KISA: A Unified Keyframe Identifier and Skill Annotator for Long-Horizon Robotics Demonstrations

Longxin Kou*[1], Fei Ni*[1], Yan Zheng[1], Jinyi Liu[1], Yifu Yuan[1], Zibin Dong[1], Jianye Hao[1]

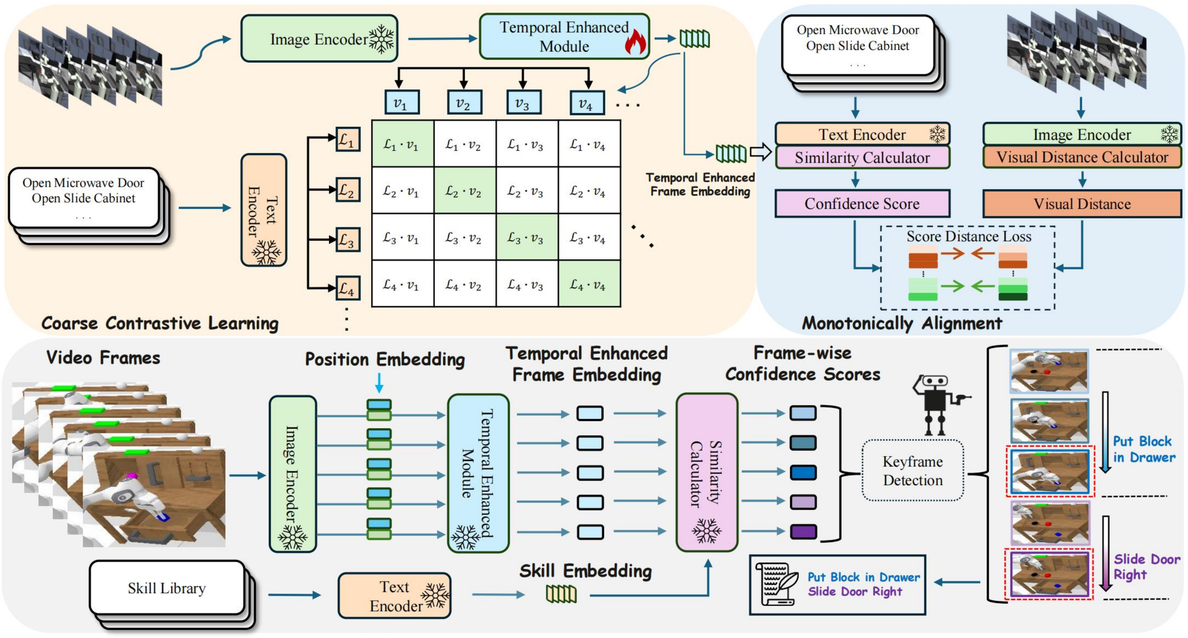
[1]Tianjin University

ICML International Conference On Machine Learning

Motivation

- **Complex robotics manipulation** tasks such as desktop tidying often span over long horizons and encapsulate multiple sub-tasks separated by keyframes. Directly learning from long-horizon demonstrations in an end-to-end manner is challenging**.**
- **Hierarchical policy** learning, by decomposing a complex demonstration into several shorter subtasks to facilitate the reusable skills and further enable modular skill composition for generalization. However, obtaining demonstrations with explicit keyframe boundaries and skill annotations is difficult, especially for real-world human videos.
- For this, we study the following open question - can we develop a framework that enables **automatic, scalable**, and **semantically meaningful** keyframe identification and skill annotation from unlabeled demonstrations?

Our Approach



- **Temporal enhancement module**
  - Relying solely on frame-level visual representations would induce training confusion for aligning them to distinct skills. So we propose a simple yet effective temporal-enhanced module on top of pretrained visual representations.

$$v_i = \Phi(h_i, o_i) = \Phi_{\text{TEMP}}(\{\phi(o_0), \ldots, \phi(o_{i-1}), \phi(o_i)\})$$

- **History-aware Contrastive Training**
  - we design coarse history-aware contrastive learning via constructing hard negative samples with mismatched historical contexts and incorrect skills.
- **Fine-grained Monotonic Alignment**
  - we additionally fine-grained monotonic alignment to encourage the capture of skill-aware progress within the sub-task, and prevent representation collapse to highly similarity within the same skill.

