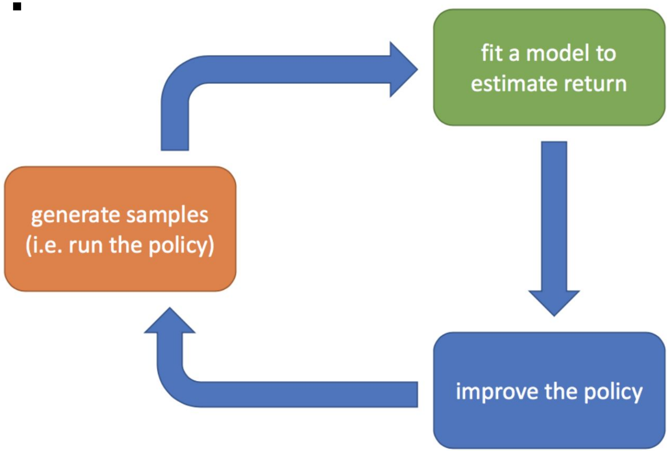
Generative and Predictive Models of Videos for Understanding the World

Oleh Rybkin

(some slides taken from Drew Jaegle, Karl Pertsch)

Can predictive objectives be useful for semantic understanding?



- Objects?
- Events?
- Affordances?

Credits: (left) Francis Vachon, found in a Chelsea Finn presentation, (right) Sergey Levine

Learning what you can do before doing anything

Oleh Rybkin*, Karl Pertsch*,
Konstantinos G. Derpanis, Kostas Daniilidis, Andrew Jaegle

ICLR 2019

Penn Engineering
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab

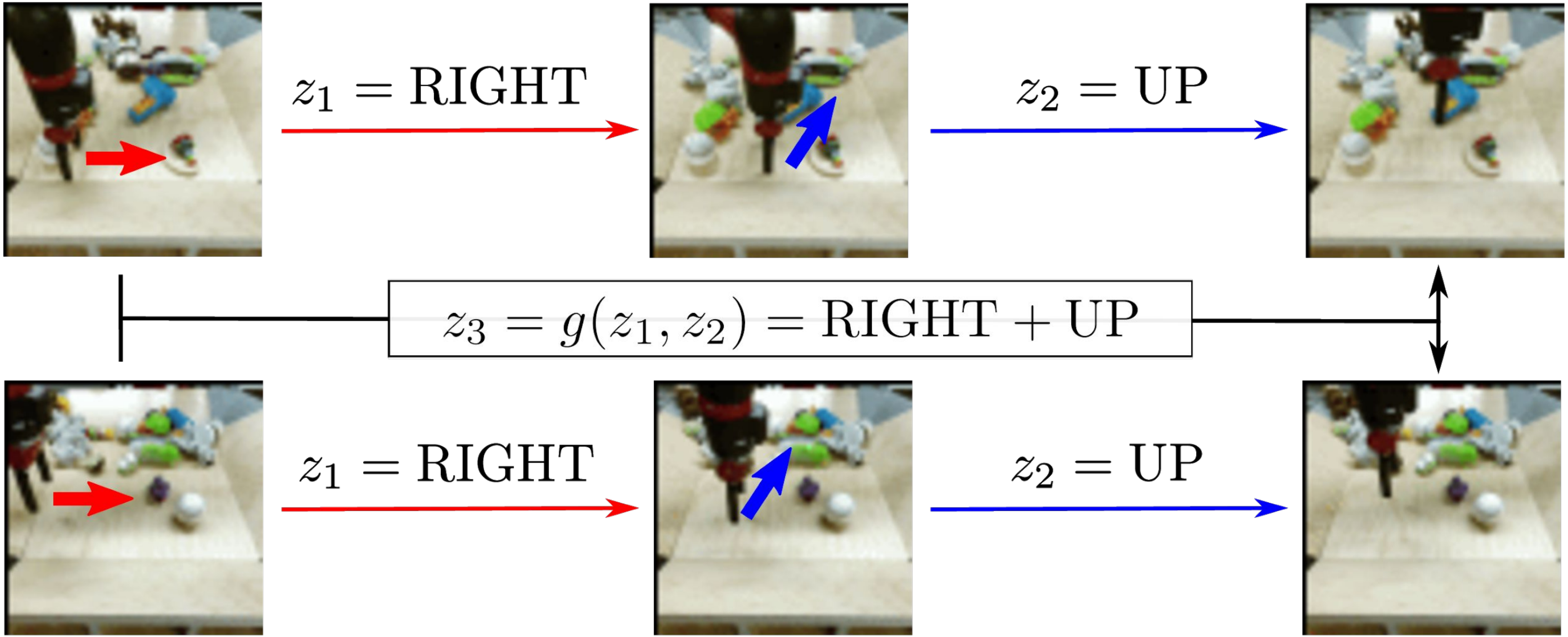
Ryerson University

University of Southern California 1880 Palmam qui meruit ferat

Understanding actions

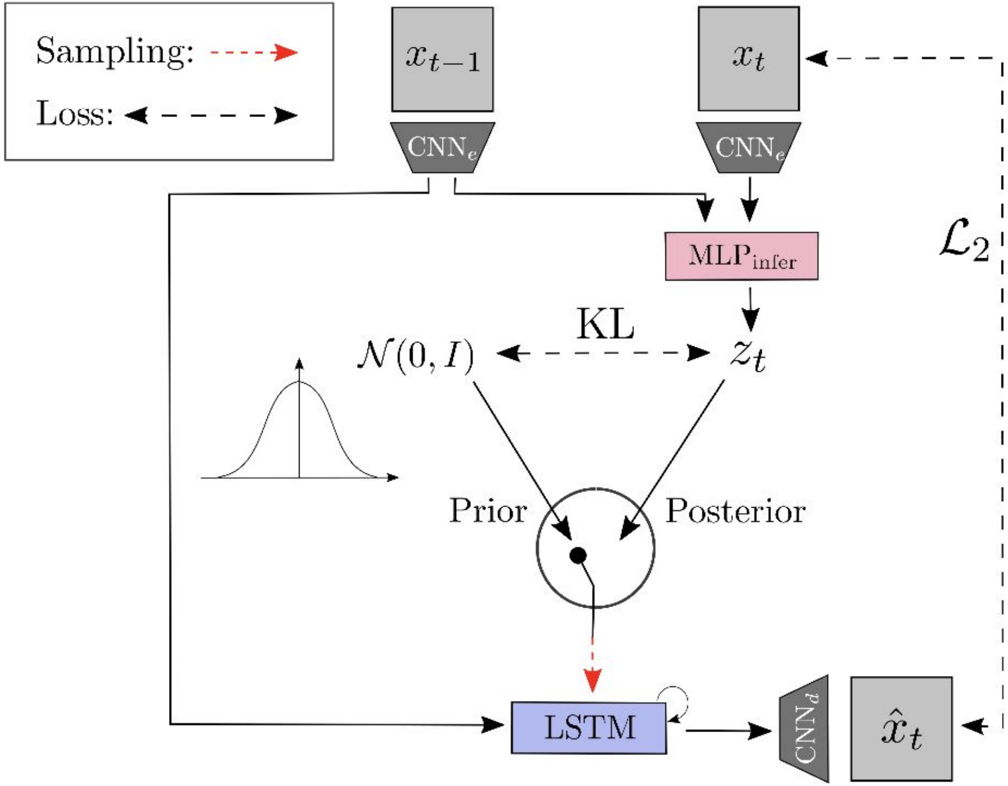
Credit: found via Nando De Freitas tweets

Understanding actions



Learning what you can do before doing anything. ICLR 2019.

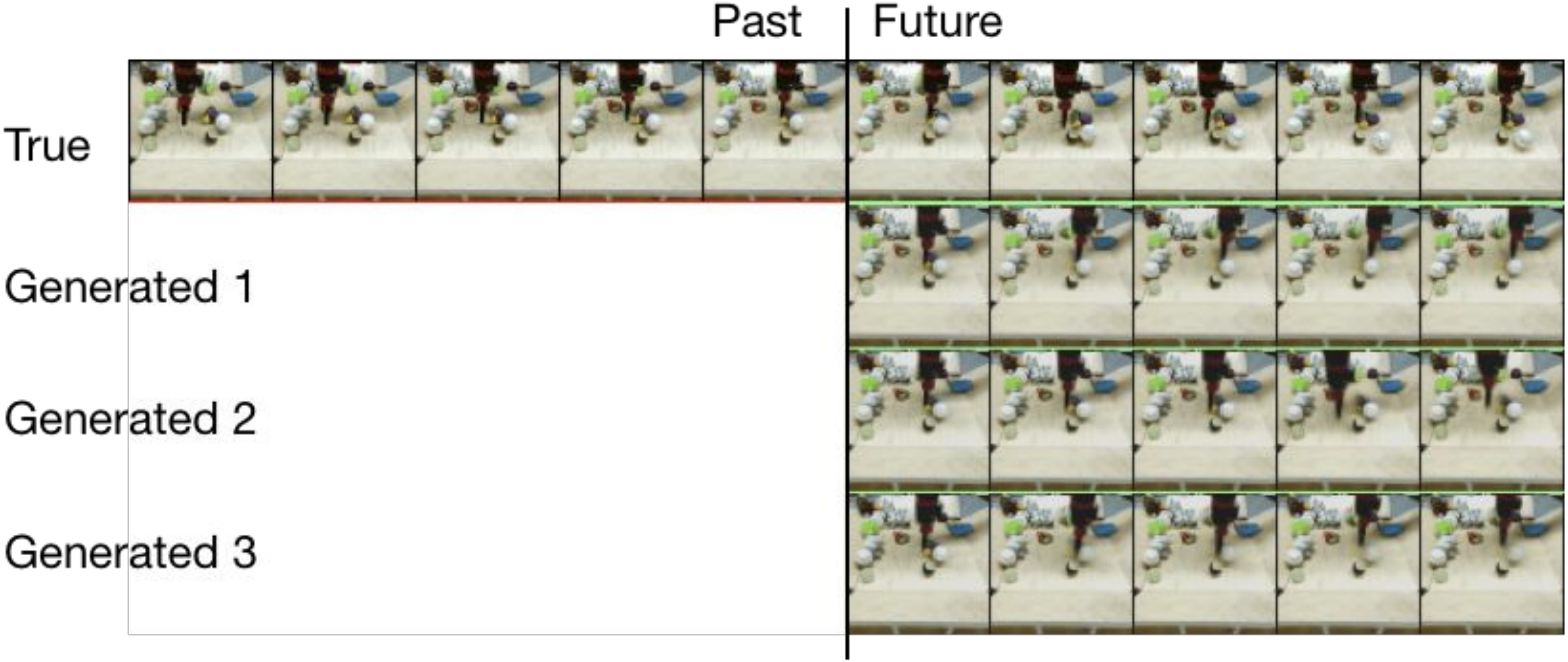
Variational Video Prediction



Denton & Fergus, 2018; Lee et al., 2018. Chung et al., 2015

Learning what you can do before doing anything. ICLR 2019.

