Focusing on what to decode and what to train: SOV Decoding with Specific Target Guided DeNoising and Vision Language Advisor

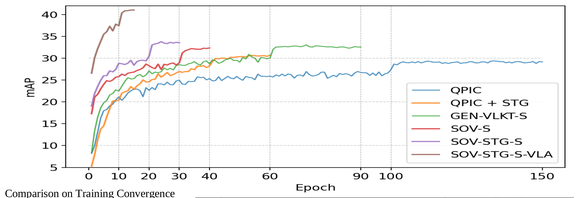
Junwen Chen Yingcheng Wang Keiji Yanai

Department of Informatics, The University of Electro-Communications, Tokyo, Japan

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Human-Object Interaction Detection

- **HOID task**: a downstream task of object detection
- Detect human-object pairs and recognize interactions
- **Transformer-based one-stage HOID method**: Predicting HOI triplets with a set of queries



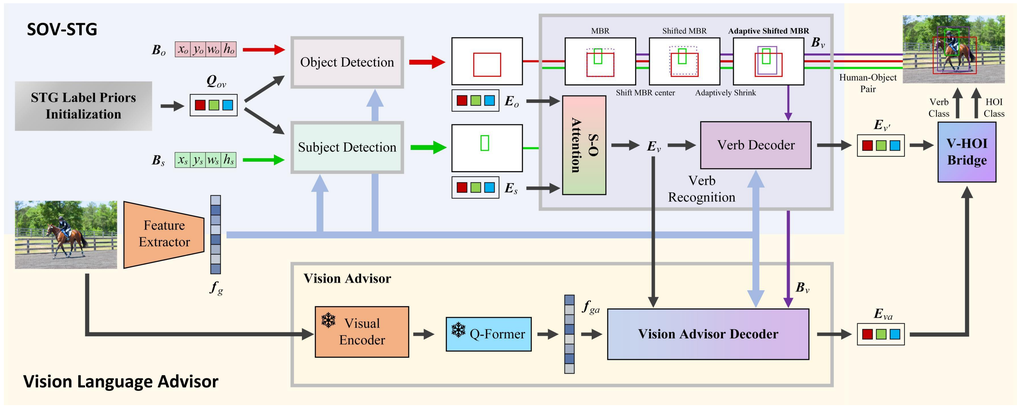
Comparison on Training Convergence

#	oDec	sDec	vDec	S-O Attn	STG	VLA	Default Full	Default Rare	Default Non-Rare
(1)	✓		✓		✓		32.68	28.21	34.02
(2)	✓	✓			✓		32.35	27.64	33.63
(3)	✓						30.14	22.82	32.32
(4)	✓	✓					30.62	24.60	32.42
(5)	✓	✓	✓				31.90	25.92	33.69
(6)	✓	✓	✓		✓		33.01	27.83	34.55
(7)	✓	✓	✓	✓	✓		33.80	29.28	35.15
(8)	✓	✓		✓	✓	✓	40.58	38.65	41.16
(9)	✓	✓	✓	✓	✓	✓	**41.16**	**39.48**	**41.67**

Contributions of each module

oDec: Object Decoder
sDec: Subject Decoder
vDec: Verb Decoder
S-O Attn: S-O Attention Module

