Perception-as-Control: Fine-grained Controllable Image Animation with 3D-aware Motion Representation

Yingjie Chen, Yifang Men, Yuan Yao, Miaomiao Cui, Liefeng Bo

arXiv 25.01

STRUCT Group Seminar
Presenter: Yifan Li
2025.1.19

Outline

- Background
- Method
- Experiments
- Conclusion

Background

Video Editing

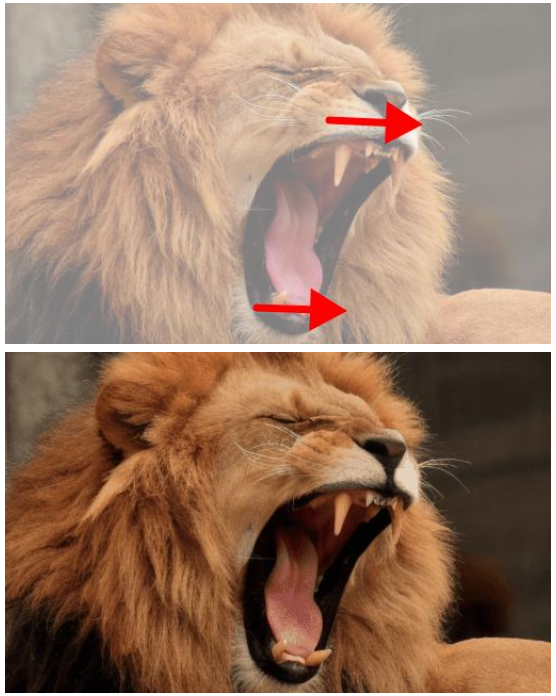
- Video to video translation
- Motion patterns are established by source videos

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. "FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation", CVPR'24.

Background

Image Animation

- Given an image + motion conditions (optional, e.g. trajectory, text, ...)
- Create new motion patterns: object **& camera**



Background

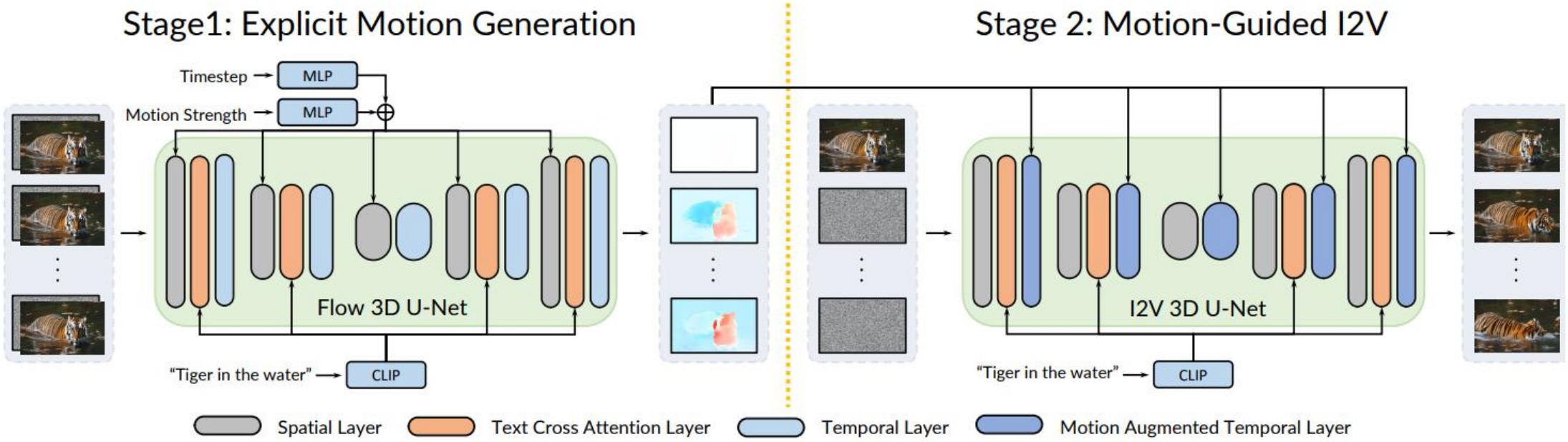
How to represent a motion?

- Optical Flow
  - Definition:
    - Offset between keyframe and i-th frame: $f_{0\to i} \in \mathbb{R}^{2\times h\times w}$
    - Update new frame with: $\mathbf{p}_i = \mathbf{p}_0 + f_{0\to i}$
  - Challenges:
    - Hard to perform long-term temporal consistency
      - pixel incompletion under huge motion
      - ill-posed prediction
    - **rigid for camera pose motion**

Background

Motion-I2V: two-stage optical flow-based generation

- Expensive data curation and training cost



"Motion-I2V: Consistent and Controllable Image-to-Video Generation with Explicit Motion Modeling", Xiaoyu Shi, Zhaoyang Huang, et al., SIGGRAPH'24

Background

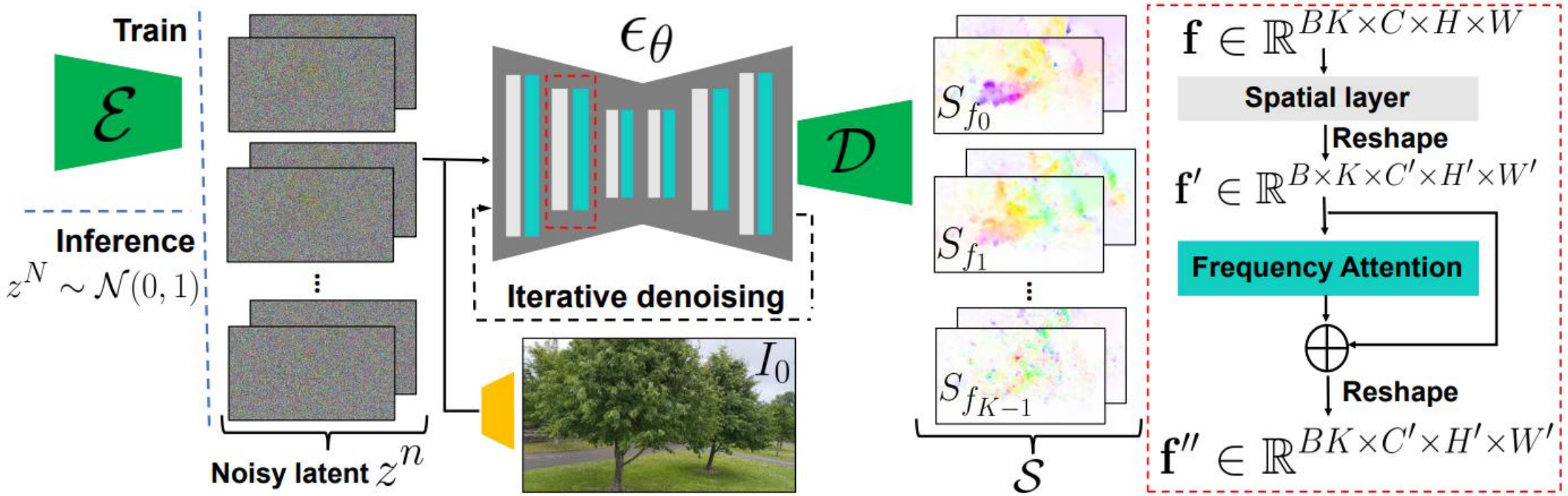
How to represent a motion?

- Spectral Volume $\boldsymbol{S}$
  - Definition**: frequencies modeled** of optical flow
    - Offsets $\{f_{0\to i}\}_{i=1}^{T} \in \mathbb{R}^{T\times 2\times h\times w}$ is too huge to handle
    - $\boldsymbol{S} = FFT(\{f_{0\to i}\}_{i=1}^{T})$
    - Discard high-freq. parts
    - Low-freq. covers natural motions basically
  - Benefits: lightweight and practical representation to generate
  - Challenges: still **rigid for camera pose motion**

Background

Generative Image Dynamics

- Collect 3,015 motion videos with natural periodic vibrations.
- Use Latent Diffusion Models (LDM) to generate spectrum volumes

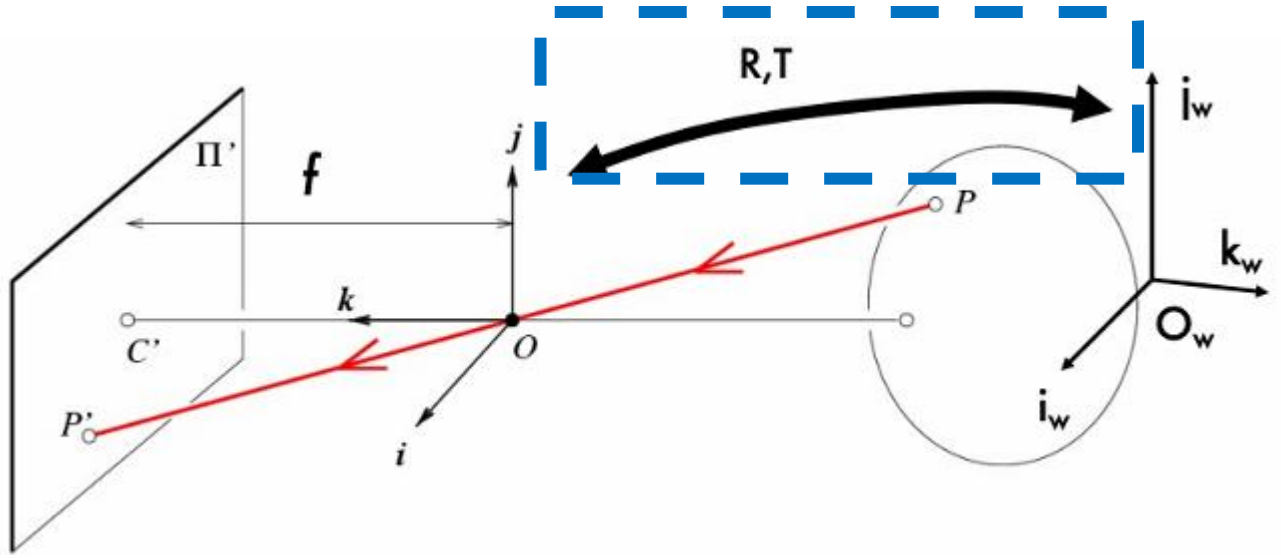


Generative Image Dynamics, Zhengqi Li Richard Tucker Noah Snavely Aleksander Holynski, CVPR'24 best paper

Background

How to represent a **camera** motion?

