Generative Hierarchical Temporal Transformer for Hand Pose and Action Modeling

Yilin Wen[1,2], Hao Pan[3], Takehiko Ohkawa[2], Lei Yang[1,4], Jia Pan[1,4], Yoichi Sato[2], Taku Komura[1] , Wenping Wang[5]

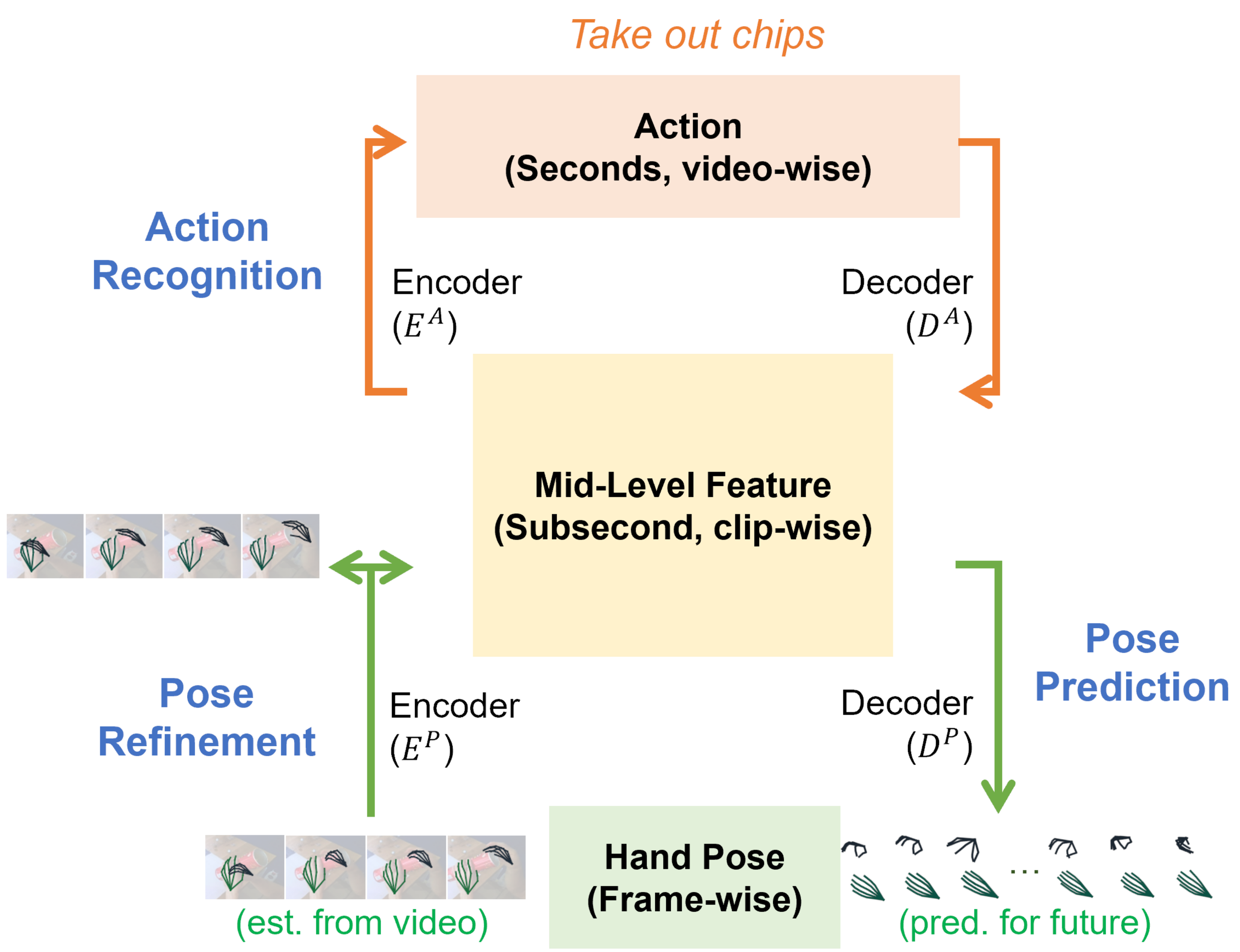
[1]The University of Hong Kong, [2]The University of Tokyo, [3]Microsoft Research Asia, [4]Centre for Garment Production Limited, Hong Kong, [5]Texas A&M University