Variational Video Prediction



Learning what you can do before doing anything. ICLR 2019.

Variational Video Prediction with Information Bottleneck

The (beta-)VAE objective for stochastic video prediction is:

$$\sum_t \left[\mathbb{E}_{q(z_t|x_{t-1:t})} \log p(x_t|Z_t, x_{t-1}) - \beta \mathrm{KL}[q(Z_t|x_{t-1:t}), p(Z)]\right]$$

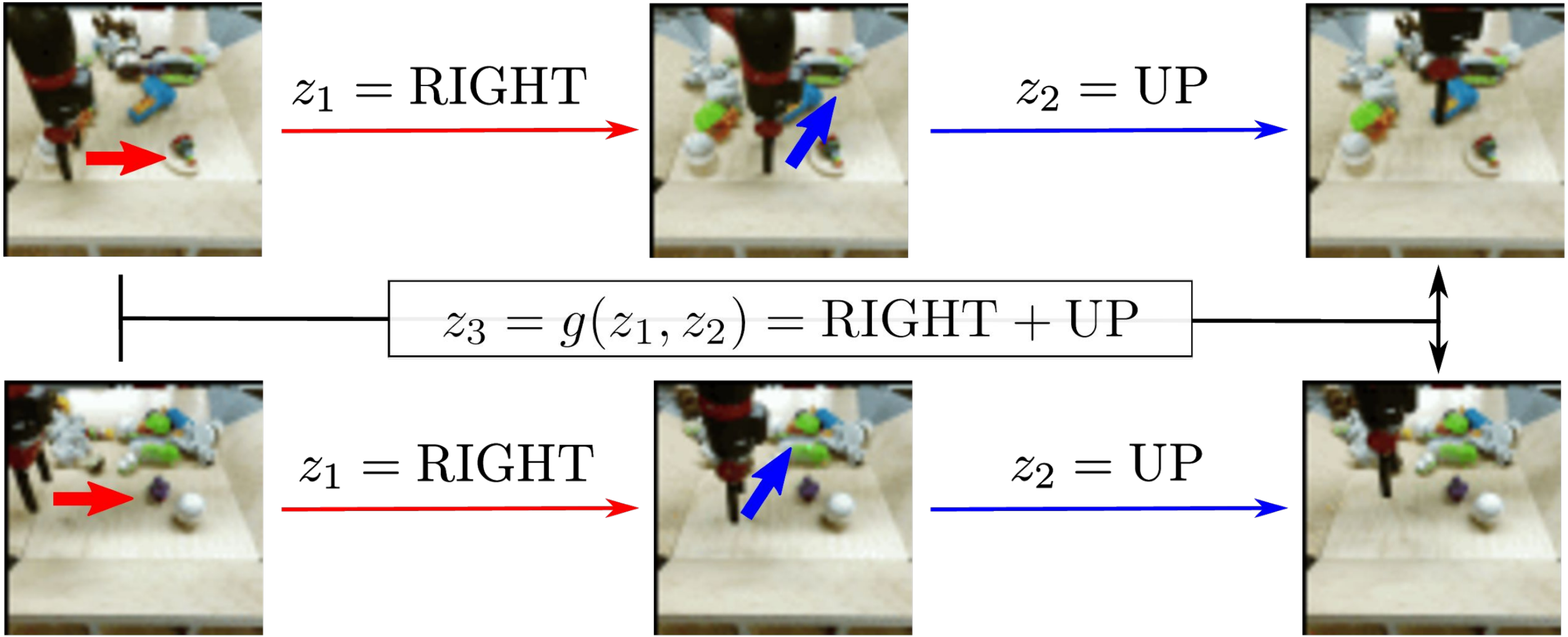
Which is equivalent to the VIB lower bound of the following:

$$\max I((z_t, x_{t-1}); x_t) \text{ s.t. } I(z_t; x_{t-1:t}) \leq I_c.$$

Alemi et al., 2017, 2018

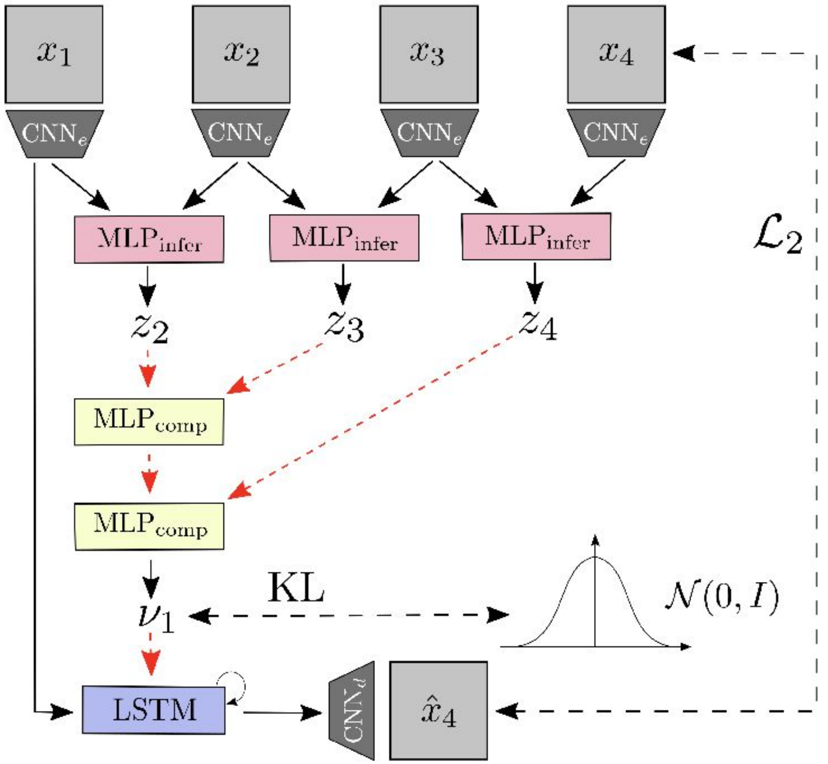
Learning what you can do before doing anything. ICLR 2019.

Enforcing structure with composability



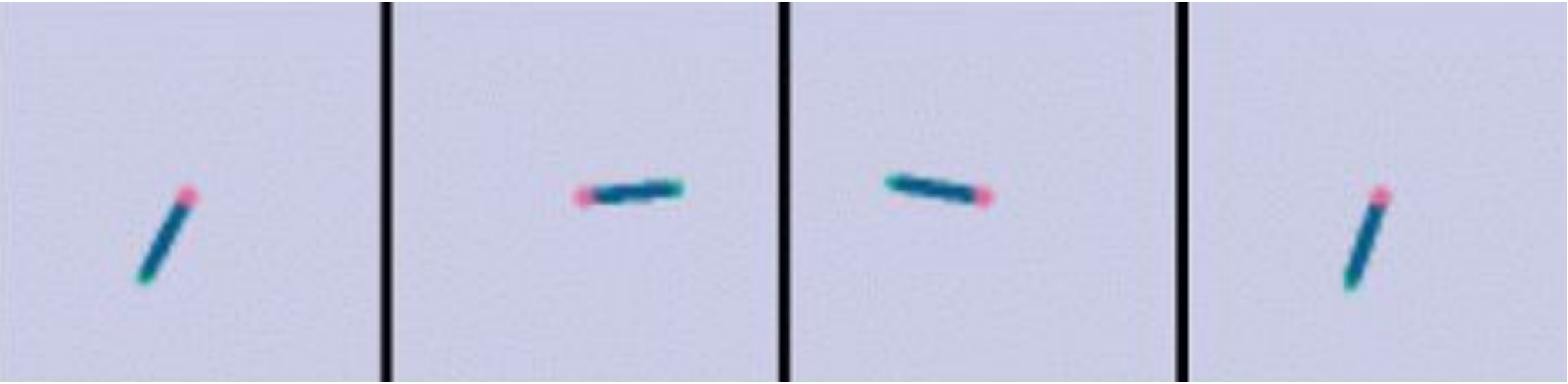
Learning what you can do before doing anything. ICLR 2019.