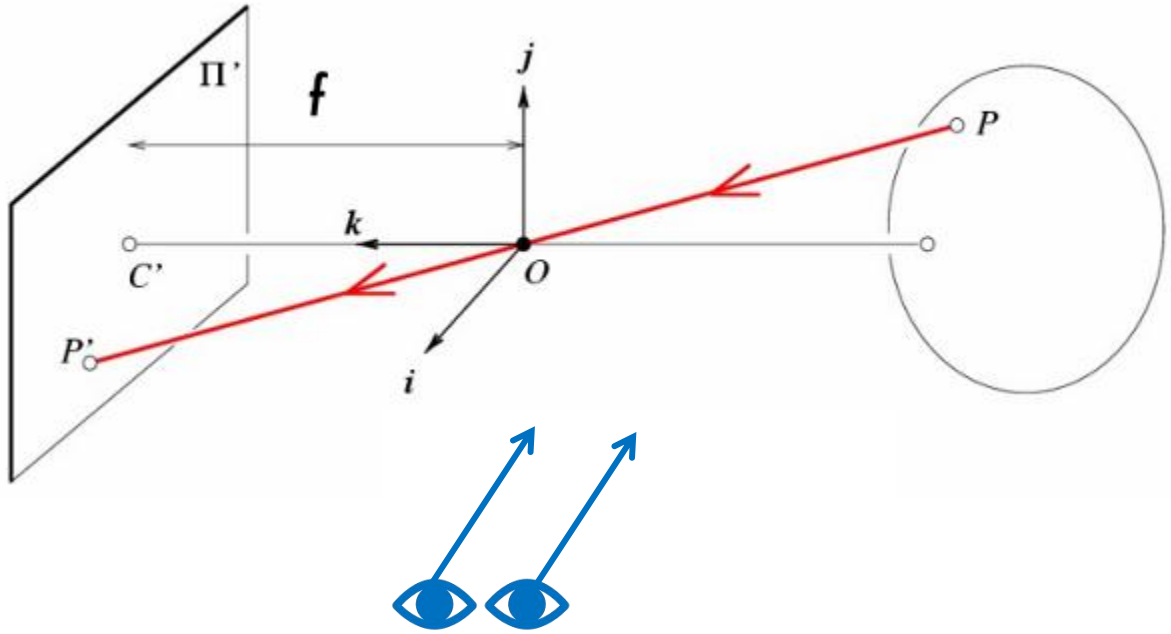
- Camera: map 3D coordinates (x, y, z) to 2D points (x', y')
  - **Intrinsic**/**Extrinsic** parameter: linear transformation matrices

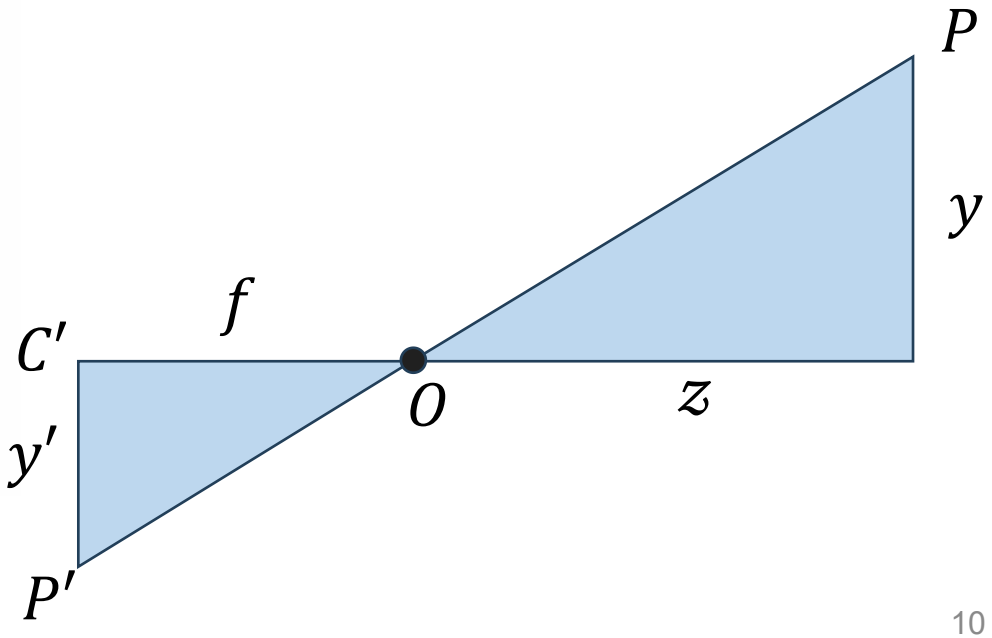


Background

Intrinsic parameter: map $P(x, y, z)$ to $P'(x', y')$

- Pinhole camera with focal length f
- Geometry relationship: $\frac{x}{z} = \frac{x'}{f}, \frac{y}{z} = \frac{y'}{f}$

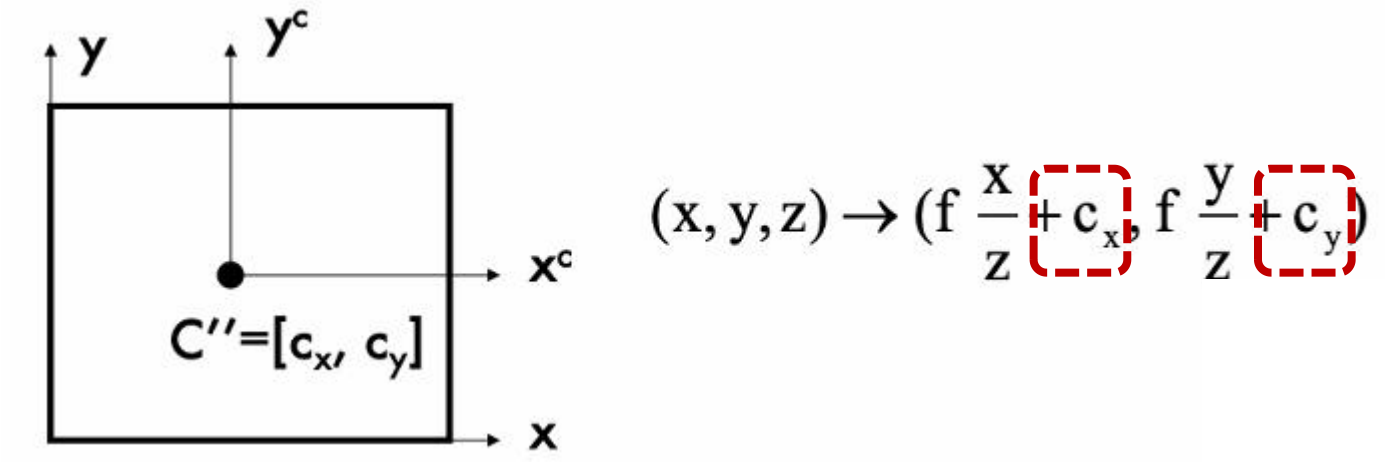




Background

Intrinsic parameter:

- Basic mapping from 3D to 2D: $\frac{x}{z} = \frac{x'}{f}, \frac{y}{z} = \frac{y'}{f}$
- 2D plane coordinates offset $c_x,\ c_y$



Background

Intrinsic parameter:

- Basic mapping from 3D to 2D: $\frac{x}{z} = \frac{x'}{f}, \frac{y}{z} = \frac{y'}{f}$
- 2D plane coordinates offset $c_x,\ c_y$
- Non-square 2D pixels:
  - $\alpha,\ \beta$ indicates pixel numbers along x-/y- axis

$$(x, y, z) \rightarrow (\underbrace{f\,k}_{\alpha}\frac{x}{z} + c_x, \underbrace{f\,l}_{\beta}\frac{y}{z} + c_y)$$

Background

Intrinsic parameter: internal properties of camera itself

- $c_x,\ c_y,\ \alpha,\ \beta$

$$P_h' = \begin{bmatrix} \alpha x + c_x z \\ \beta y + c_y z \\ z \end{bmatrix} = \begin{bmatrix} \alpha & 0 & c_x & 0 \\ 0 & \beta & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

