target camera trajectory, we aim to synthesize a target video sharing the same dynamic scene (4D consistent) and adhering to the input trajectory.



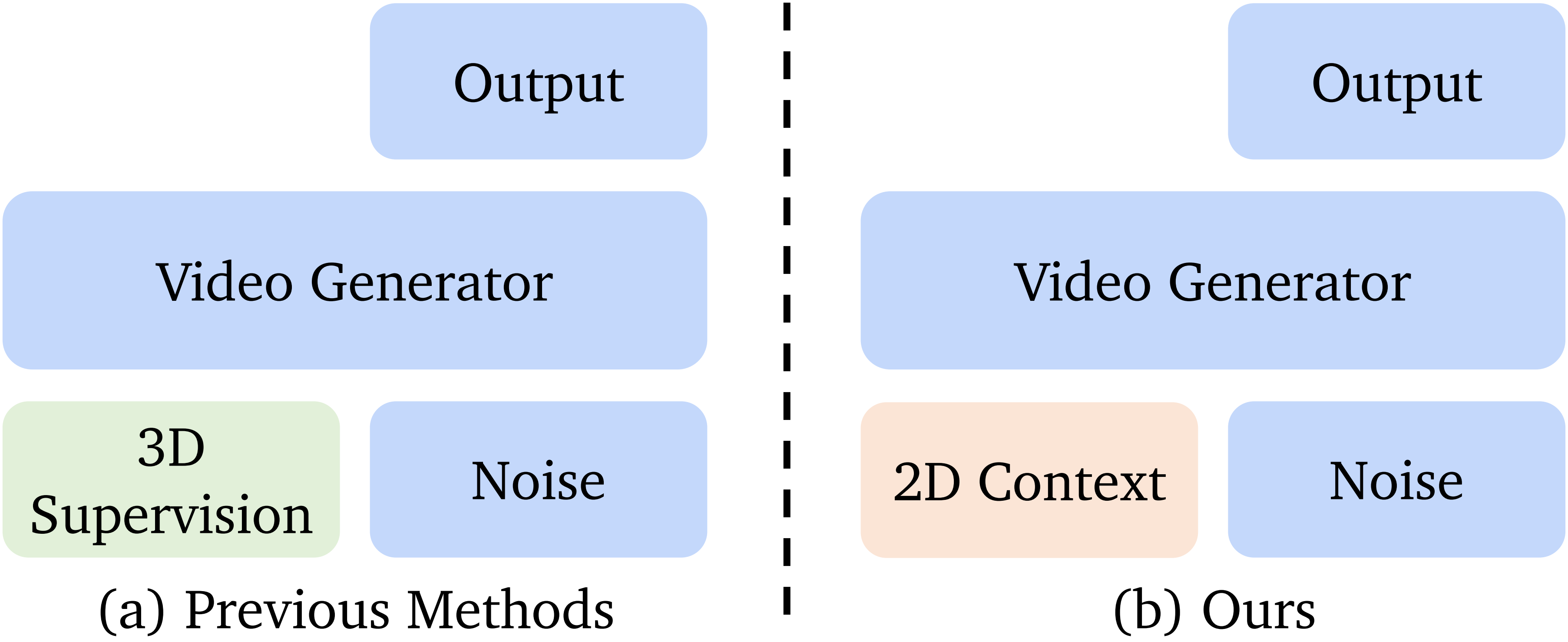

Source Video + Target Camera Trajectory
4D Consistent
Noise
Video Diffusion Transformer
Target Video

How to inject conditions?
How to obtain training data?