CLASP: Enforcing structure with composability



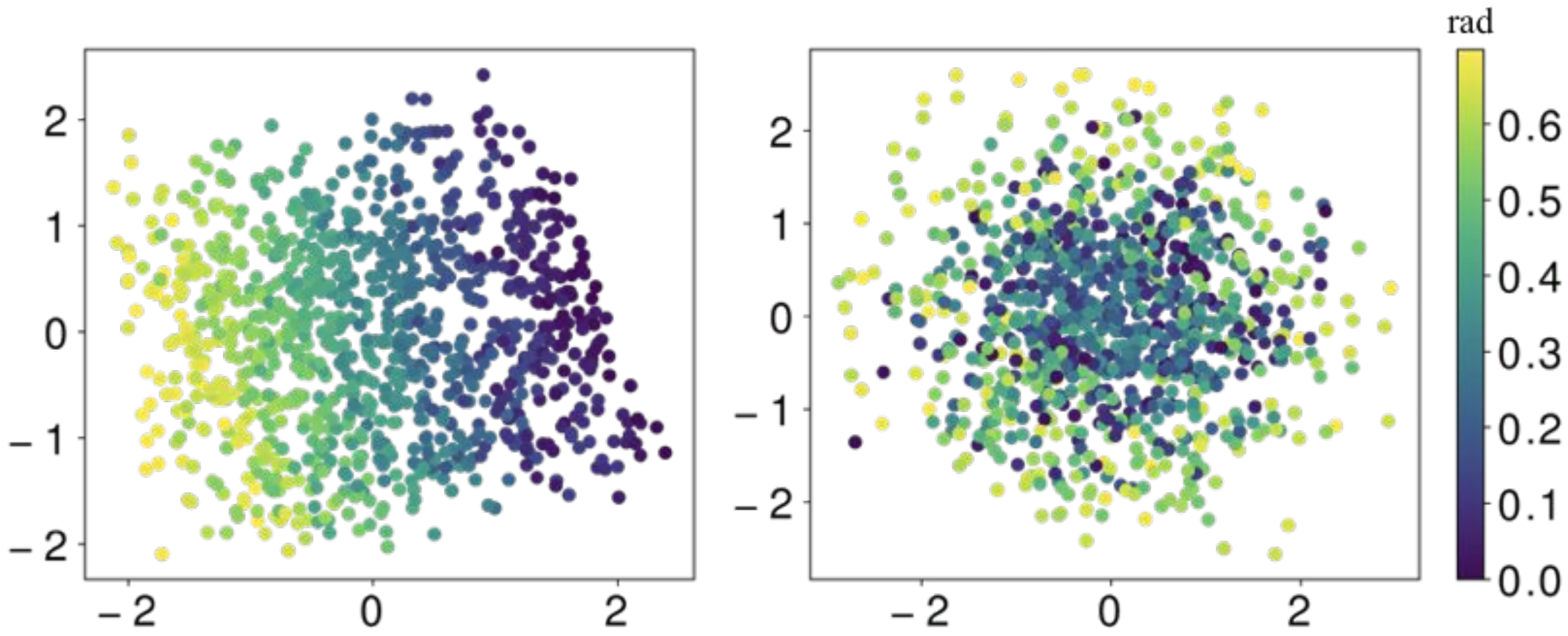
Learning what you can do before doing anything. ICLR 2019.

Reacher environment



Learning what you can do before doing anything. ICLR 2019.

Understanding actions

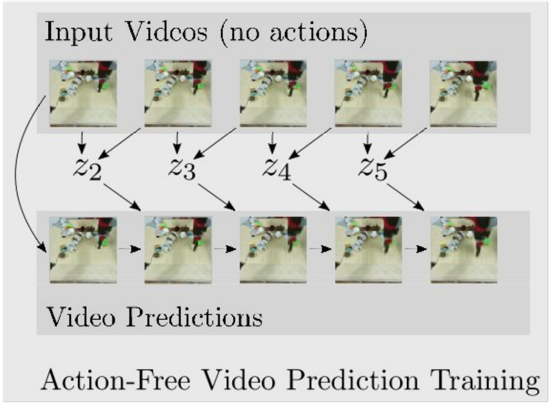


a) CLASP (Ours) **b)** Denton & Fergus (2018)

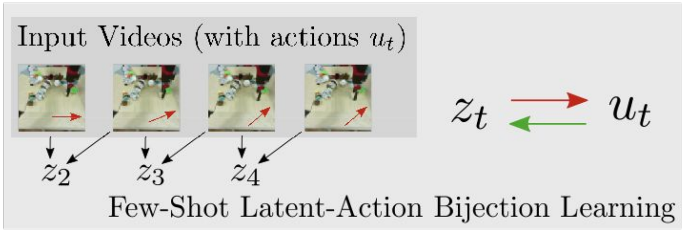
Learning what you can do before doing anything. ICLR 2019.

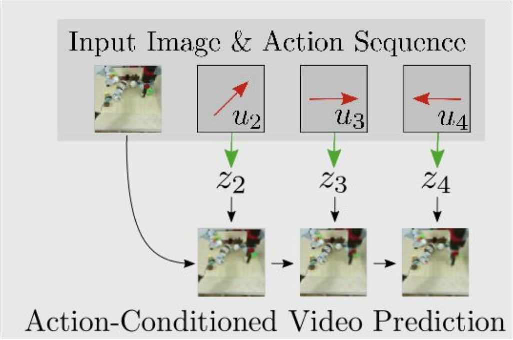
Applications of CLASP

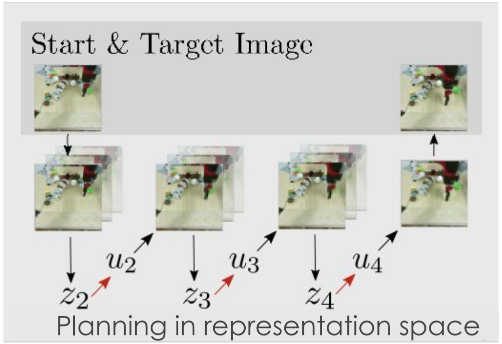
Passive learning



Active learning







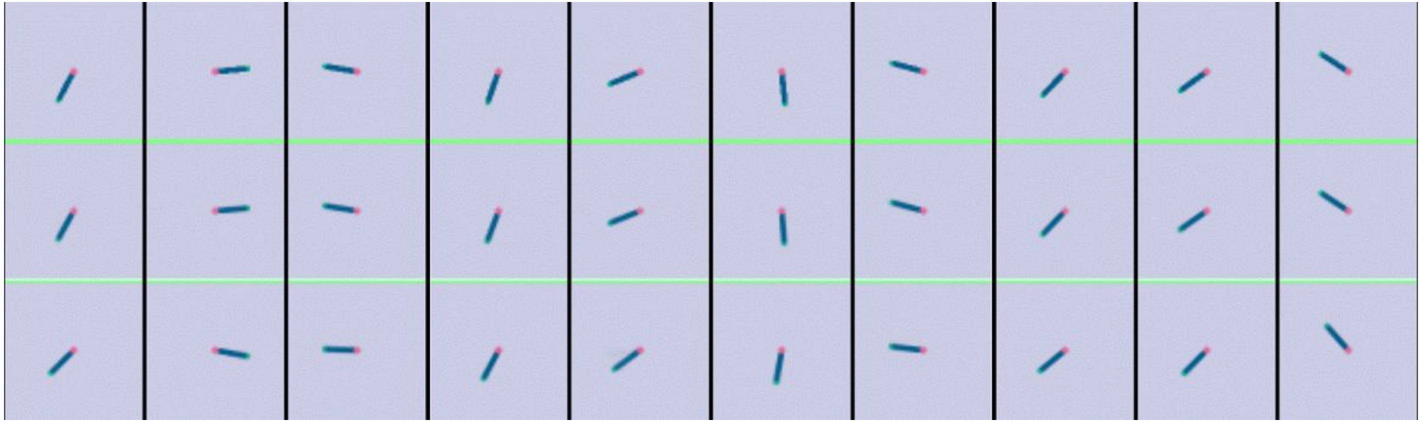
Learning what you can do before doing anything. ICLR 2019.

Action-conditioned prediction

Ground Truth:

CLASP (ours):

Denton & Fergus:

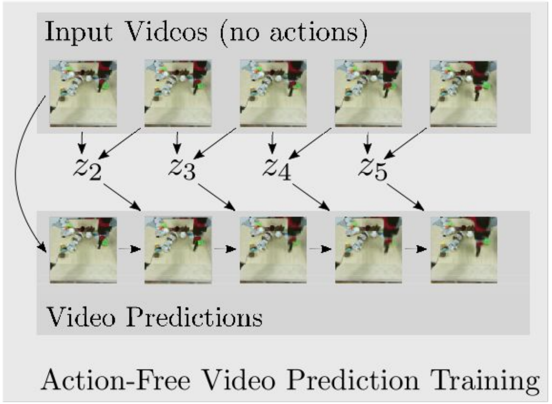


	Reacher	BAIR
Method	Error [deg]	Error [px]
Random	26.6 ± 21.5	-
Baseline	22.6 ± 17.7	3.6 ± 4.0
Ours	$\mathbf{2.9 \pm 2.1}$	$\mathbf{3.0 \pm 2.1}$
Supervised	2.6 ± 1.8	2.0 ± 1.3

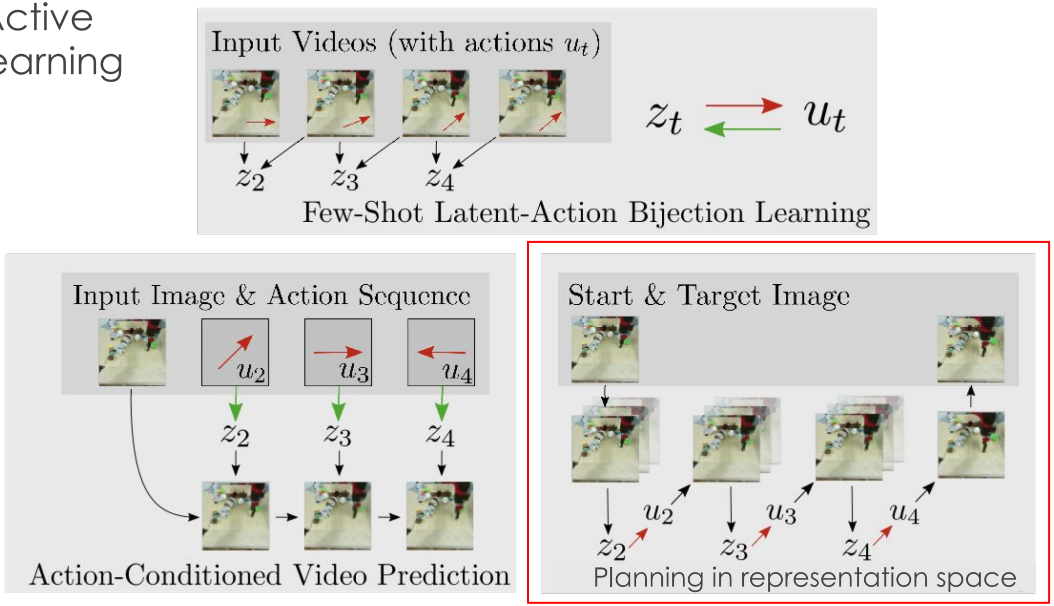
Learning what you can do before doing anything. ICLR 2019.