Intrinsic Matrix

Background

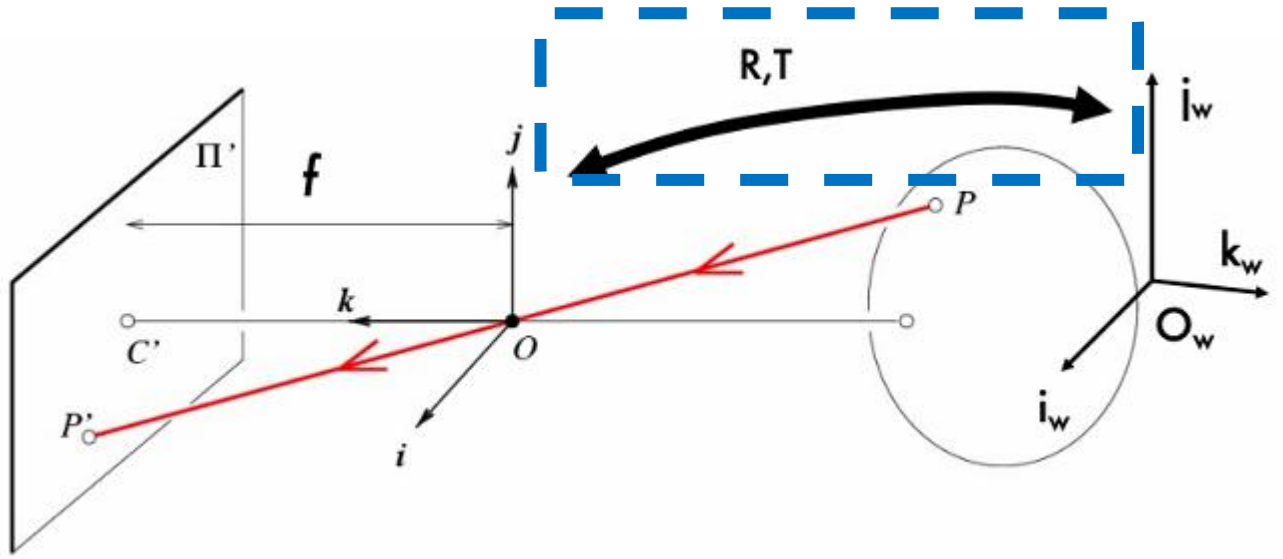
Extrinsic parameter

- A set of rotation and translation matrices

$$R_x(\alpha)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\alpha & -\sin\alpha\\ 0 & \sin\alpha & \cos\alpha\end{bmatrix}$$

$$R_y(\beta)=\begin{bmatrix}\cos\beta & 0 & \sin\beta\\ 0 & 1 & 0\\ -\sin\beta & 0 & \cos\beta\end{bmatrix}$$

$$R_z(\gamma)=\begin{bmatrix}\cos\gamma & -\sin\gamma & 0\\ \sin\gamma & \cos\gamma & 0\\ 0 & 0 & 1\end{bmatrix}$$



Background

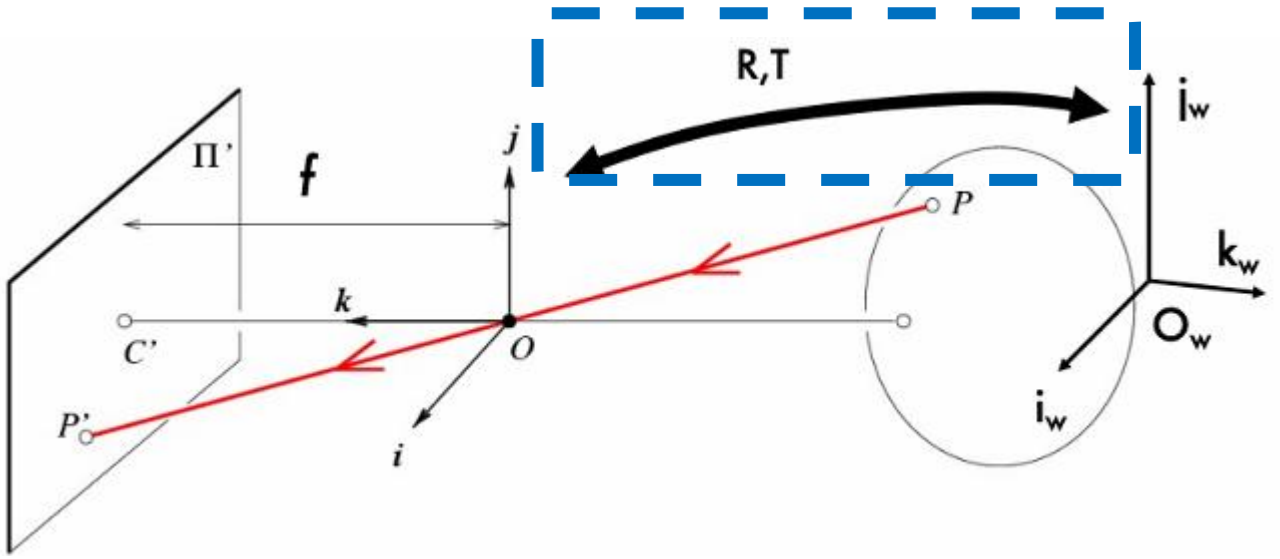
Extrinsic parameter

- Degrees of freedom: 6
  - 3 (rotation angles) + 3 (move along axes)

$$R_x(\alpha)=\begin{bmatrix}1 & 0 & 0\\ 0 & \cos\alpha & -\sin\alpha\\ 0 & \sin\alpha & \cos\alpha\end{bmatrix}$$

$$R_y(\beta)=\begin{bmatrix}\cos\beta & 0 & \sin\beta\\ 0 & 1 & 0\\ -\sin\beta & 0 & \cos\beta\end{bmatrix}$$

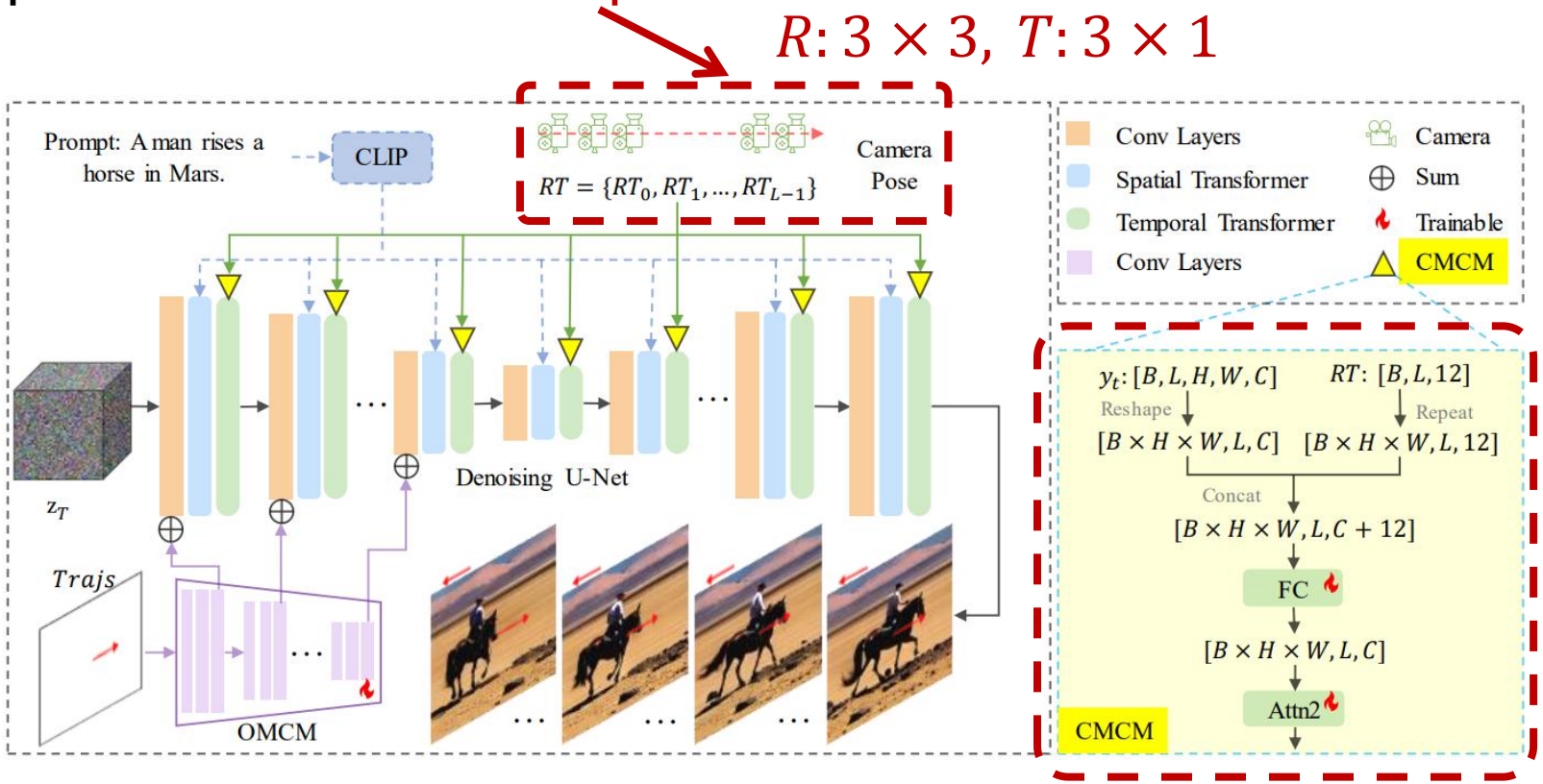
$$R_z(\gamma)=\begin{bmatrix}\cos\gamma & -\sin\gamma & 0\\ \sin\gamma & \cos\gamma & 0\\ 0 & 0 & 1\end{bmatrix}$$