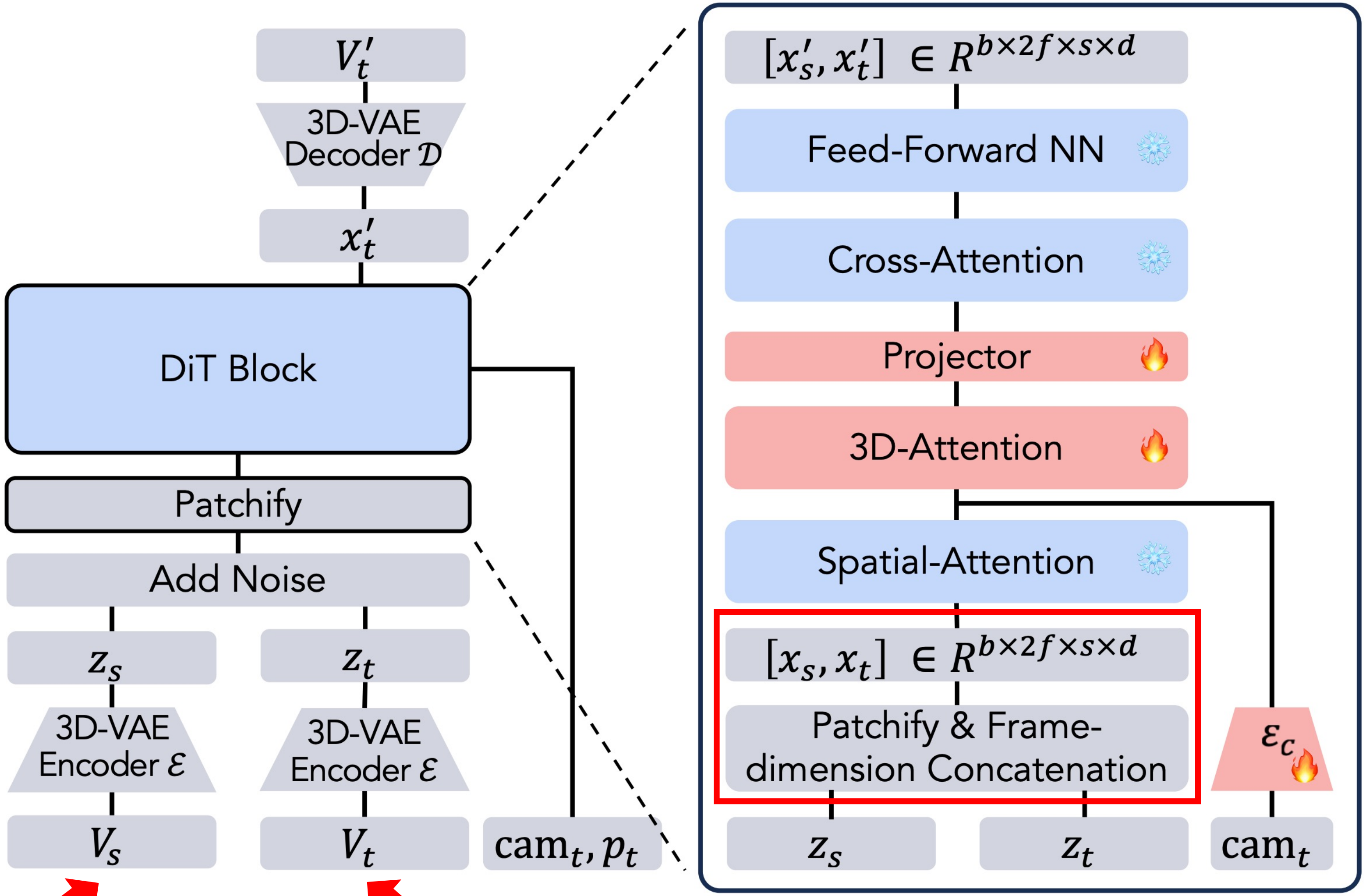


Transformer-based latent diffusion model as the base model.

How to inject conditions?