Applications of CLASP

Passive learning



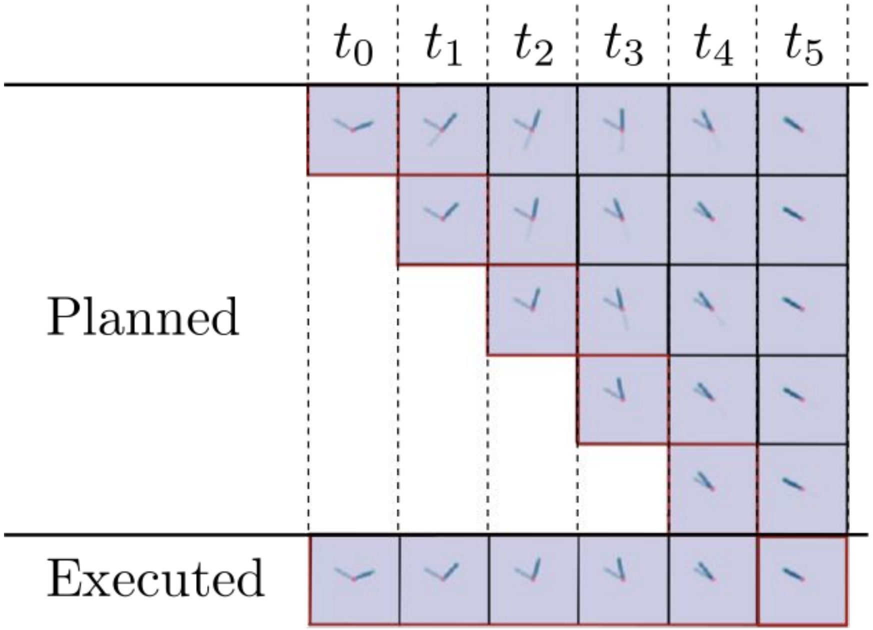
Active learning

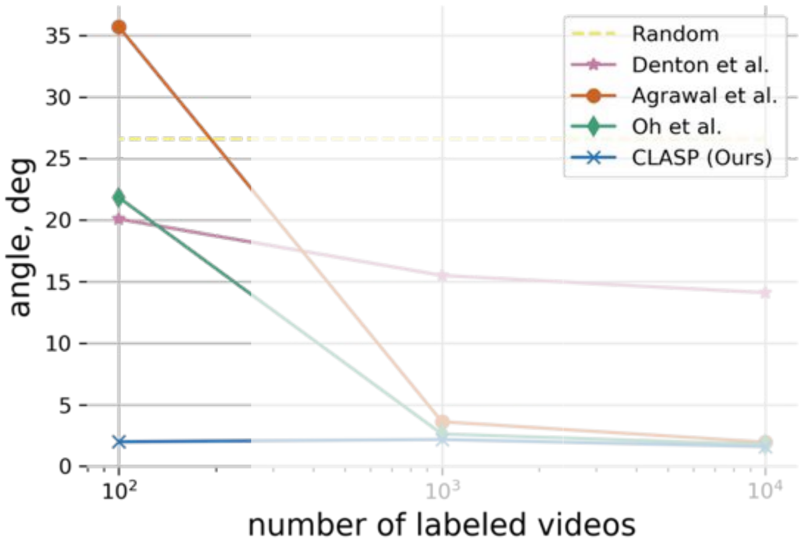


Ebert et Finn et al., 2018

Learning what you can do before doing anything. ICLR 2019.

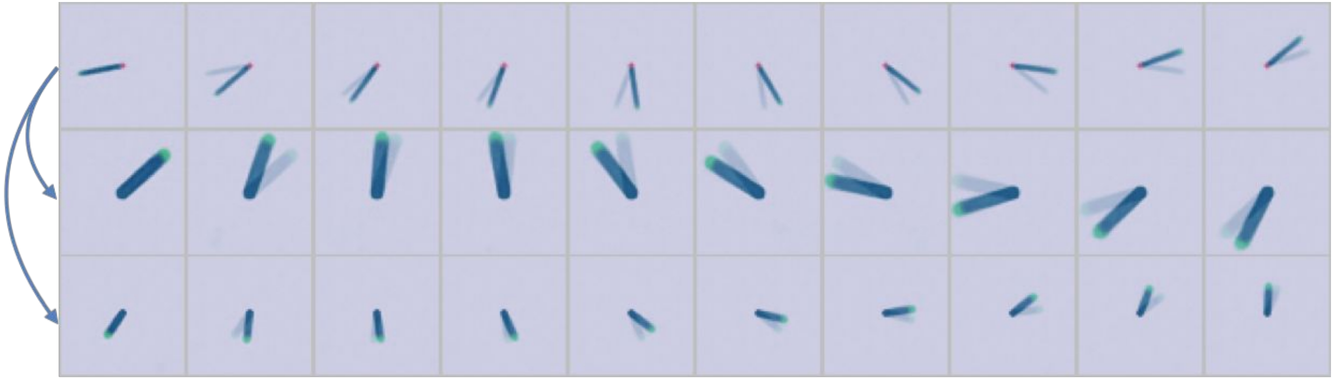
Planning in learned latent space

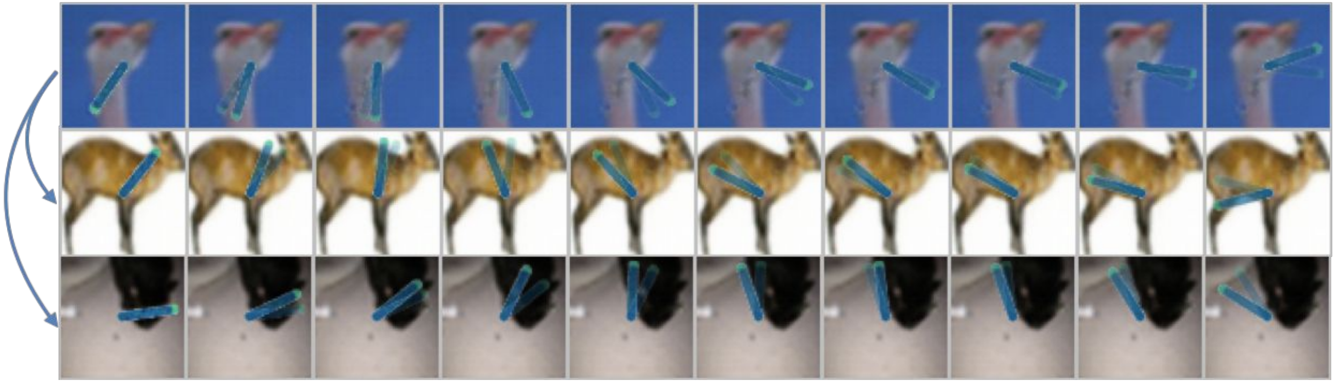




Learning what you can do before doing anything. ICLR 2019.

Varying visual characteristics

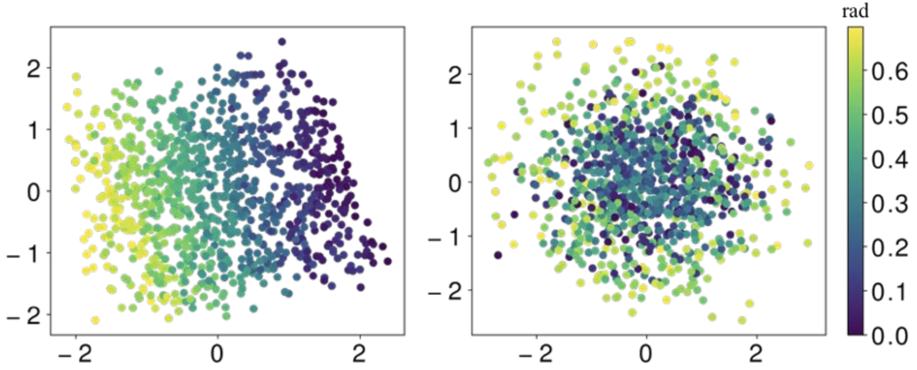




Learning what you can do before doing anything. ICLR 2019.

Learning what you can do before doing anything

1. The *inductive biases* of minimality and composability provide sufficient constraints for learning action representations just from visual observations
2. The learned representation is *disentangled* from the static scene content and visual characteristics of the environment.
3. The representation to be used for *planning* and *action-conditioned prediction* while requiring orders of magnitude less action-labeled videos.



a) CLASP (Ours) **b)** Denton & Fergus (2018)

Karl Pertsch*

Kosta Derpanis

Kostas Daniilidis

Andrew Jaegle

Learning what you can do before doing anything. ICLR 2019.