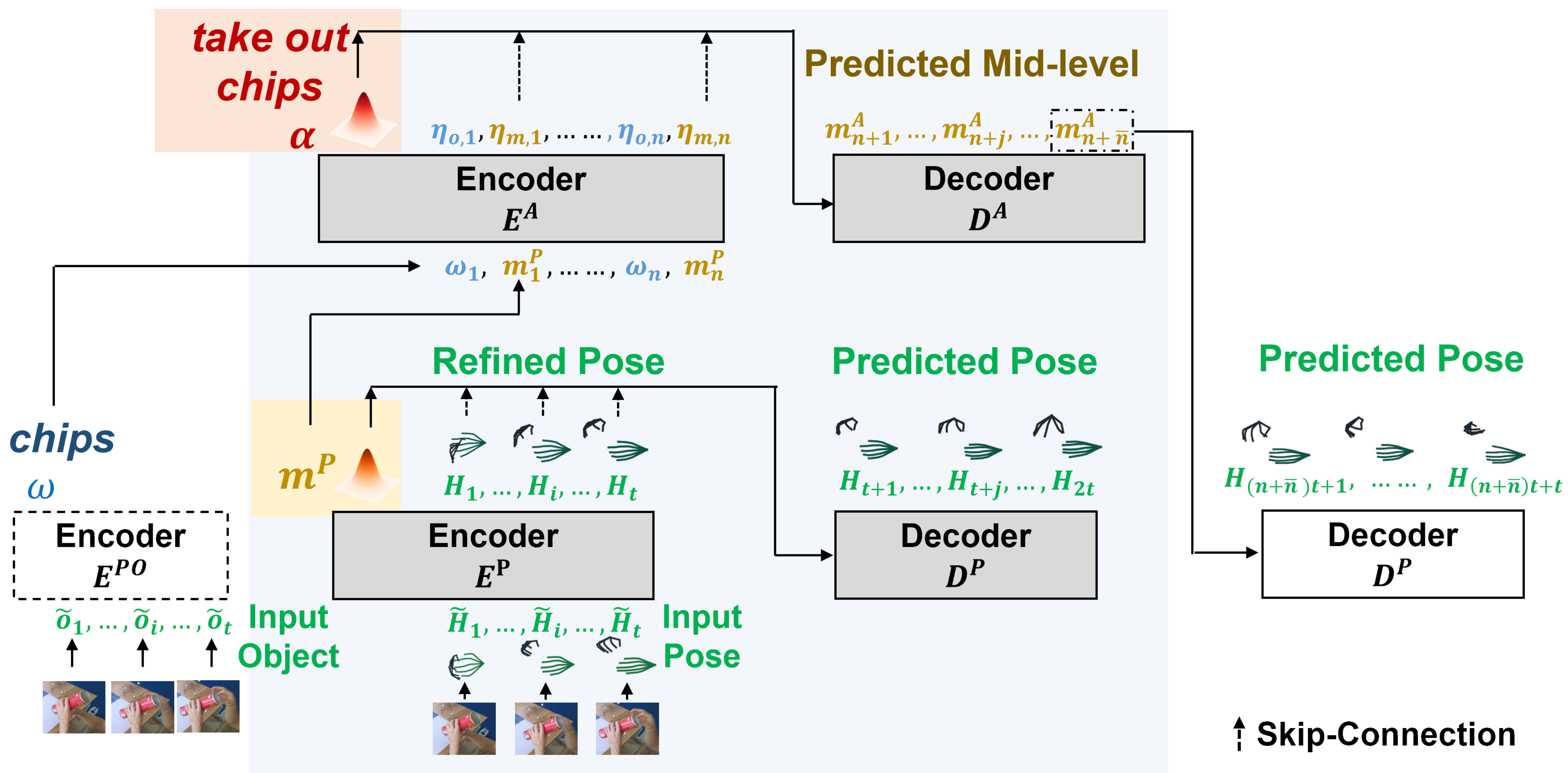
Project Page

Summary: A Unified Framework

- **Concurrently tackles recognition and generation.**
  - ✓ Exploits the synergy of both sides, thus improving over separate models.
- **Models semantic dependency and temporal granularity between pose and action.**
  - ✓ Captures both short-term and long-term temporal regularities via hierarchical temporal transformer blocks.
  - ✓ Trains the two blocks separately to fully utilize datasets with annotations of different temporal granularities.

G-HTT: Hierarchical Transformer VAE