Background

MotionCtrl

- directly repeat extrinsic camera parameters to feature size as control signals

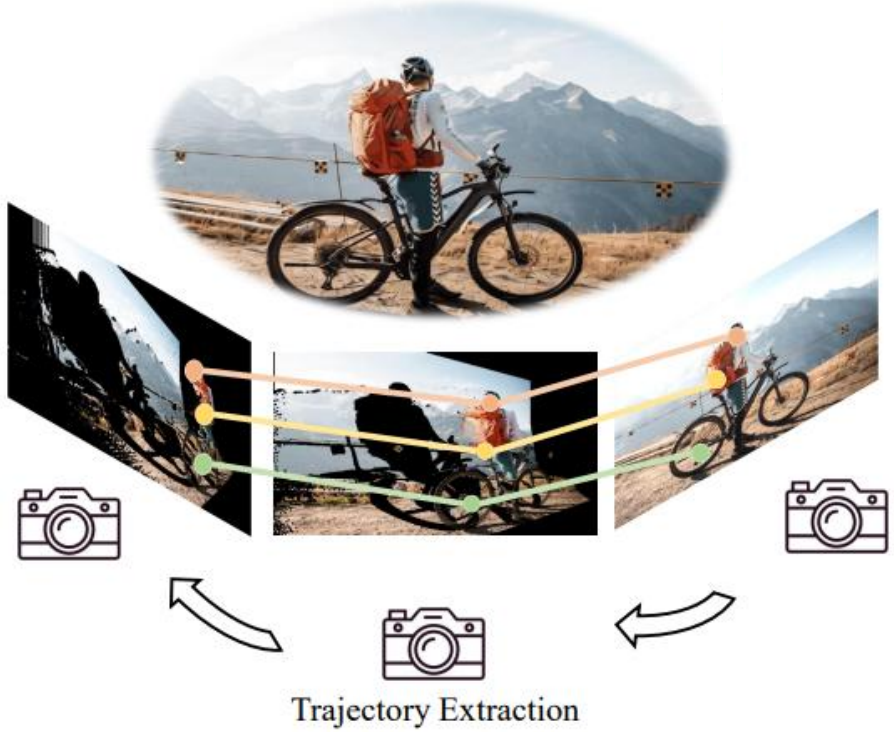


"MotionCtrl: A Unified and Flexible Motion Controller for Video Generation", Zhouxia Wang et al., SIGGRAPH'24

Background

TrajectoryAttention

- Extract trajectory as motion guidance across frames

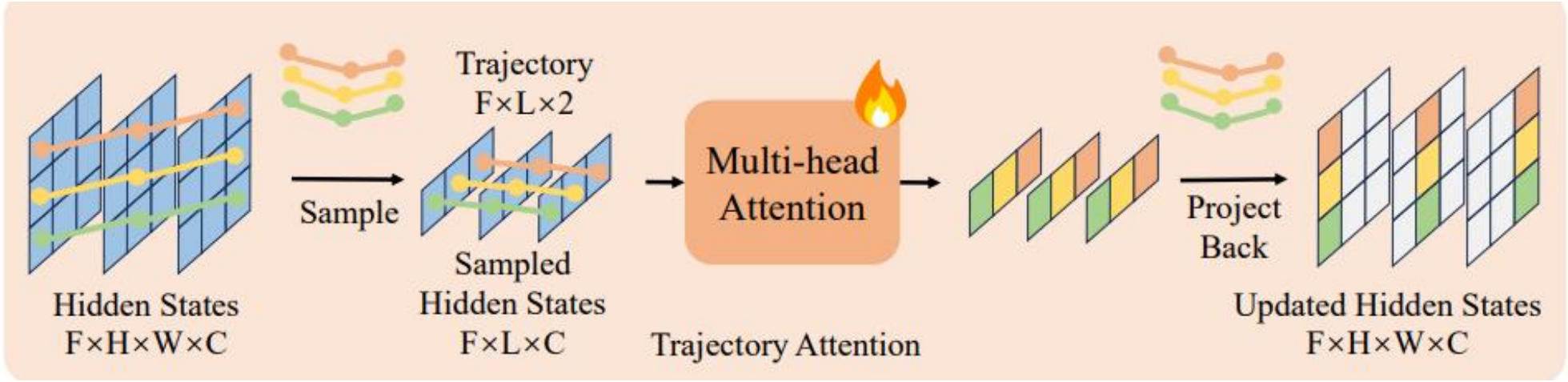


“Trajectory attention for fine-grained video motion control”, Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan, under submission of ICLR’25, score: 6,6,6,8,8

Background

TrajectoryAttention

- Explicit cross-frame regularization with trajectory guidance
- Cross-frame attention across points along the trajectories

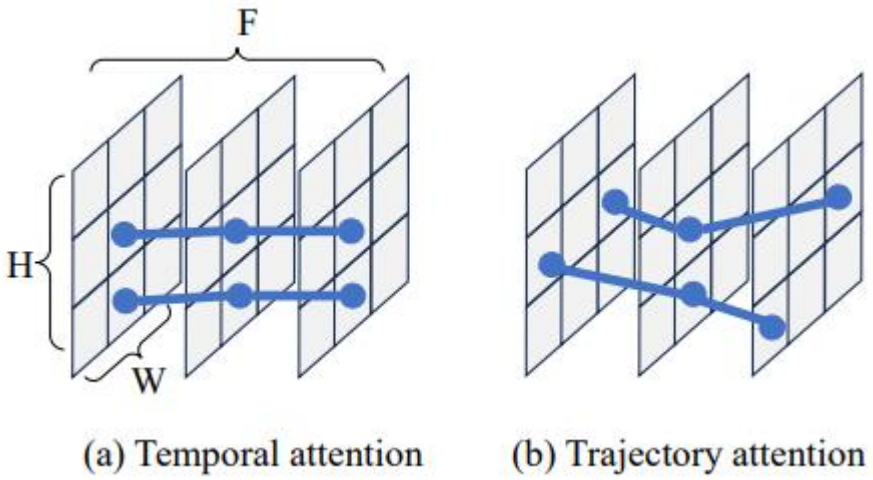


"Trajectory attention for fine-grained video motion control", Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan, under submission of ICLR'25, score: 6,6,6,8,8