KeyIn: Discovering Subgoal Structure with Keyframe-based Video Prediction

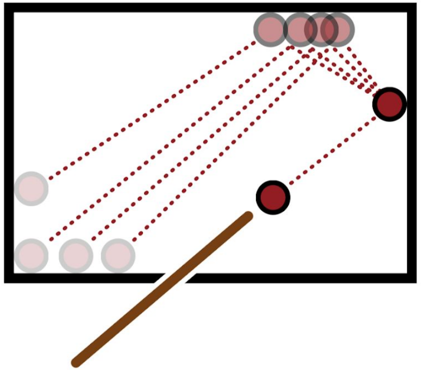
Karl Pertsch*, Oleh Rybkin*, Jingyun Yang,
Konstantinos G. Derpanis, Joseph Lim, Kostas Daniilidis,
Andrew Jaegle

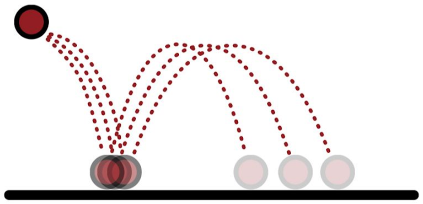
Penn Engineering
GRASP LABORATORY
General Robotics, Automation, Sensing & Perception Lab

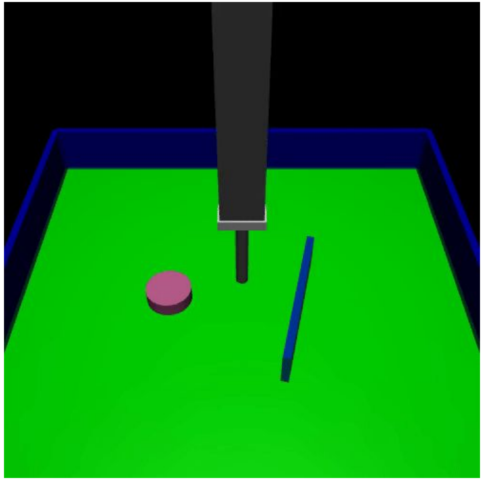
Ryerson University

University of Southern California

Keyframes in natural sequences



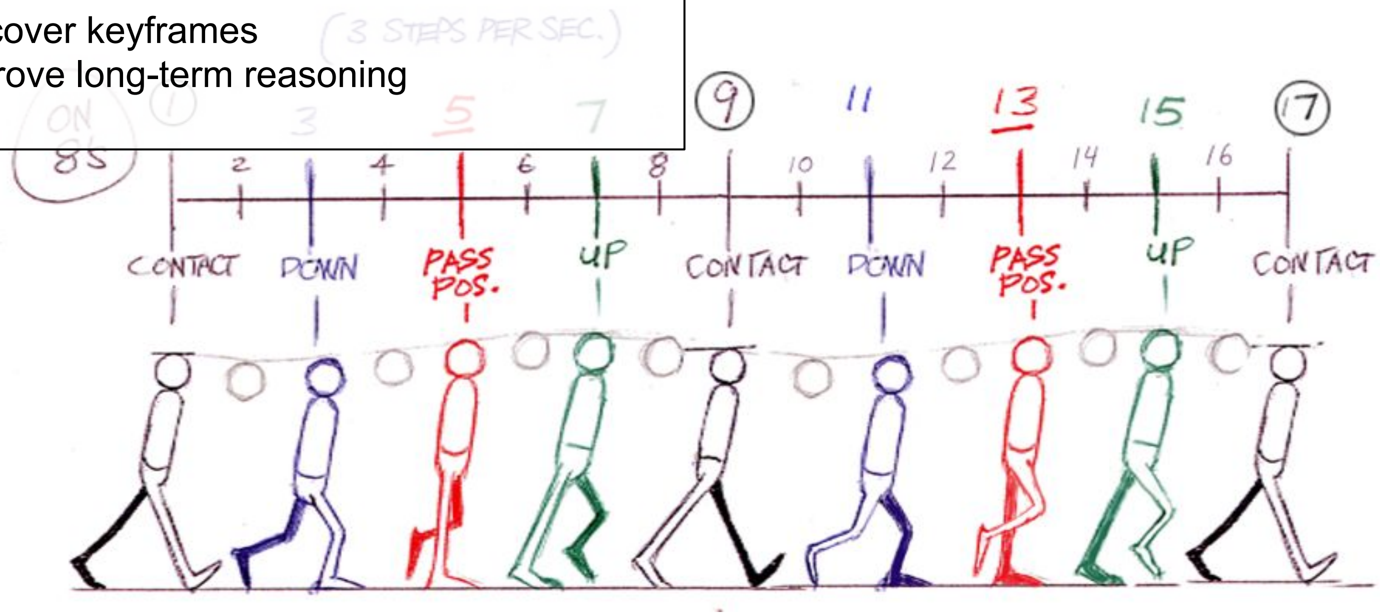




- Dynamics in complex scenes are stochastic. But not uniformly so!
- How can we exploit this structure to improve long-term reasoning?
- **Keyframes**: capture interesting structure in time, but also allow reconstruction of the full dynamics.

Keyframing

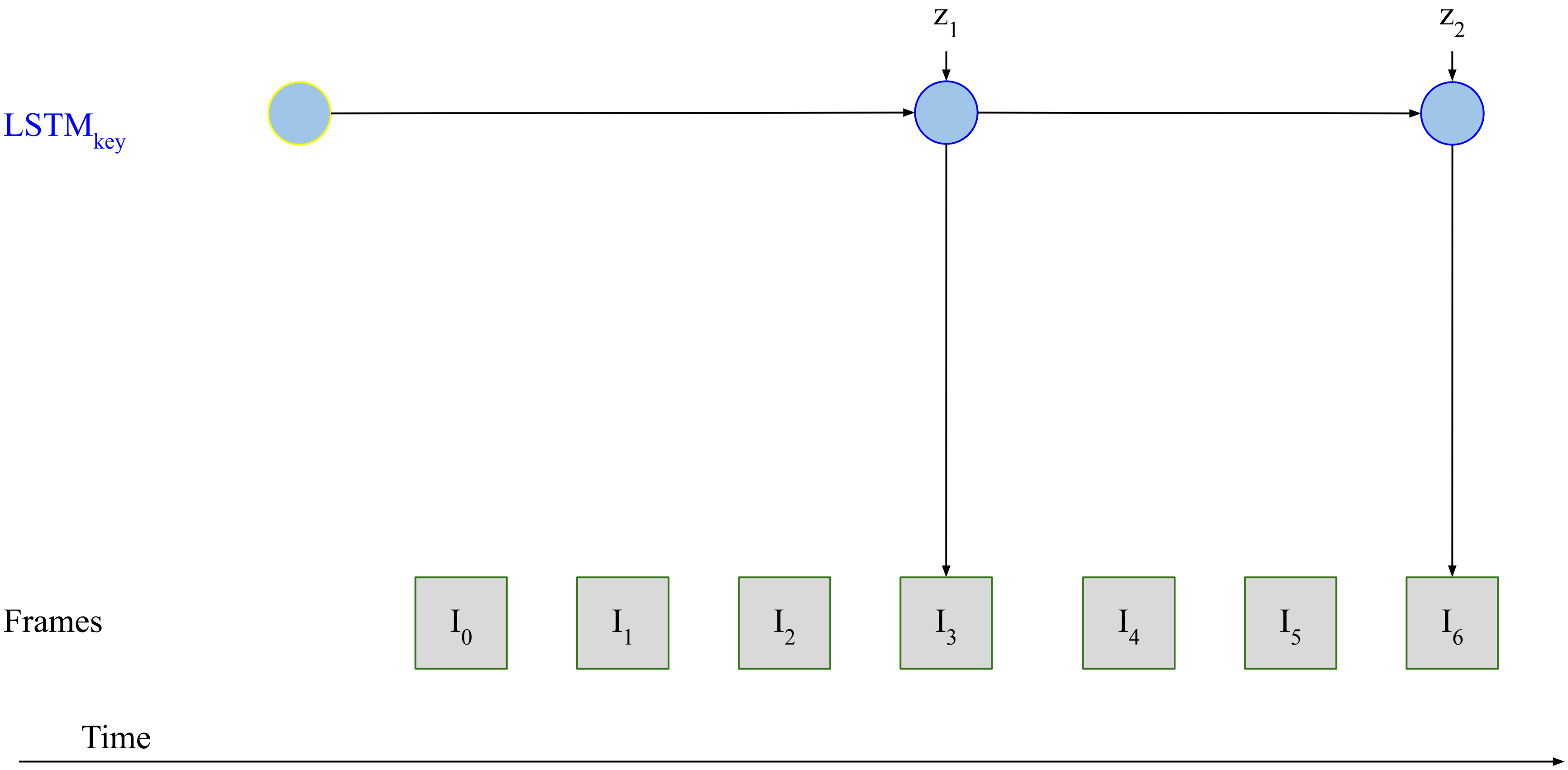
- Discover keyframes
- Improve long-term reasoning