SOV-STG-VLA Inference Pipeline



Results

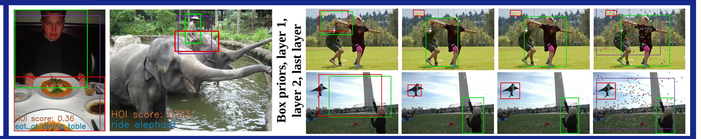


HICO-DET

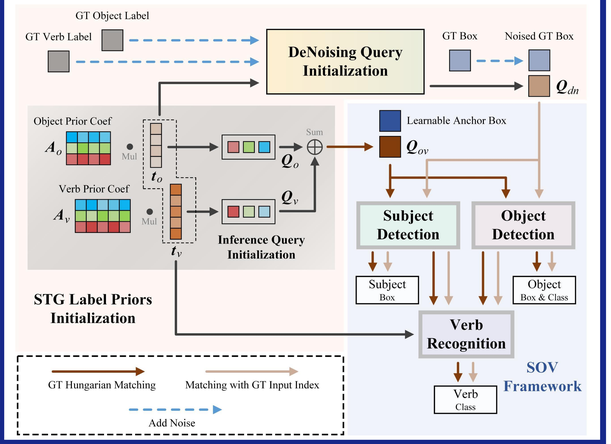
Method	Epoch	Backbone	Default Full	Default Rare	Default Non-Rare	Known Object Full	Known Object Rare	Known Object Non-Rare
Two-stage								
CASN [13]	12	ResNet-50	31.86	25.15	33.84	34.44	27.69	36.45
STIP [77]	20	ResNet-50	32.22	28.15	33.43	35.29	31.43	36.45
UPT [71]	20	ResNet-101-DC5	32.62	28.62	33.81	36.08	31.41	37.43
CDN-L [74]	10	ResNet-101-DC5	32.52	27.39	33.95	35.91	29.27	37.38
VPLG [11]	8	ViT-B/32	34.95	33.83	35.28	36.15	36.77	36.56
One-stage								
QAHOI [1]	150	ResNet-50	29.18	18.06	28.85	-	-	-
QPIC [57]	150	ResNet-50	29.07	21.85	31.23	31.68	24.14	33.93
MSTR [19]	50	ResNet-50	31.17	25.31	32.92	34.02	28.83	35.57
CDN-L [74]	100	ResNet-101	32.07	27.19	33.53	34.79	29.48	36.38
HQM (CDN-S) [30]	80	ResNet-50	32.47	28.15	33.56	35.17	30.73	36.50
RLIP-ParSe [63]	90	ResNet-50	32.84	24.63	26.85	-	-	-
RLIP-ParSe [63]	90	ResNet-50	33.63	26.32	34.37	36.78	32.73	37.99
GEN-VLKT-S [34]	90	ResNet-50	33.75	29.25	35.10	36.78	32.75	37.99
HOICLIP [17]	90	ResNet-50	34.69	31.12	35.74	37.61	34.47	38.54
GEN-VLKT-M [34]	90	ResNet-101	34.78	31.50	35.77	38.07	34.94	39.01
GEN-VLKT-L [34]	90	ResNet-101	34.95	31.18	36.08	38.22	34.36	39.37
CLIP4HOI [41]	100	ResNet-50	35.33	32.95	35.74	37.19	35.27	37.77
CQL+GEN-VLKT-S [40]	90	ResNet-50	35.36	32.97	36.07	38.43	34.85	39.50
UniHOI-S [1]	90	ResNet-50	40.06	**39.91**	40.11	42.20	42.01	42.08
QAHOI-Swin-L [1]	150	Swin-Large-22K	35.78	29.80	37.56	37.59	31.36	39.36
HQM (CDN-L) [30]	160	Swin-Large-22K	37.11	33.79	37.77	38.97	35.80	39.57
DiffHOI-Swin-L [15]	90	Swin-Large-22K	41.50	39.16	41.94	43.82	41.73	44.28
PViC-Swin-L [44]	30	Swin-L	44.32	**44.61**	44.24	47.81	**48.38**	47.64
RLIP-ViT-PaSt-Swin-L [63]	20	Swin-Large	45.09	45.23	45.64	-	-	-
SOV-STG-S	30	ResNet-50	33.80	29.28	35.15	36.22	30.99	37.78
SOV-STG-M	30	ResNet-101	34.87	30.39	36.20	37.35	32.46	38.81
SOV-STG-L	30	ResNet-101	35.01	30.63	36.32	37.66	32.77	39.05
SOV-STG-VLA-S	15	ResNet-50	**41.16**	39.48	**41.67**	**43.83**	**42.63**	**44.17**
SOV-STG-Swin-L	30	Swin-Large-22K	43.35	42.25	43.69	45.53	43.82	46.11
SOV-STG-VLA-Swin-L	15	Swin-Large-22K	**45.64**	**46.35**	**46.03**	**48.22**	**47.12**	**48.55**

VCOCO

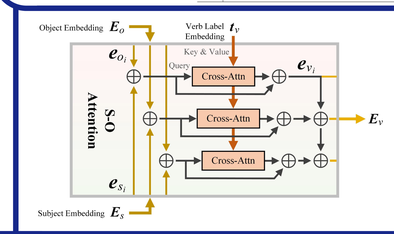
Method	Backbone	AP^{S2}_{role}	AP^{S2}_{role}
FGAHOI [33]	ResNet-50	59.0	59.3
RLIP-ParSe [63]	ResNet-50	61.9	64.2
MSTR [19]	ResNet-50	62.0	65.2
GEN-VLKT-S [34]	ResNet-50	62.4	64.5
GEN-VLKT-M [34]	ResNet-101	63.3	65.6
HORCLIP [17]	ResNet-50	63.5	64.8
GEN-VLKT-L [34]	ResNet-101	63.6	65.9
CDN-L [74]	ResNet-101	63.9	65.9
UniHOI-S [1]	ResNet-50	65.6	68.3
SOV-STG-S	ResNet-50	61.2	62.5
SOV-STG-M	ResNet-101	63.7	65.2
SOV-STG-L	ResNet-101	63.9	65.4
SOV-STG-VLA-S	ResNet-50	63.8	65.7



SOV-STG Training Pipeline



S-O Attention



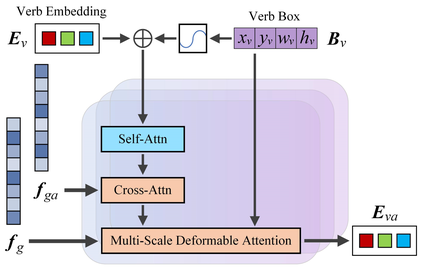
SOV-STG consists of STG label priors initialization, subject and object detection, and the verb recognition.
VLA enriches the verb embeddings by **Vision Advisor**, **V-HOI Bridge** connects the prediction of HOI and verb labels.

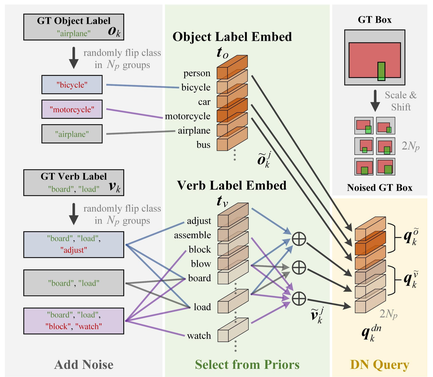
Challenge & Motivation

- **Decoding target**: Query embeddings used to predict both labels and boxes suffer from ambiguous representations.
- **Hard training**: Recent DETR-based methods need long session training to recognize long-tailed HOI.
- **Verb label or HOI label?** The knowledge gap between the prediction of HOI and verb labels.

Vision Advisor Decoder

Vision Advisor Decoder is designed to release the burden of aligning the pretrained VLM knowledge with the verb recognition part of SOV-STG.





Adding Noise to GT

V-HOI Bridge

Verb-HOI bridge predicts the HOI classes in a two-step manner to fill the gap between the verb recognition and HOI recognition.