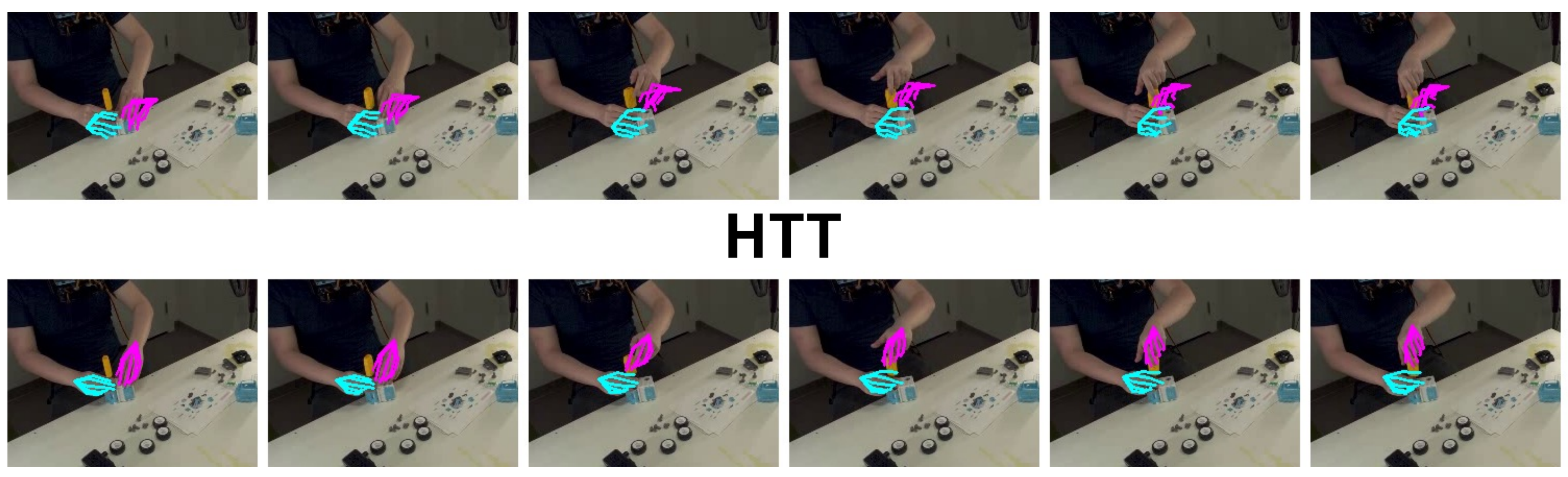


- **Generative Transformer VAE architecture** to jointly model recognition and prediction.
  - Encoder and decoder respectively capture recognition and prediction.
  - VAE bottleneck mandates the learning of consistent hand motion from the past to the future and vice versa.
- **Block cascades** to capture the semantic dependency and temporal granularity of hand pose-action.
  - Lower block and upper block respectively model hand poses over short time spans and action over long time spans.
  - Two blocks are bridged by a **middle-level representation.**