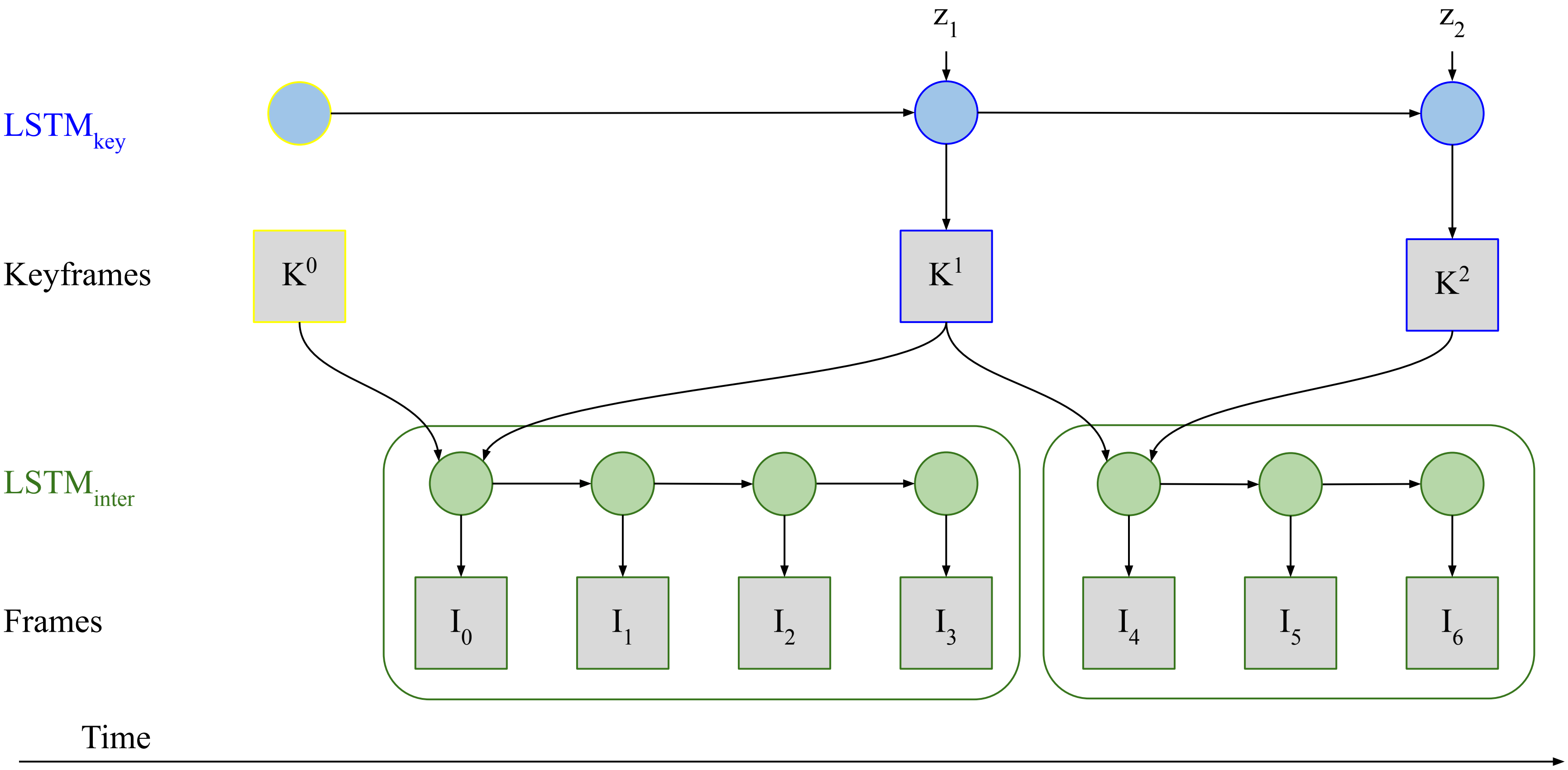
1. Draw the start and end points of all motions: define the stochastic long-term sequence dynamics (*lead animator*).
2. Interpolate between the start and end points: make the local, deterministic dynamics explicit (*inbetweener*).

Image credit: Emily Wakefield (source)