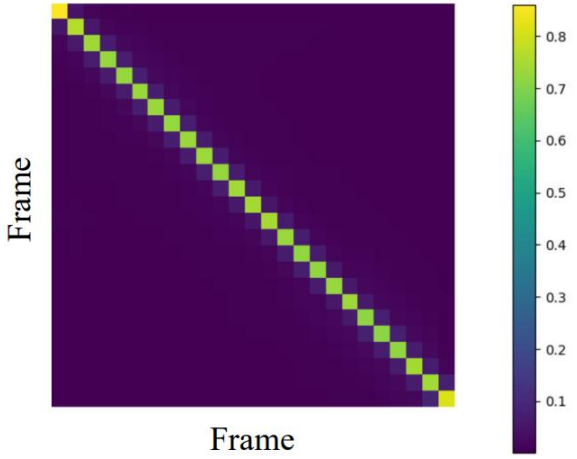
Background

TrajectoryAttention

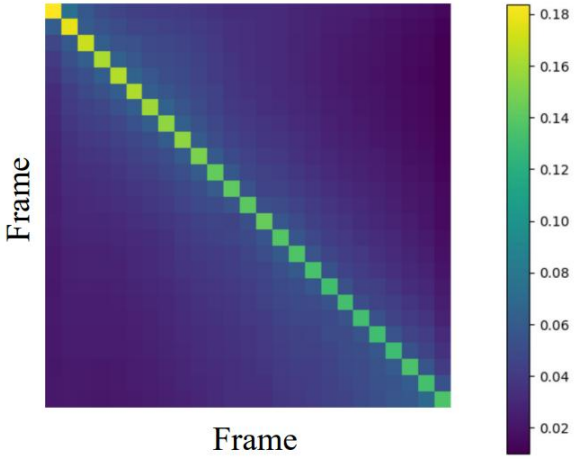
- Re-arranged temporal attention



(a) Temporal attention (b) Trajectory attention



(a) Attention map of temporal attention



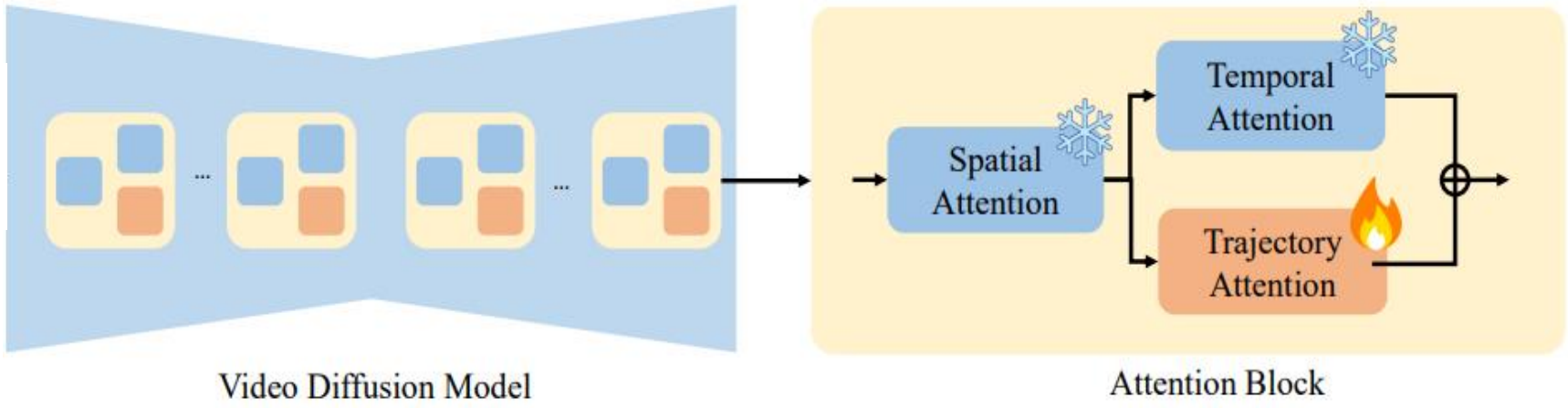
(b) Attention map of trajectory attention

"Trajectory attention for fine-grained video motion control", Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan, under submission of ICLR'25, score: 6,6,6,8,8

Background

TrajectoryAttention

- Applied as an auxiliary residual branch
- Stabilize training from the beginning



"Trajectory attention for fine-grained video motion control", Zeqi Xiao, Wenqi Ouyang, Yifan Zhou, Shuai Yang, Lei Yang, Jianlou Si, and Xingang Pan, under submission of ICLR'25, score: 6,6,6,8,8

Background

TrajectoryAttention

- Optical flows for training
- **Motion trajectory** extracted from camera explicit parameters for testing

Algorithm 3: Trajectory extraction from single image

Input: Image $\mathbf{I} \in \mathbb{R}^{H_p \times W_p \times 3}$, A set of camera pose with intrinsic and extrinsic parameters,$\{\mathbf{K} \in \mathbb{R}^{3\times3}\}$ and $\{\mathbf{E}[\mathbf{R};\mathbf{t}]\}$, where $\mathbf{R} \in \mathbb{R}^{3\times3}$ representations the rotation part of the extrinsic parameters, and $\mathbf{t} \in \mathbb{R}^{3\times1}$ is the translation part. The length of the camera pose equals frame number F. H_p and W_p are the height and width of the pixel space

1 Estimate the depth map $\mathbf{D} \in \mathbb{R}^{H_p \times W_p}$ from $\mathbf{I}$ given camera pose parameters.

2 Get the translation of pixels $\mathbf{T} \in \mathbb{R}^{F \times H_p \times W_p \times 2}$ based on $\mathbf{I}$ using using $\mathbf{D}$, $\mathbf{K}$, and $\mathbf{E}$.

3 Get trajecories $\mathbf{Tr} = \mathbf{T} + \mathbf{C}$, where $\mathbf{C} \in \mathbb{R}^{H_p \times W_p \times 2}$ is pixel-level grid coordinates of image with shape $H_p \times W_p$.

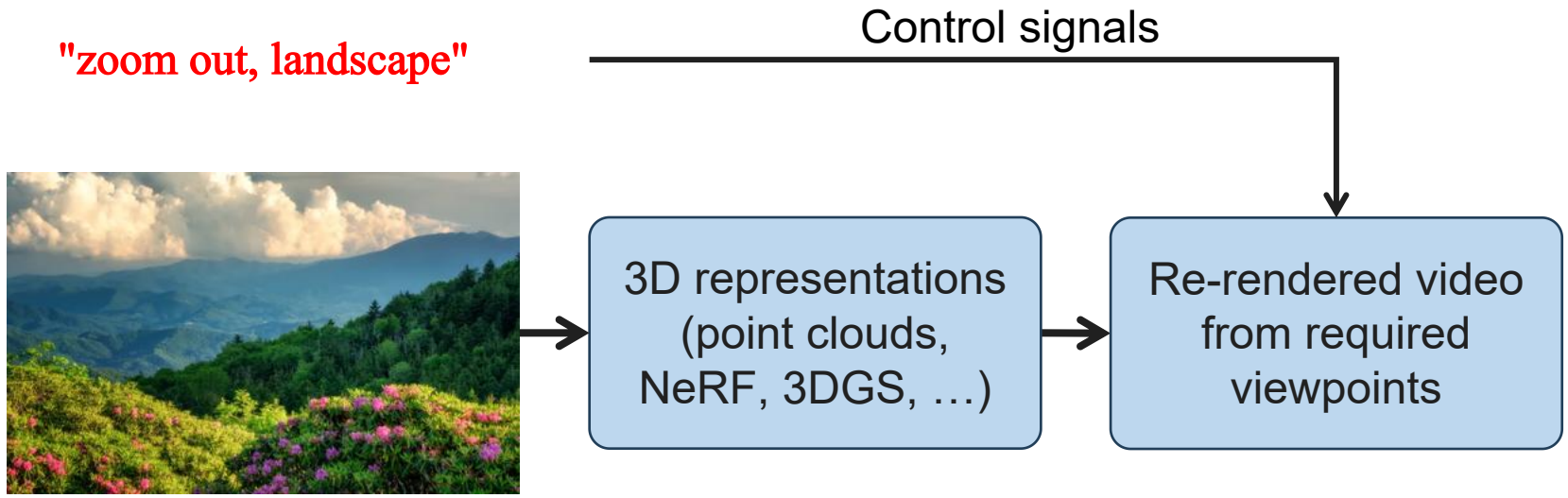
4 Get valid trajectory mask $\mathbf{M}$ for pixels that within the image space.

Output: Trajectories $\mathbf{Tr}$, Trajectory Masks $\mathbf{M}$

Background

How to represent a camera motion?

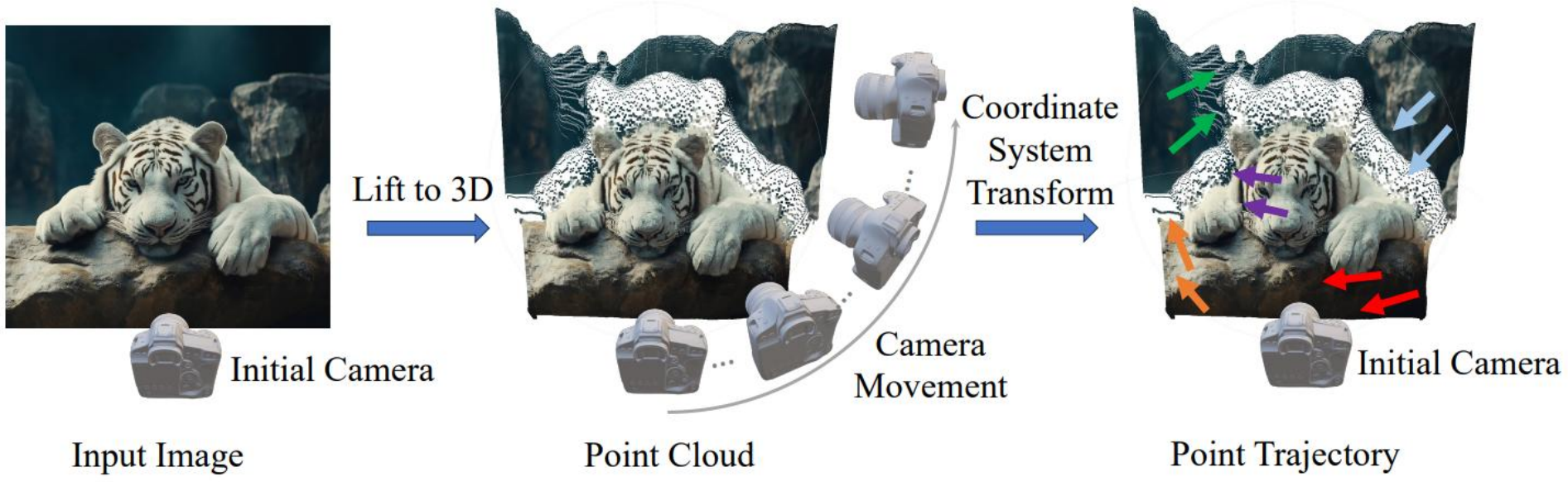
- A great effort for large scale **clean** camera extrinsic parameters
- An intermediate perspective: 3D reconstruction



Background

How to find a better 3D representation?