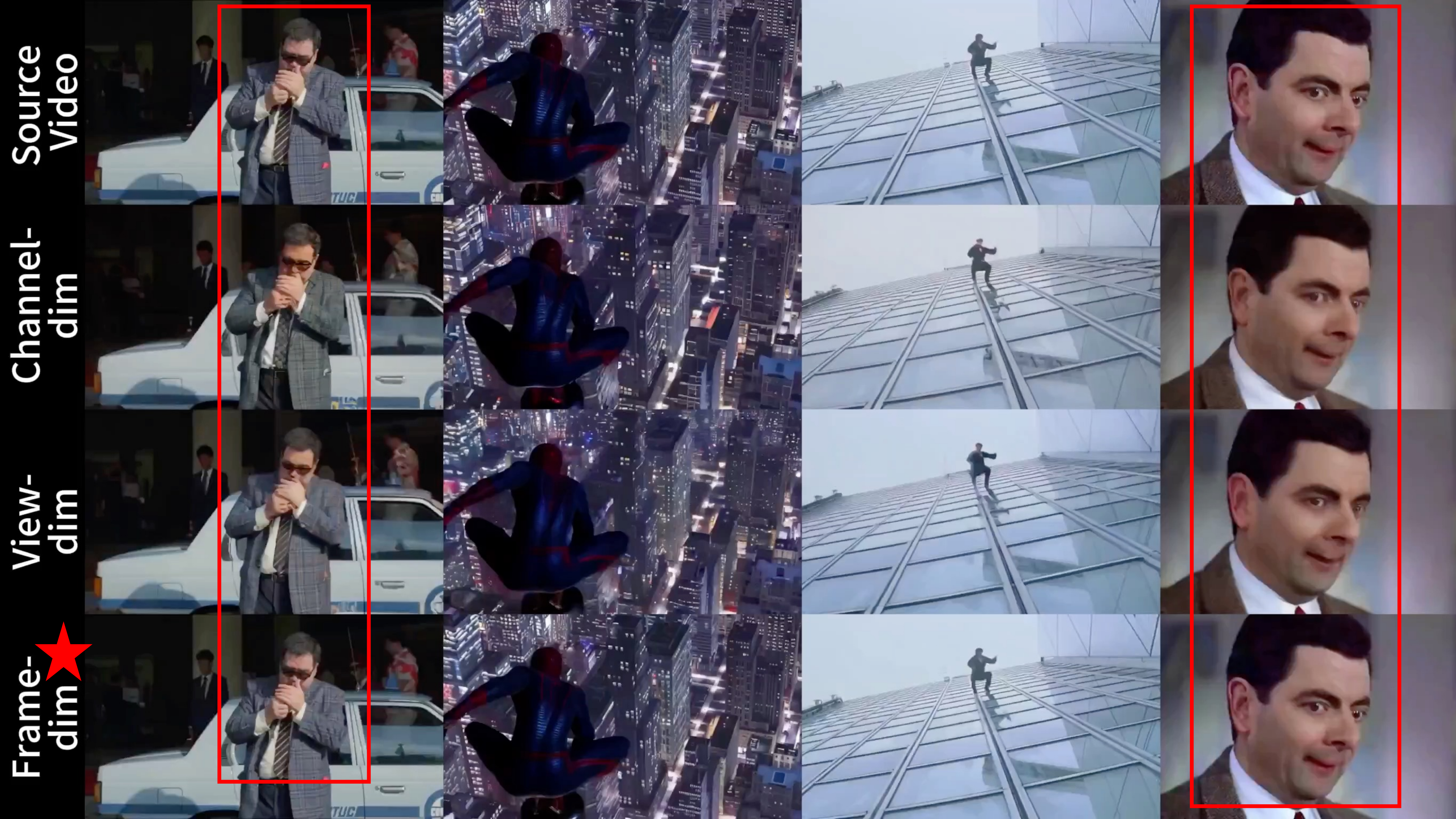


- 3D consistency can be learned without explicit 3D supervision.
- Video generators are effective in-context learners, in-context conditioning achieves better performance.



V'_t
3D-VAE Decoder $\mathcal{D}$
x'_t
DiT Block
Patchify
Add Noise
z_s
z_t
3D-VAE Encoder $\mathcal{E}$
3D-VAE Encoder $\mathcal{E}$
V_s
V_t
cam_t, p_t
$[x'_s, x'_t] \in R^{b\times 2f\times s\times d}$
Feed-Forward NN
Cross-Attention
Projector
3D-Attention
Spatial-Attention
$[x_s, x_t] \in R^{b\times 2f\times s\times d}$
Patchify & Frame-dimension Concatenation
$\mathcal{E}_c$
z_s
z_t
cam_t

In-context conditioning is much more effective.