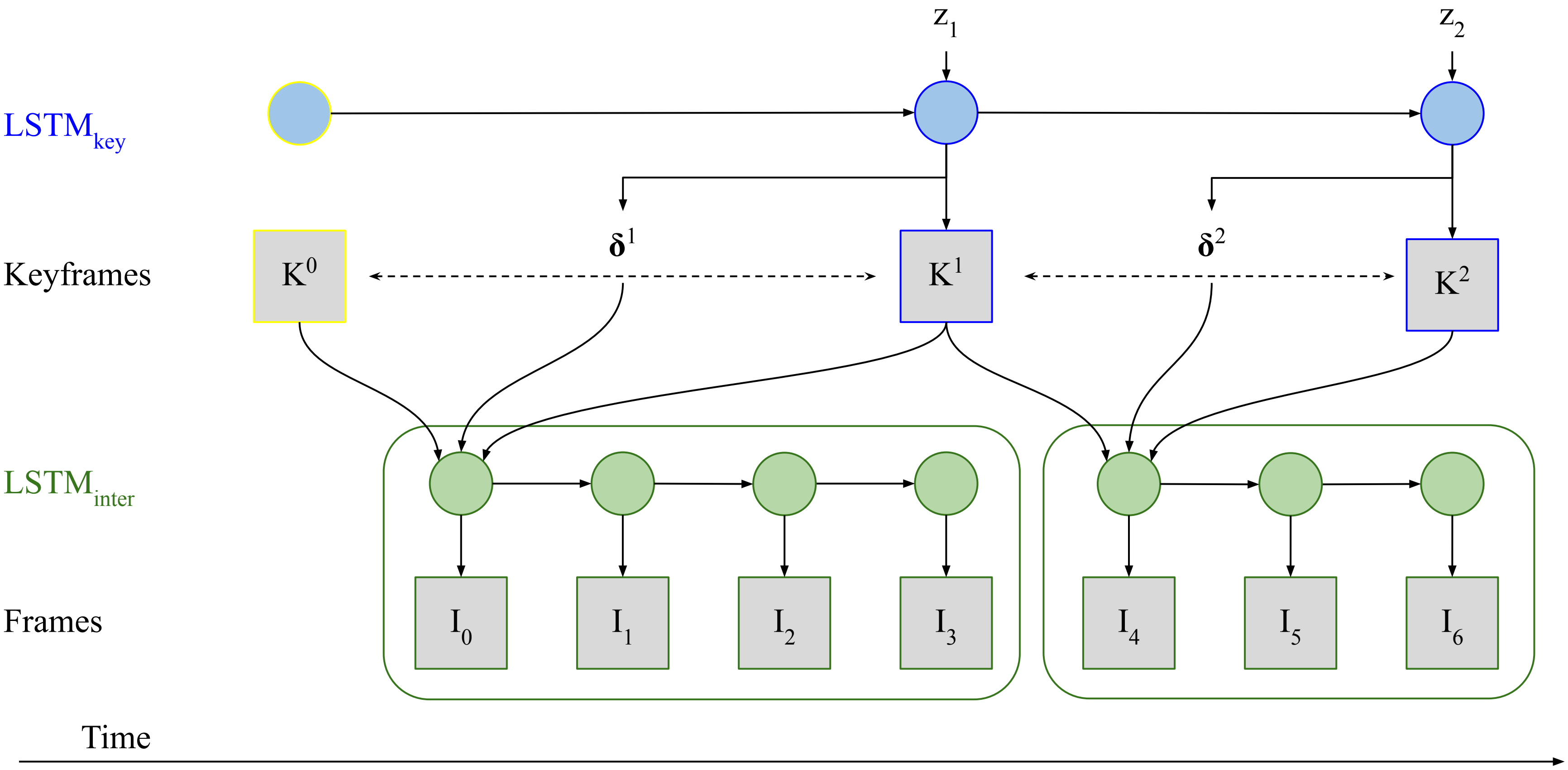
KeyIn - keyframe prediction



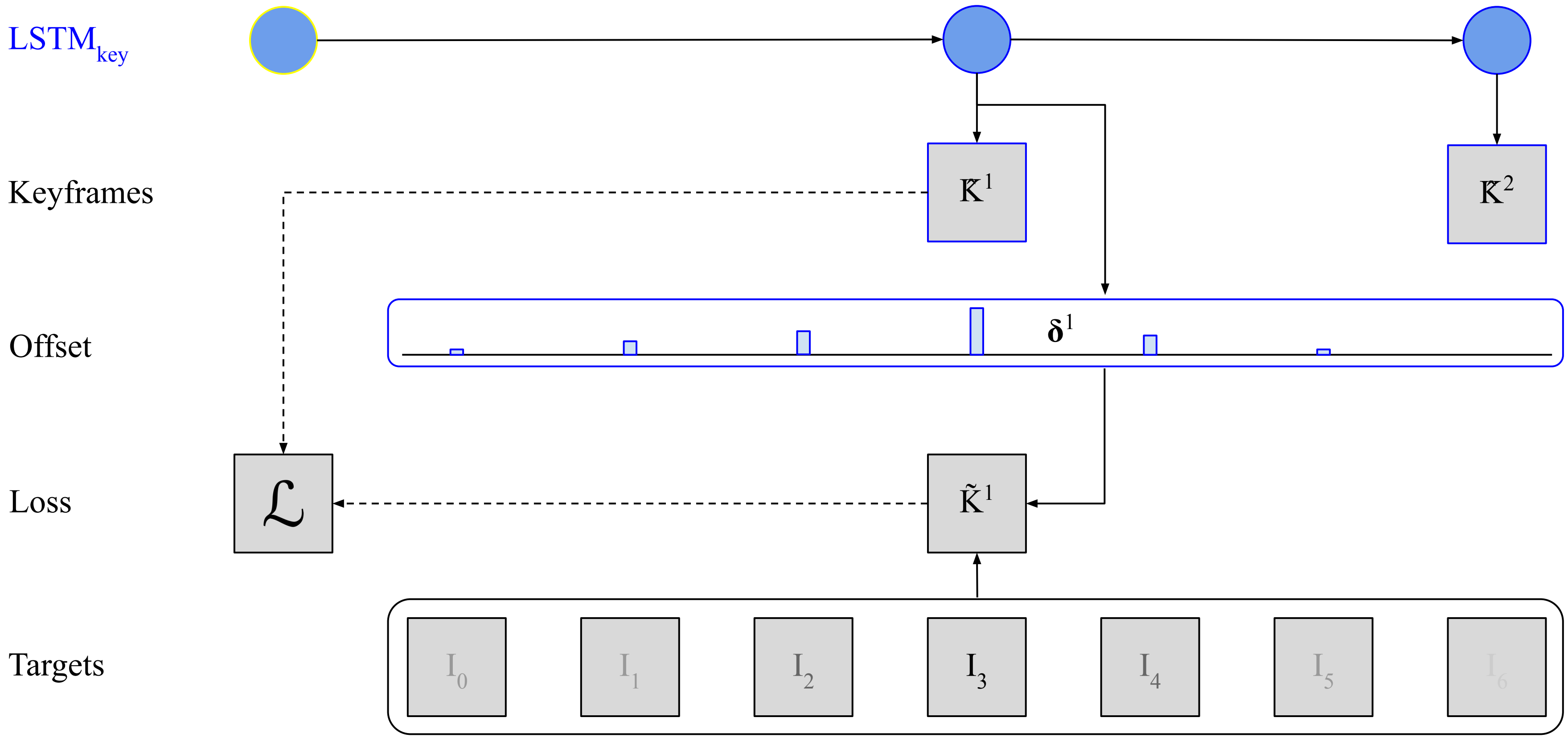
KeyIn - keyframe-based prediction



KeyIn - predicting interframe offsets



KeyIn - Continuous relaxation



KeyIn -Full loss

Soft Keyframe targets

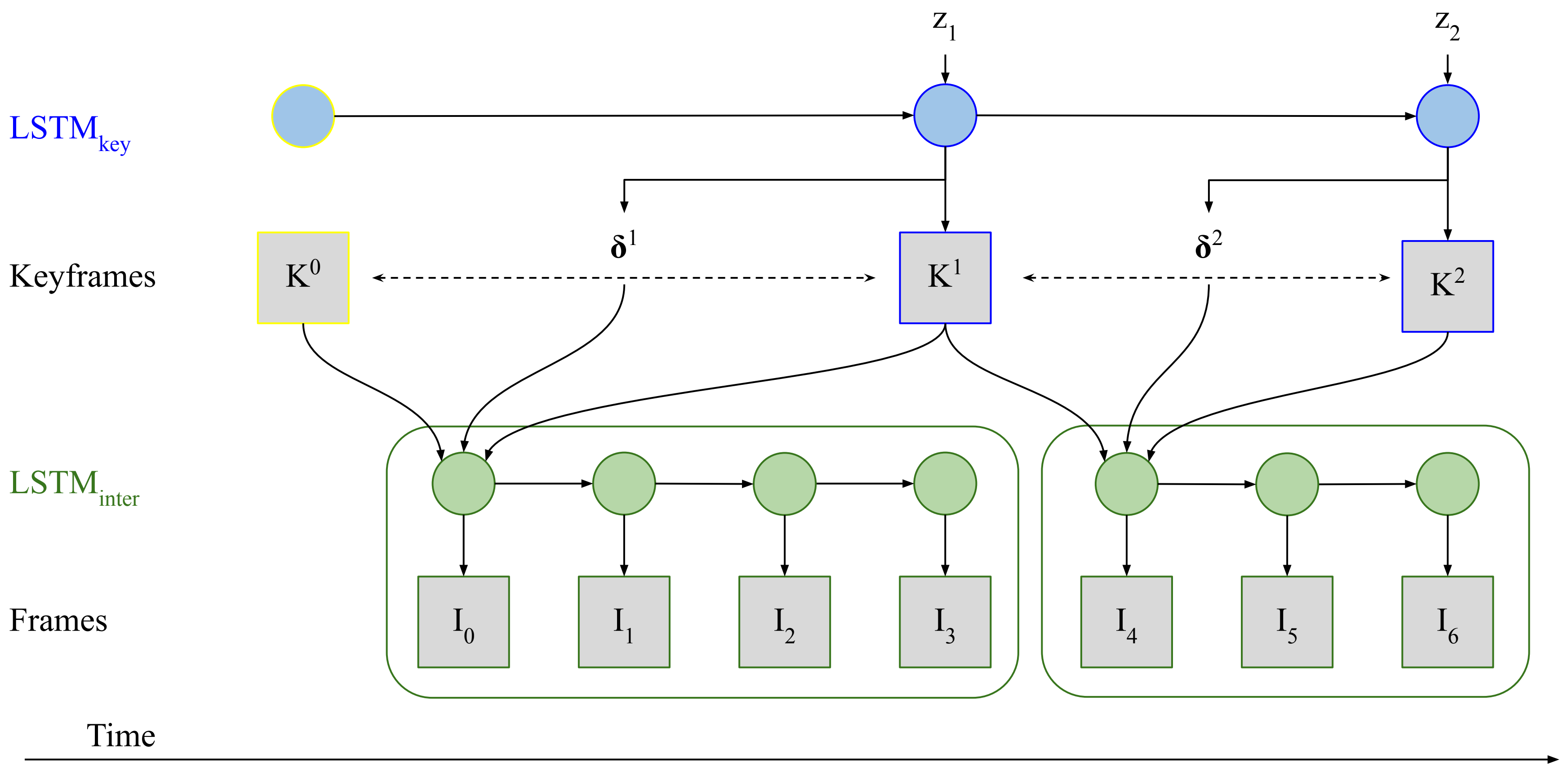
$$\mathcal{L}_{key} =(\sum_t c^t \beta_{ki} ||\hat{K}^t - \tilde{K}^t||^2$$