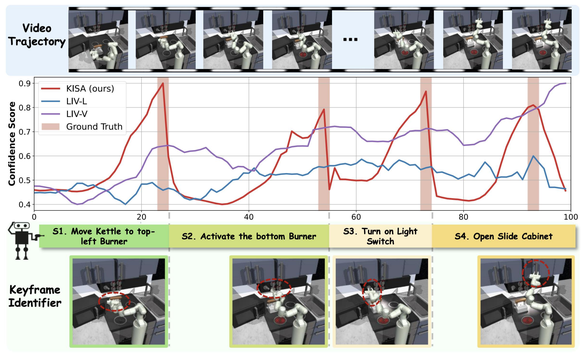
$$\mathcal{L}_{\text{contrastive}} = \underbrace{-\log \frac{e^{\mathcal{C}(o^+,h^+,\ell^+)}}{\sum_{j=1}^{k} e^{\mathcal{C}(o^+,h^+,\ell_j^-)}}}_{\text{Incorrect Skill Misalignments}} \underbrace{-\log \frac{e^{\mathcal{C}(o^+,h^+,\ell^+)}}{\sum_{z=1}^{k} e^{\mathcal{C}(o^+,h_z^-,\ell^+)}}}_{\text{Disjoint Frame-History Compositions}} \underbrace{-\log \frac{e^{\mathcal{C}(o^+,h^+,\ell^+)}}{\sum_{w=1}^{k} e^{\mathcal{C}(\{(o^+,h^+)_w^{rev},l^+\})}}}_{\text{Semantic Reversals via Video Inversion}}$$

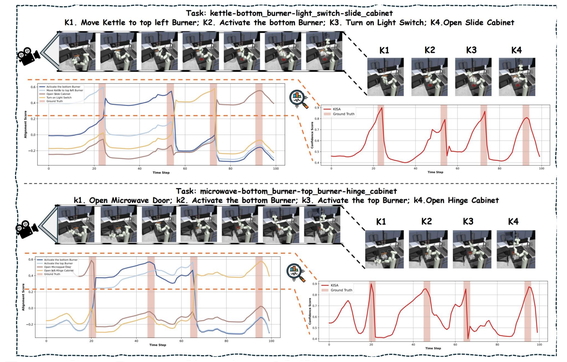
Experiments

- We conduct experiments on various benchmarks to answer the following questions: 1) Can KISA achieve better **accuracy** and **interpretable** skill alignment compared to other competitive baselines? 2) Does KISA exhibit robust zero-shot generalization across objects, compositional tasks, and cross-embodiments? 3) Is KISA a flexible framework for incorporating pretrained robotics representations?

➢ **The Accuracy of Keyframes and Skills Annotation**

Model	Maniskill Number Error↓	Maniskill F1 Score↑	Maniskill MAE↓	CALVIN Number Error↓	CALVIN F1 Score↑	CALVIN MAE↓	FrankaKitchen Number Error↓	FrankaKitchen F1 Score↑	FrankaKitchen MAE↓
VideoRLCS	10.1 ± 2.1	$15.2 \pm 0.3\%$	34.4 ± 0.4	9.5 ± 2.4	$15.4 \pm 0.5\%$	54.8 ± 1.0	0.6 ± 0.0	$5.5 \pm 0.8\%$	39.3 ± 0.7
KTS	5.1 ± 0.0	$15.7 \pm 4.0\%$	24.2 ± 6.8	0.9 ± 0.6	$20.2 \pm 0.7\%$	50.9 ± 7.4	0.5 ± 0.2	$13.8 \pm 3.4\%$	35.6 ± 6.9
R3M	5.0 ± 0.2	$17.1 \pm 0.8\%$	38.2 ± 3.0	5.4 ± 0.2	$21.1 \pm 1.3\%$	63.2 ± 1.1	1.0 ± 0.1	$53.7 \pm 0.8\%$	30.4 ± 0.1
VIP	4.0 ± 0.2	$31.7 \pm 2.5\%$	24.8 ± 1.4	4.6 ± 0.2	$24.3 \pm 1.6\%$	63.4 ± 1.2	0.9 ± 0.1	$57.2 \pm 1.1\%$	31.4 ± 0.1
LIV	3.9 ± 0.4	$30.3 \pm 2.2\%$	23.9 ± 1.4	5.6 ± 0.1	$25.9 \pm 1.1\%$	61.7 ± 1.5	1.4 ± 0.0	$64.2 \pm 0.8\%$	30.7 ± 0.1
UVD	0.7 ± 0.1	$40.2 \pm 1.1\%$	20.3 ± 0.1	0.8 ± 0.1	$36.9 \pm 0.5\%$	40.6 ± 1.7	0.6 ± 0.1	$64.8 \pm 2.4\%$	31.1 ± 0.2
KISA	$\mathbf{0.0 \pm 0.0}$	$\mathbf{99.7 \pm 0.2\%}$	$\mathbf{0.2 \pm 0.1}$	$\mathbf{0.1 \pm 0.1}$	$\mathbf{85.2 \pm 0.9\%}$	$\mathbf{11.2 \pm 2.4}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{98.7 \pm 0.6\%}$	$\mathbf{0.4 \pm 0.0}$