- **I2VControl-Camera**: Point cloud
  - Estimate depth maps from 2D video: narrow the gap between 2D and 3D
- Drawback: only handle camera motion, not object motions

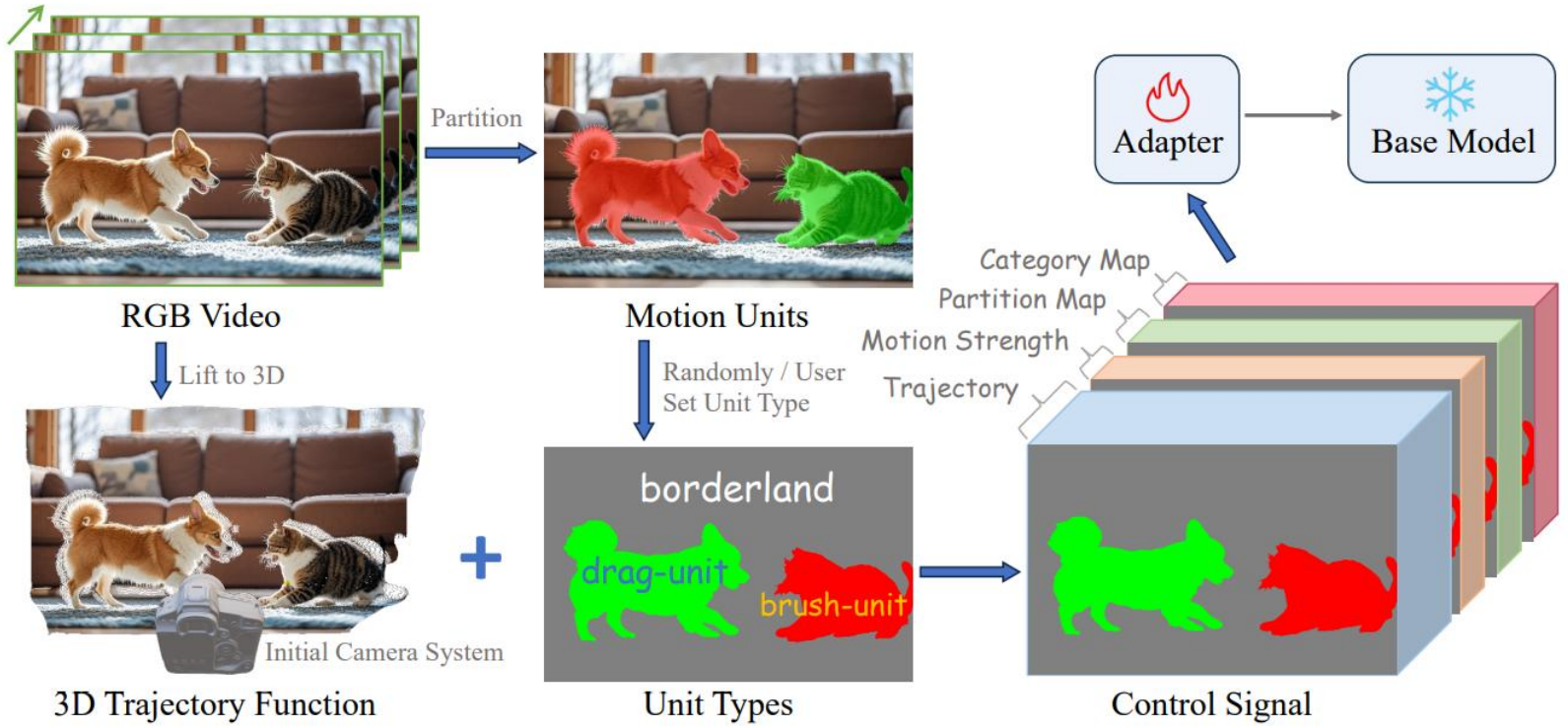


"I2VControl-Camera: Precise Video Camera Control with Adjustable Motion Strength", Wanquan Feng, et al., under submission of ICLR'25, score: 6,6,6,8

Background

How to find a better 3D representation?

- **I2VControl:** separated point clouds with human priors



"I2VControl: Disentangled and Unified Video Motion Synthesis Control", Wanquan Feng et al, arXiv 24.11

Background

I2VControl

- Borderland: background with minor motions
- Drag units:
  - controlled by rotations and translations
  - camera motions

Background

I2VControl

- Borderland
- Drag units
- Brush units: controlled by a motion strength scaler

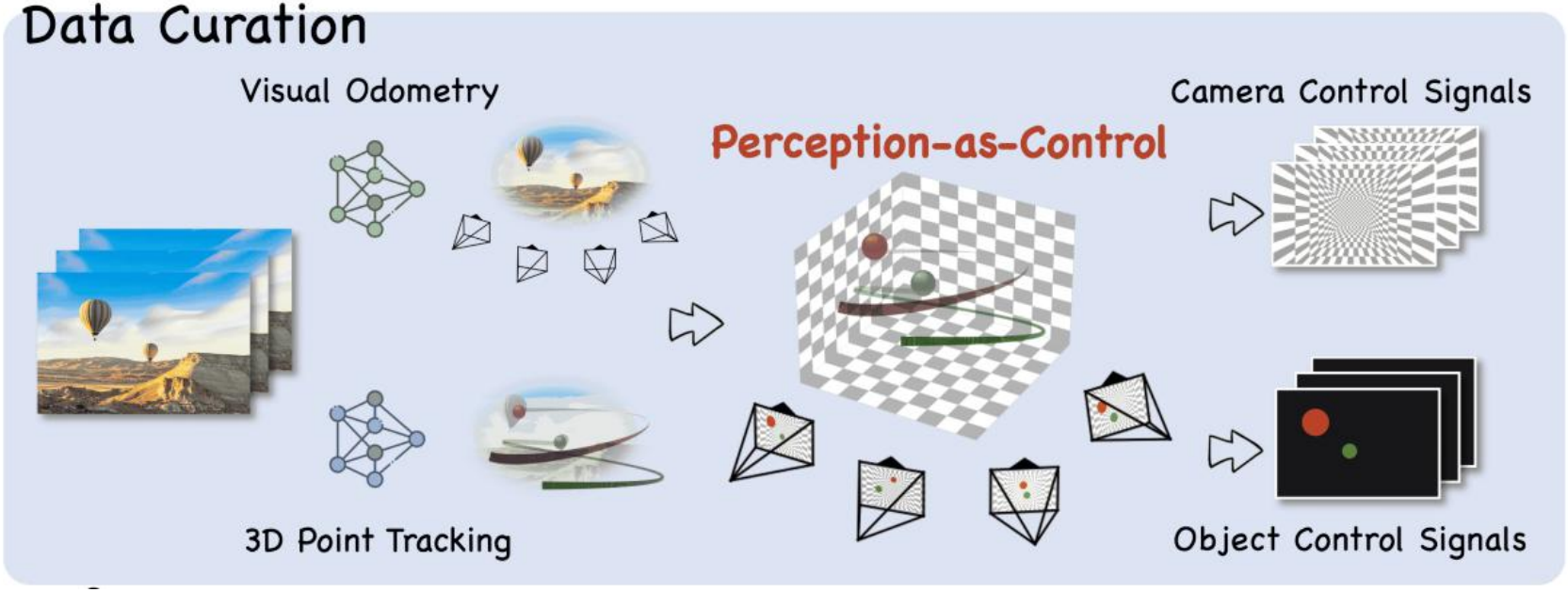
Outline

- Background
- Method
- Experiments
- Conclusion

Method

Perception-as-control

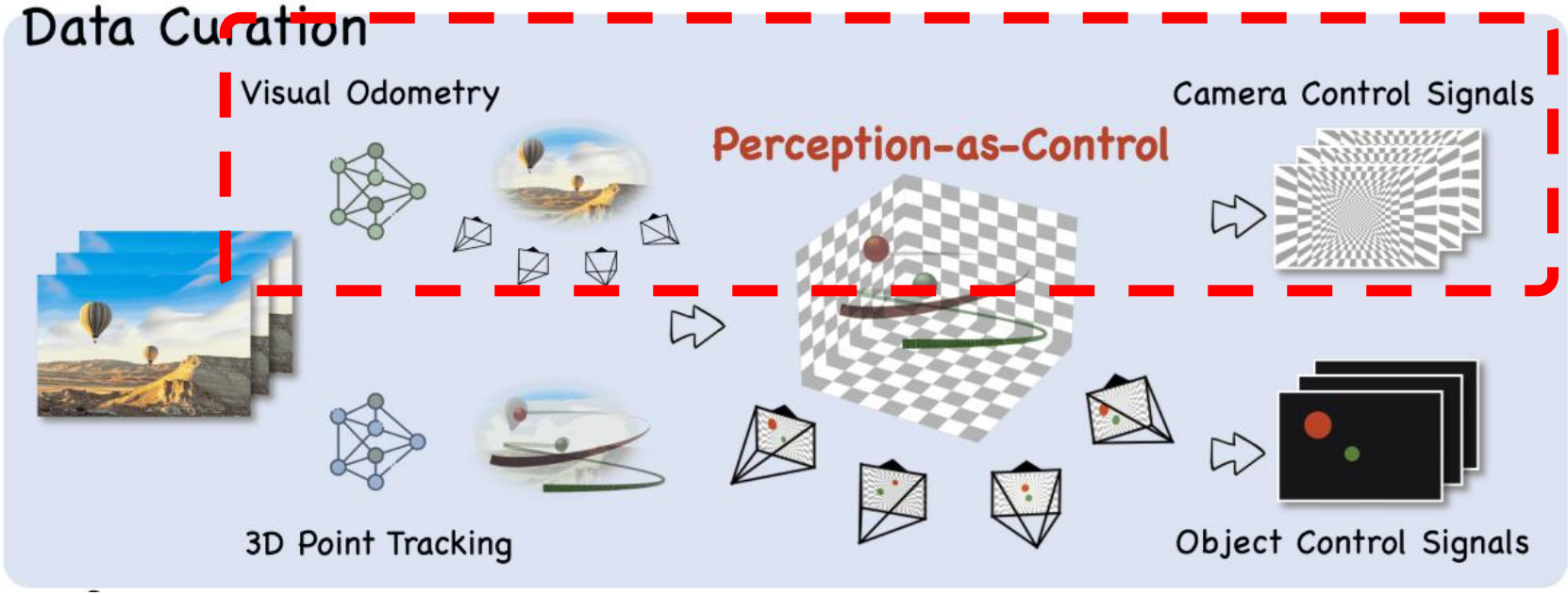
- Approximation of 3D world can support motion generation
  - Camera motion: extrinsic camera parameters sequence
  - Object motion: 3D sphere tracking