Soft embedding targets

Prior divergence

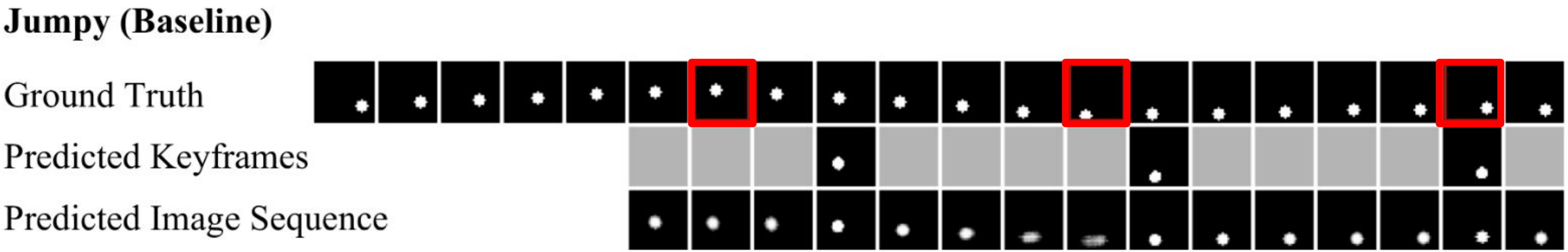
Interpolation targets

KeyIn - full method

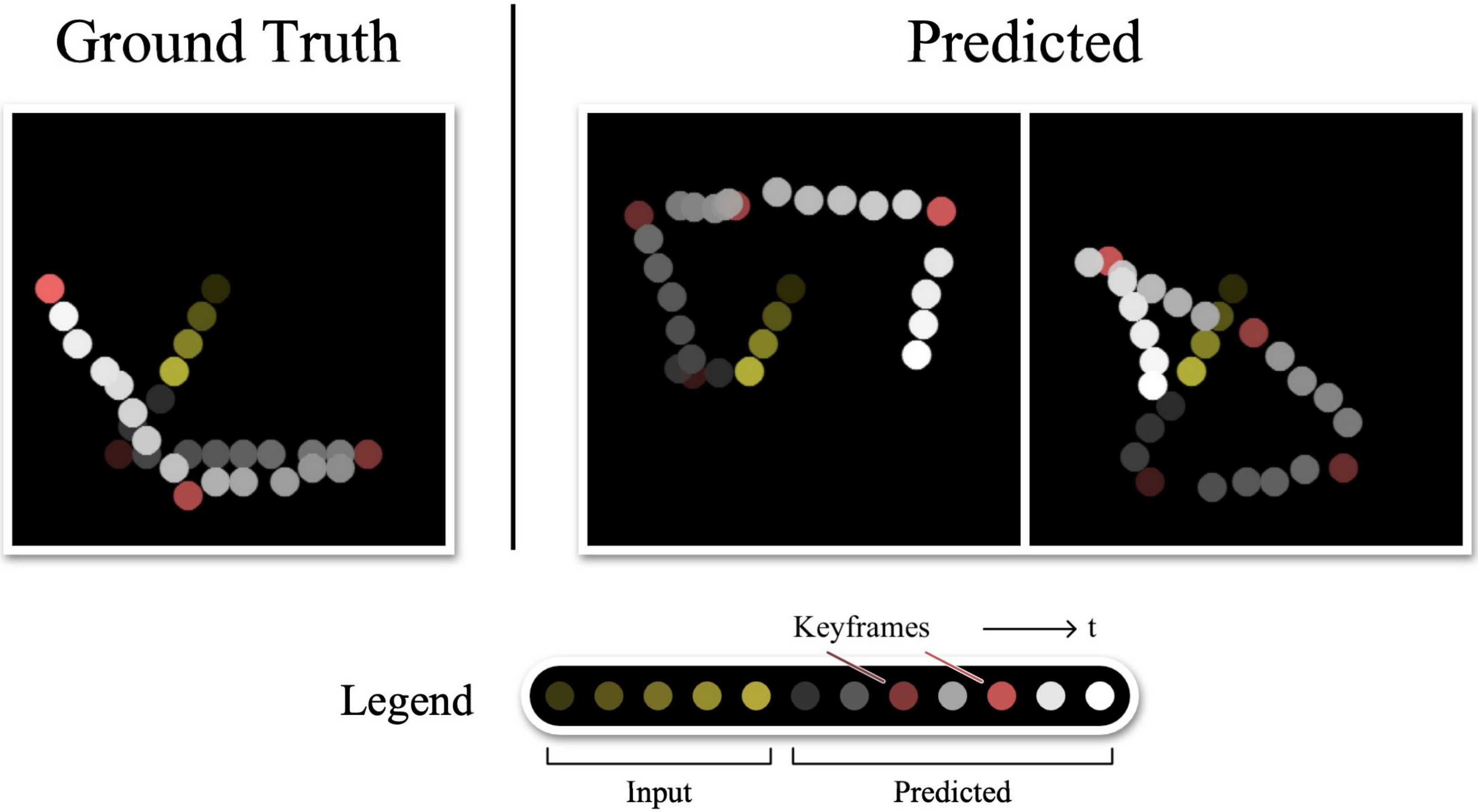


Structured Brownian motion data

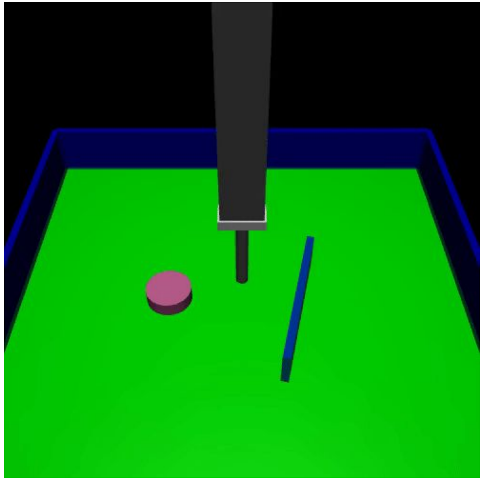
Enforcing descriptive Keyframes

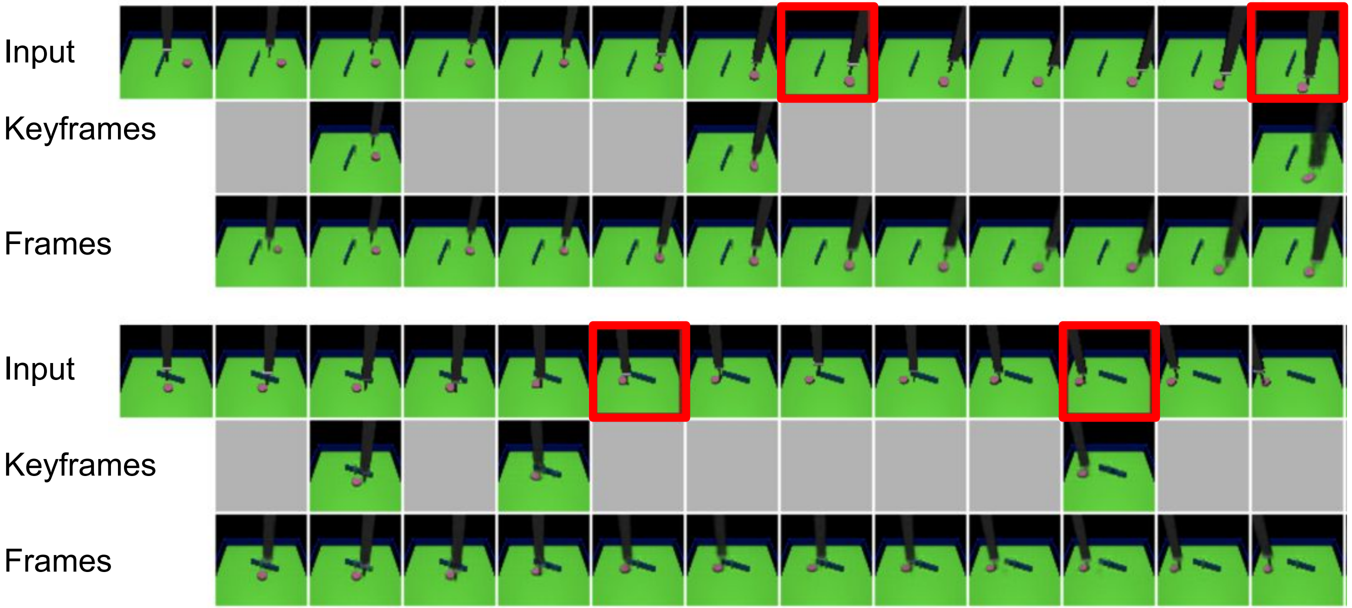


Generative model of trajectories via keyframes

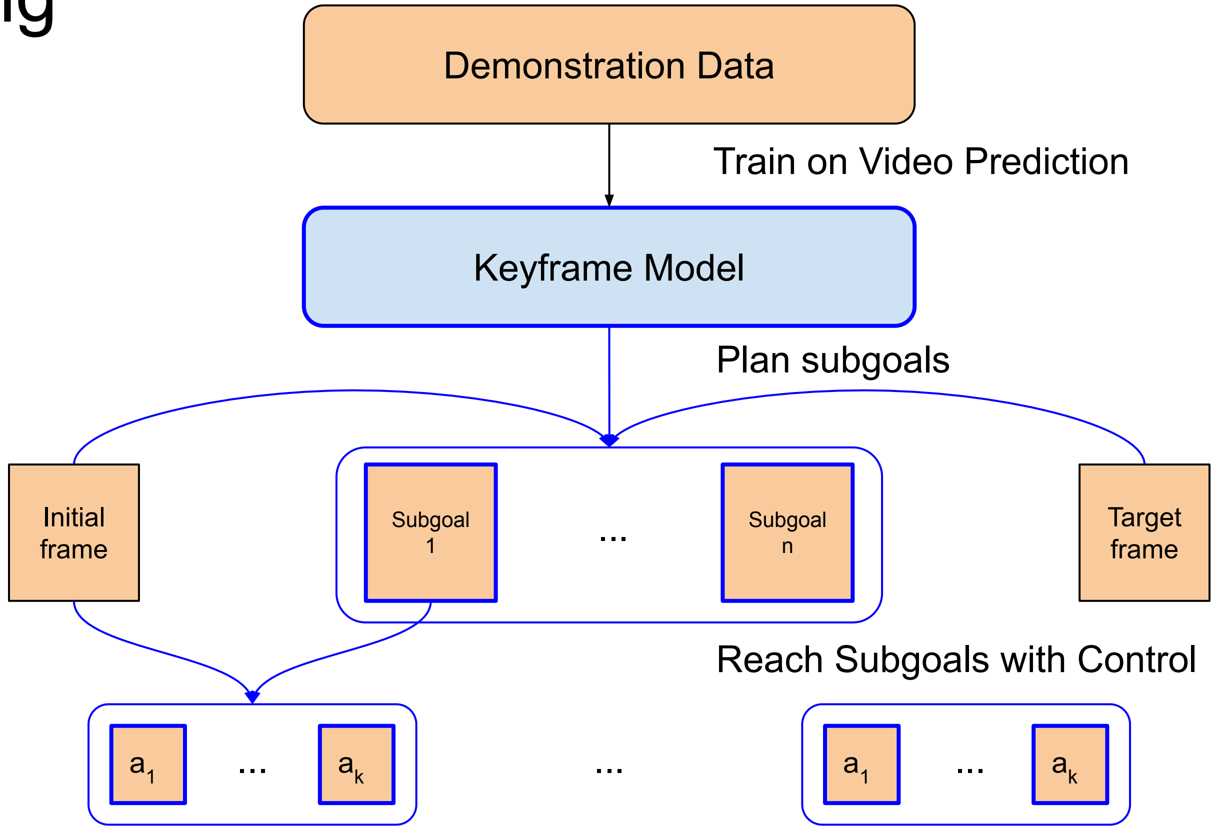


Pushing data





Planning



Ebert et Finn et al., 2018

Planning

Algorithm 1 Planning in the subgoal space.

Input: Keyframe model $\textsc{KeyIn}(.,.)$, cost function c
Input: Start and target images I_0 and I_{target}

Sample L sequences of latent variables:
$$z^{0:M} \sim \mathcal{N}(\mu_n, \sigma_n)$$
Produce subgoal plans: $\hat{K}^{0:M} = \textsc{KeyIn}(I_0, z^{0:M})$
Compute cost between produced and true target:
$$c(\hat{K}^M)$$
Choose L' best plans,

end for
Return: Best subgoal plan $K^{0:M}$

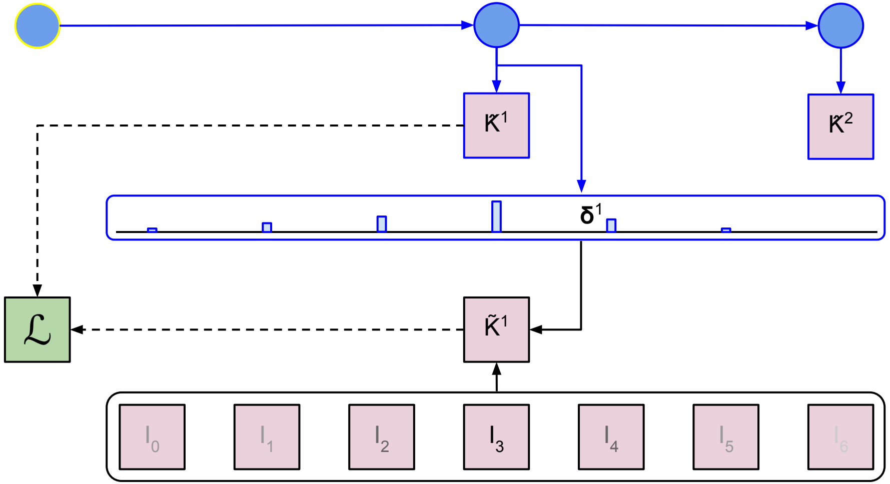
Ebert et Finn et al., 2018

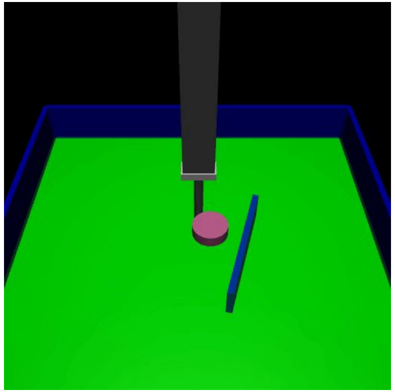
Planning on the pushing task

Method	Final position error	Success Rate
Intitial	1.32 ± 0.06	-
Random	1.32 ± 0.07	-
No subgoals	0.90 ± 0.14	$15.0\,\%$
TAP	0.80 ± 0.16	$23.3\,\%$
Jumpy	0.62 ± 0.33	$58.8\,\%$
KeyIn (Ours)	$\mathbf{0.50 \pm 0.26}$	**64.2 %**

KeyIn: Discovering Subgoal Structure with Keyframe-based Video Prediction

- The model learns to predict videos by first predicting a set of descriptive keyframes
- A differentiable loss allows to train the model to select the most descriptive keyframes
- The keyframes the model discovers are useful as subgoals for a planning task





Karl Pertsch*

Me*

Jingyun Yang

Kosta Derpanis

Joseph Lim

Kostas Daniilidis

Andrew Jaegle