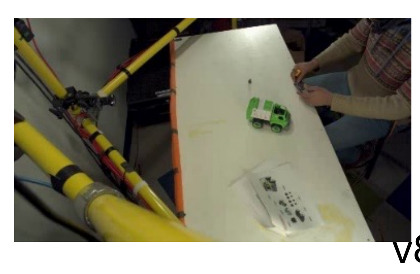
Results: G-HTT Improves over Separate Models

Hand Pose Estimation and Action Recognition

- Baselines:
  - **Resnet-18 for image-based hand pose estimation**, which provides frame-wise inputs for G-HTT and HTT.
  - **HTT [Wen+, CVPR'23] for hand pose estimation and action recognition.** Note that HTT is trained on the pre-trained Resnet-18, where camera view v3 is leveraged in training; our G-HTT is never trained on the pre-trained Resnet-18.
- Through evaluation on different camera viewpoints, our G-HTT shows enhanced generalization by learning regular motion priors across tasks; HTT is more likely to overfit particular data distributions.

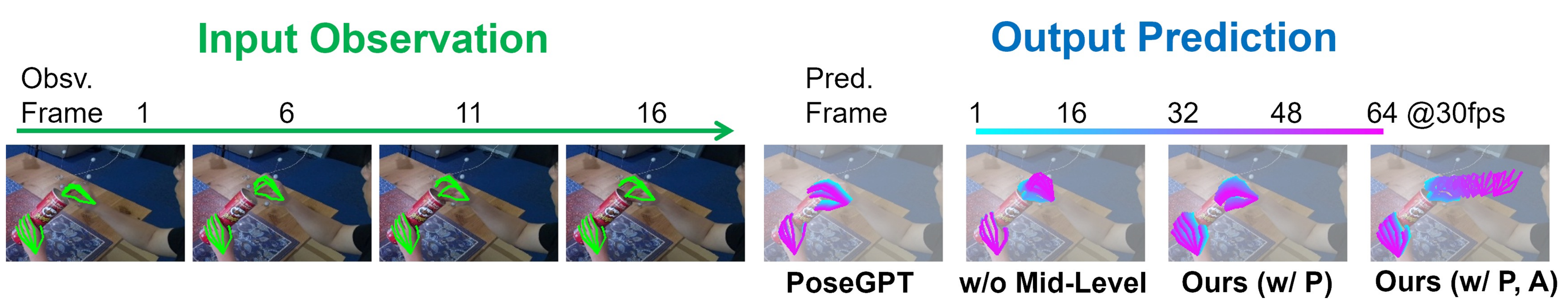


On camera view v1 of AssemblyHands [Ohkawa+, CVPR'23].

	 v1		 v3*		 v8	
	MPJPE-RA↓ <L, R> in mm	Action Accuracy↑	MPJPE-RA↓ <L, R> in mm	Action Accuracy↑	MPJPE-RA↓ <L, R> in mm	Action Accuracy↑
Resnet-18	35.4, 22.7	-	27.5, 27.2	-	26.1, 30.4	-
HTT	55.6, 39.0	16.55	**26.7**, 27.3	**39.42**	91.3, 88.5	9.98
Ours	**35.1, 22.4**	**36.01**	27.3, **26.9**	34.79	**25.9, 30.0**	**36.74**

Hand Motion Prediction

- Ours shows better generation quality across actions than a prediction-only network (i.e., PoseGPT [Lucas+, ECCV'22])



On a case of *taking out chips* from the H2O dataset. Ours shows globally consistent action.

	H2O-test		AssemblyHands-val	
	FID ↓	APD ↑ <L, R> in mm	FID ↓	APD ↑ <L, R> in mm
PoseGPT	11.70	**24.1, 48.6**	16.07	25.3, **33.0**
Ours (with P, A)	**8.19**	20.1, 33.9	**5.04**	**28.1**, 32.8

On action sequences that are longer than 1 sec.