Method

Camera motion

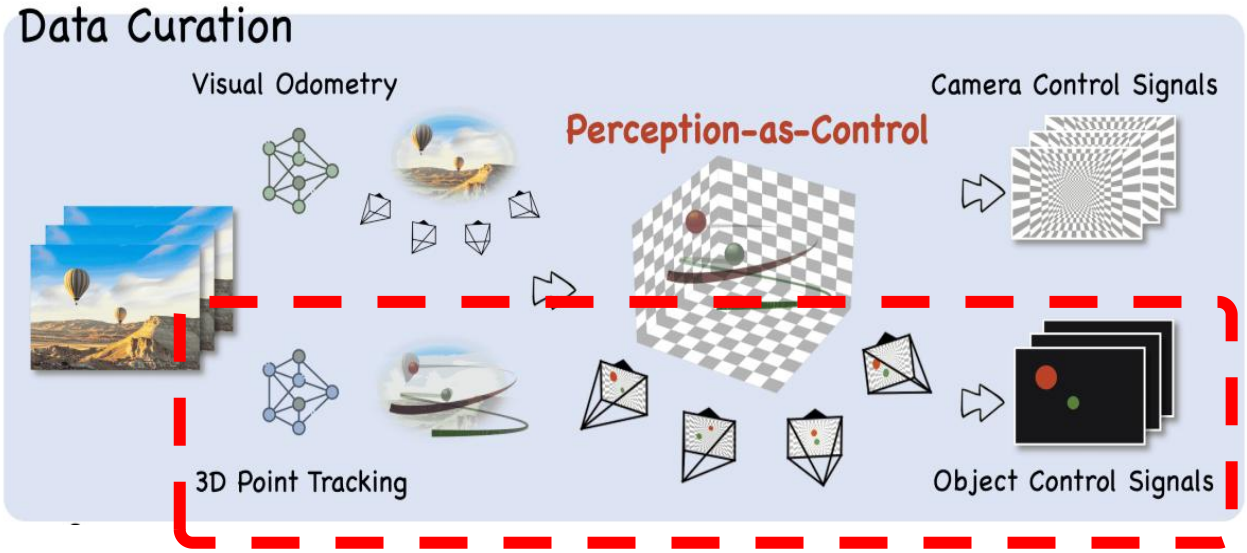
- Coarse estimation of camera extrinsic sequences $\{E_i\}_{i=1}^{T}$, $E_i = R_i T_i$
- Pre-defined intrinsic parameters
- Off-the-shelf toolbox: TartanVO [PMLR'21]

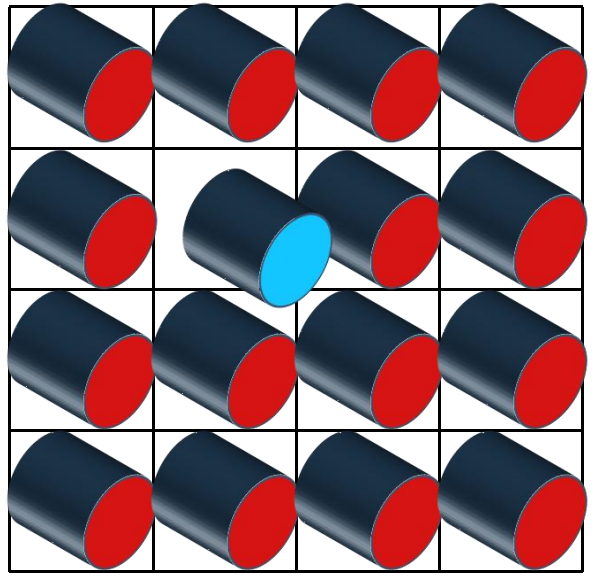


Method

Object motion

- Simplify 3D objects to **unit spheres** with adjustable grid points
- 3D point tracking → 2D point tracking & depth estimation
- Off-the-shelf toolbox: SpaTracker [CVPR'24]

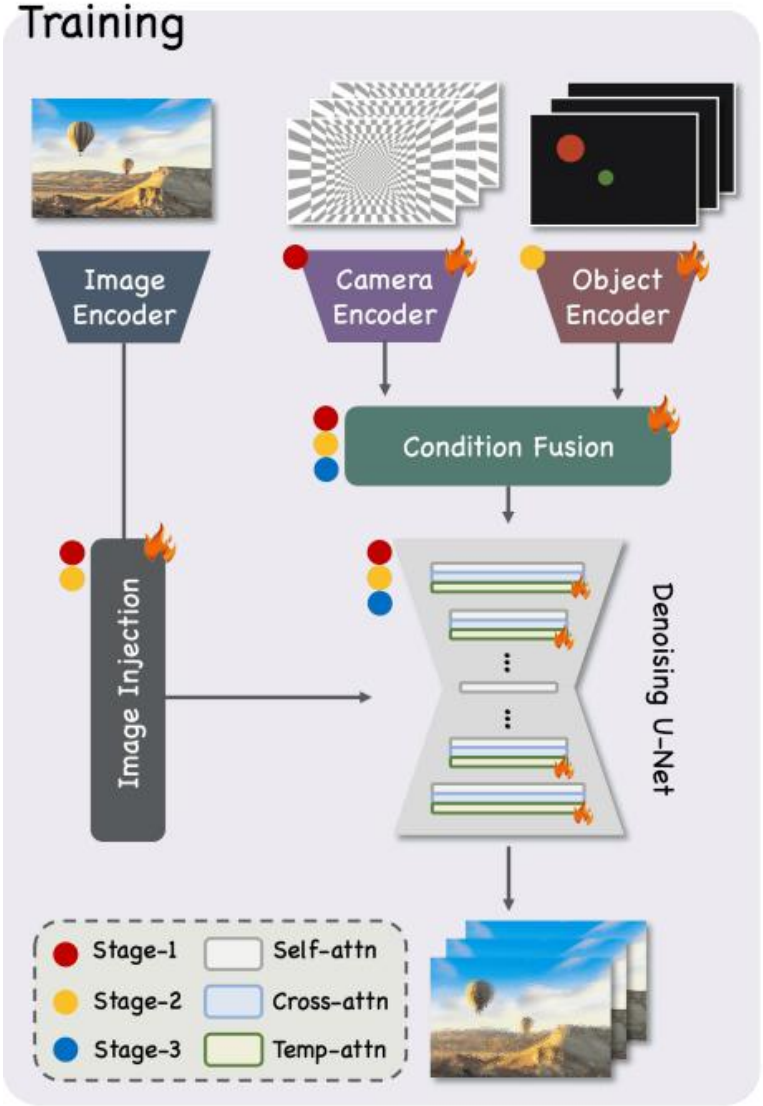




Method

Training strategy: 3 stage

- Condition-specific encoder
  - image, camera poses, object points
- Stage1: camera motion insertion
- Stage2: collaborative motion
- Stage3: dense-to-sparse fine-tune
  - For fine-grained control
  - Randomly sample points from long object motion trajectories



Method

Dataset

- Source:
  - RealEstate10K: with annotated camera intrinsic and explicit parameters
  - WebVid10M: large-scale in-the-wild motion video clips
- Filtering:
  - Estimate optical flow by RAFT [ECCV'20]
  - Compute F-Norm, discard small motion ones.
- Final: 6k with camera labels + 35k clips with rich scenes