Source Video
Channel-dim
View-dim
Frame-dim

Data Curation

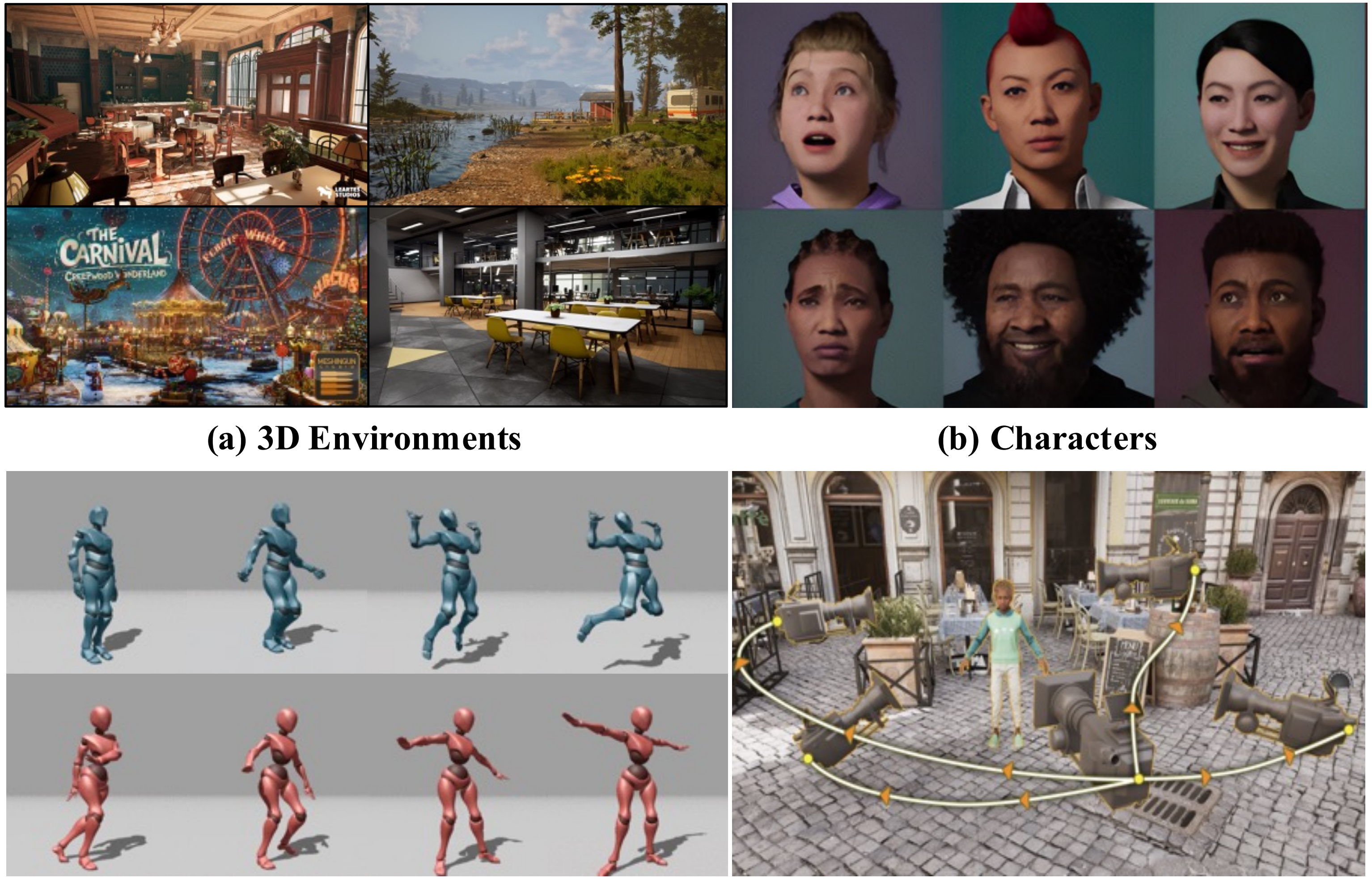
What kind of data do we need?

1. Video and camera parameters.
2. Multi-camera synchronized.
3. Diverse camera trajectories.

→ Hard to collect in the real world!

Our Solution: Rendering with Synthetic Data Engine

Data Curation



(a) 3D Environments

(b) Characters

(c) Animations

(d) Camera Trajectories

The MultiCamVideo Dataset is a multi-camera synchronized video dataset rendered using Unreal Engine 5.

MultiCamVideo Dataset (open source on HuggingFace)

OK STOR
OK STOR

TOMATO SAUCE
PINEAPPLE SLICES

Supreme
eturn
Please Pull Forward
Please Pu
STOP

Boarder Impacts & Takeaway Messages

1. The video generation model can understand the 4D scene and can be used as a renderer to generate 3D/4D-consistent content (towards spatial intelligence and world models).
2. Learning with minimal 3D bias can be easier to scale up and have better performance (e.g., Genie3, RTFM, etc.).
3. In-context conditioning is effective for transformer-based generative models, it could be generalized to more tasks.

Code and dataset are open source.

Thanks for your attention!