➢ **The Flexibility for Pre-trained Representations**

Methodology	Maniskill2	CALVIN	FrankaKitchen
KISA-R3M	$\mathbf{71.8 \pm 3.9\%}$	$\mathbf{53.6 \pm 1.1\%}$	$\mathbf{88.9 \pm 0.6\%}$
- w/o monotonic align	$63.0 \pm 3.2\% \downarrow$	$50.1 \pm 0.8\% \downarrow$	$81.5 \pm 1.0\% \downarrow$
- w/o historical contrastive	$41.3 \pm 2.2\% \downarrow$	$45.7 \pm 2.3\% \downarrow$	$76.6 \pm 0.5\% \downarrow$
- w/o temporal enhance	$23.1 \pm 0.5\% \downarrow$	$24.0 \pm 2.2\% \downarrow$	$21.9 \pm 0.8\% \downarrow$
KISA-VIP	$\mathbf{99.6 \pm 0.1\%}$	$\mathbf{70.9 \pm 2.7\%}$	$\mathbf{96.4 \pm 0.3\%}$
- w/o monotonic align	$88.9 \pm 0.5\% \downarrow$	$64.0 \pm 0.4\% \downarrow$	$90.1 \pm 0.3\% \downarrow$
- w/o historical contrastive	$58.9 \pm 1.5\% \downarrow$	$53.1 \pm 1.5\% \downarrow$	$81.8 \pm 0.9\% \downarrow$
- w/o temporal enhance	$24.0 \pm 0.3\% \downarrow$	$23.4 \pm 0.8\% \downarrow$	$21.4 \pm 0.8\% \downarrow$
KISA-LIV	$\mathbf{99.2 \pm 0.1\%}$	$\mathbf{94.7 \pm 1.1\%}$	$\mathbf{96.1 \pm 0.2\%}$
- w/o monotonic align	$90.2 \pm 0.3\% \downarrow$	$82.1 \pm 0.4\% \downarrow$	$89.1 \pm 0.7\% \downarrow$
- w/o historical contrastive	$59.1 \pm 1.9\% \downarrow$	$58.1 \pm 1.3\% \downarrow$	$73.7 \pm 0.4\% \downarrow$
- w/o temporal enhance	$22.0 \pm 1.6\% \downarrow$	$25.0 \pm 1.6\% \downarrow$	$19.0 \pm 0.3\% \downarrow$

➢ **The Zero-shot Generalization Ability.**

Model	Maniskill2(L1) F1 Score↑	Maniskill2(L1) MAE↓	CALVIN(L2) F1 Score↑	CALVIN(L2) MAE↓	RealKitchen(L3) F1 Score↑	RealKitchen(L3) MAE↓
KTS	$12.3 \pm 2.7\%$	25.2 ± 5.3	$20.2 \pm 0.7\%$	54.9 ± 7.4	$11.9 \pm 5.4\%$	44.9 ± 21.4
R3M	$16.8 \pm 1.4\%$	38.8 ± 2.5	$20.9 \pm 1.1\%$	63.4 ± 1.2	$20.7 \pm 0.5\%$	44.9 ± 21.4
VIP	$30.6 \pm 1.8\%$	23.3 ± 1.8	$23.6 \pm 1.9\%$	63.2 ± 1.1	$20.5 \pm 17.8\%$	34.8 ± 11.0
LIV	$30.2 \pm 1.7\%$	23.6 ± 1.0	$25.4 \pm 1.3\%$	63.2 ± 1.1	$21.7 \pm 20.9\%$	44.9 ± 21.4
UVD	$39.2 \pm 1.1\%$	21.5 ± 0.2	$36.9 \pm 0.5\%$	40.6 ± 1.7	$26.3 \pm 11.9\%$	30.2 ± 6.2
KISA	$\mathbf{80.7 \pm 0.9\%}$	$\mathbf{6.4 \pm 0.7}$	$\mathbf{89.4 \pm 1.8\%}$	$\mathbf{14.2 \pm 0.9}$	$\mathbf{40.7 \pm 14.8\%}$	$\mathbf{27.8 \pm 5.0}$

➢ **Visualization of Keyframe Identification**