Outline

- Background
- Method
- Experiments
- Conclusion

Experiments

Camera motions

- Regular move
- Arbitrary move

https://chen-yingjie.github.io/projects/Perception-as-Control/

Experiments

Regular move

https://chen-yingjie.github.io/projects/Perception-as-Control/

Experiments

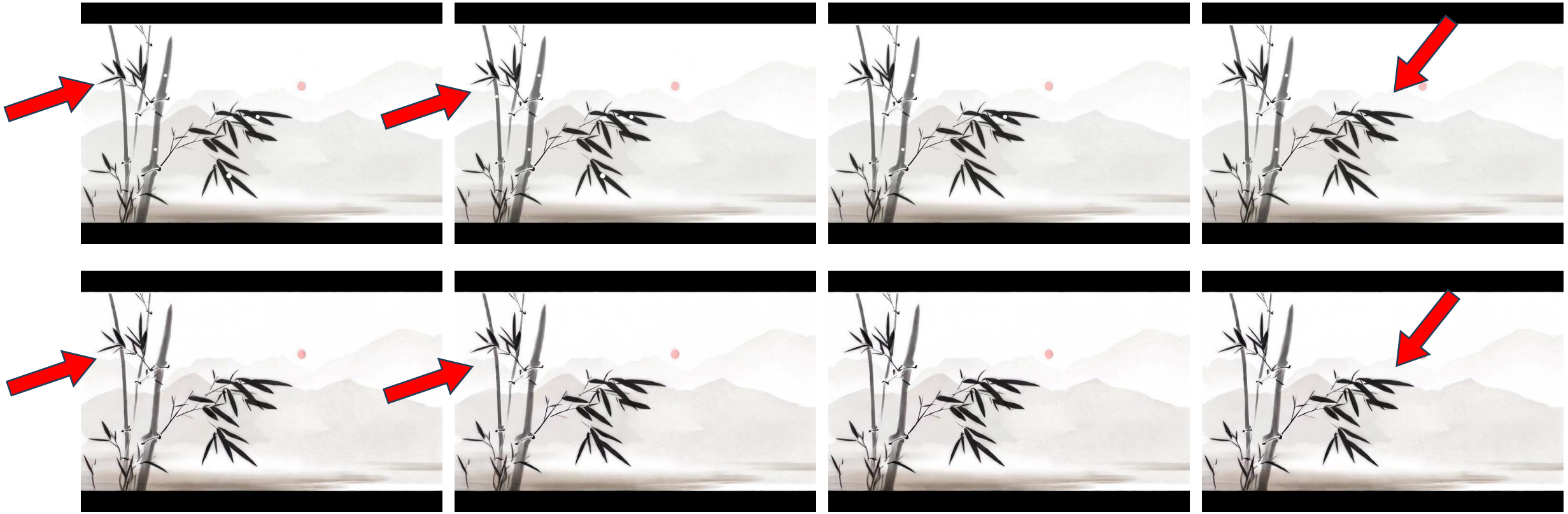
Arbitrary move

https://chen-yingjie.github.io/projects/Perception-as-Control/

Experiments

Object motions

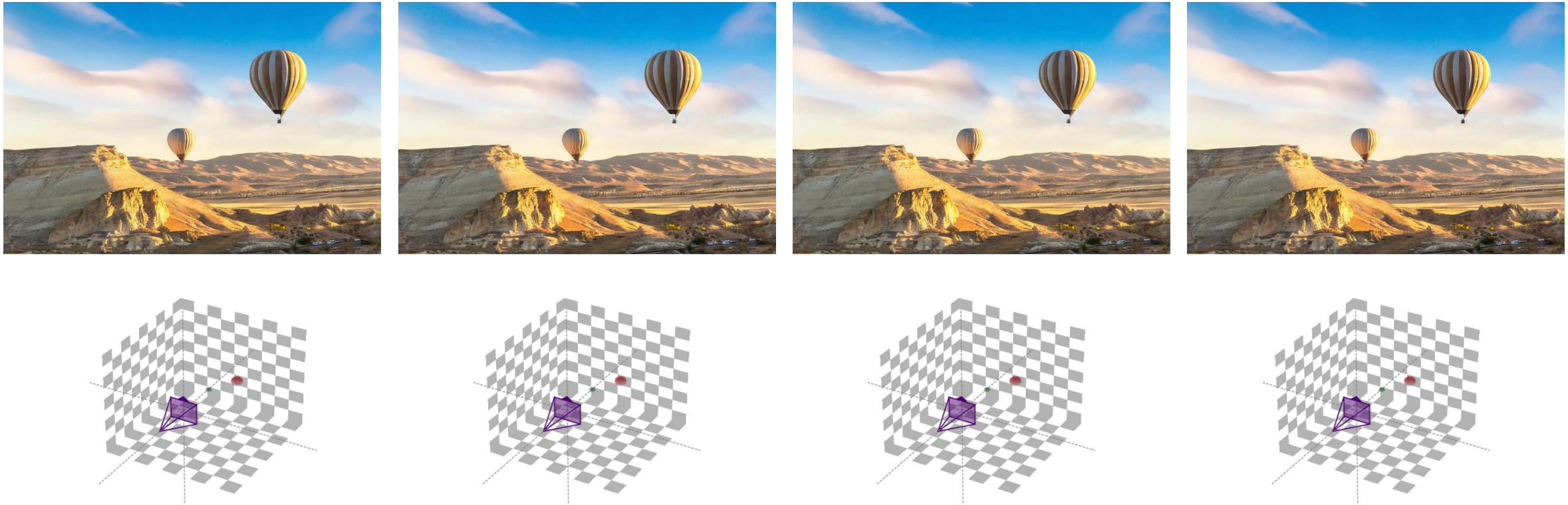
- Fine-grained control



https://chen-yingjie.github.io/projects/Perception-as-Control/

Experiments

Collaborative motions

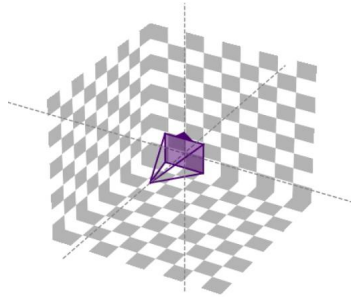


https://chen-yingjie.github.io/projects/Perception-as-Control/

Experiments

Further Applications: background motion control

- Even foreground is not static

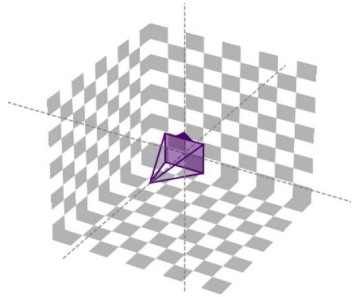


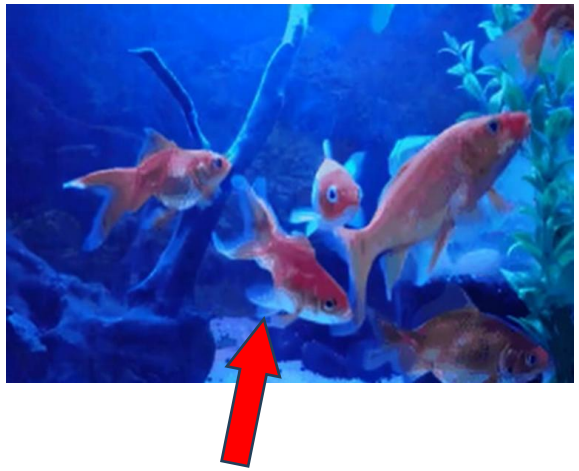
https://chen-yingjie.github.io/projects/Perception-as-Control/

Experiments

Further Applications: background motion control

- Even foreground is not static





https://chen-yingjie.github.io/projects/Perception-as-Control/

Outline

- Background
- Method
- Experiments
- Conclusion

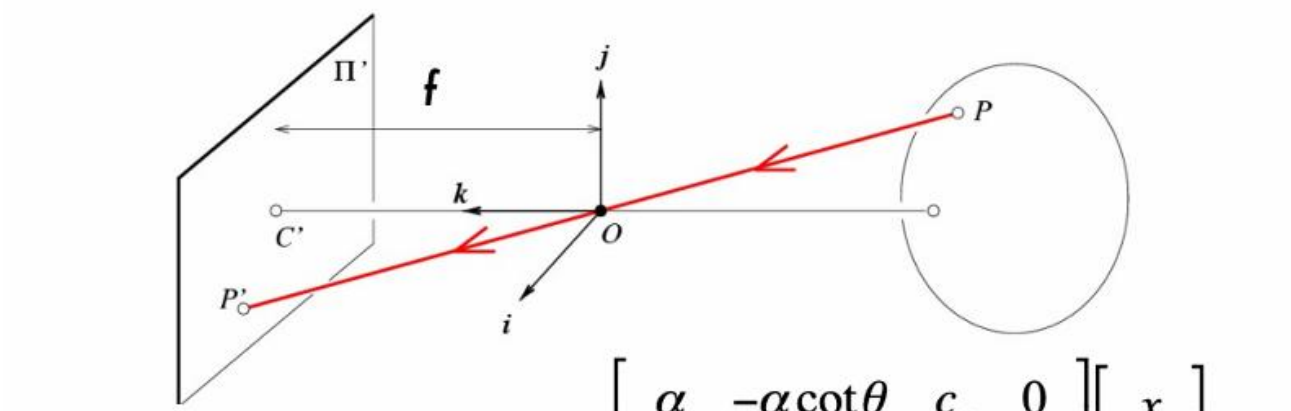
Conclusion

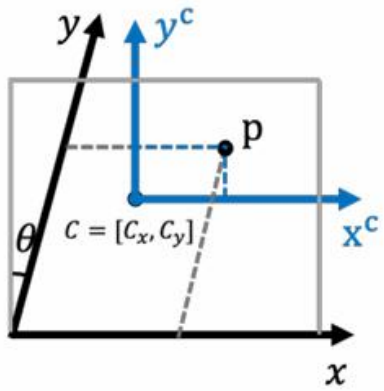
- Overview of camera basics
- An effective image animation method based on coarse 3D envelop
- Potential trending topics beyond video editing
- Potential applications to more low-/high-level tasks
  - Novel view synthesis, Multi-view understanding, …

Thanks for listening!

Conclusion: Potential Extension

Camera Skewness inside Intrinsic Matrix





$$P' = \begin{bmatrix} \alpha & -\alpha\cot\theta & c_x & 0 \\ 0 & \frac{\beta}{\sin\theta} & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

How many degrees of freedom does K have?

5 degrees of freedom!