EgoDTM: Towards 3D-Aware Egocentric Video-Language Pretraining

Boshen Xu[1] **Yuting Mei**[1] **Xinbi Liu**[1] **Sipeng Zheng**[2] **Qin Jin**[1]*
[1] AIM3 Lab, Renmin University of China [2] BeingBeyond

AI·M³
www.ruc-aim3.com

中國人民大學
RENMIN UNIVERSITY OF CHINA

Egocentric Videos: How Humans See the World

VR/AR

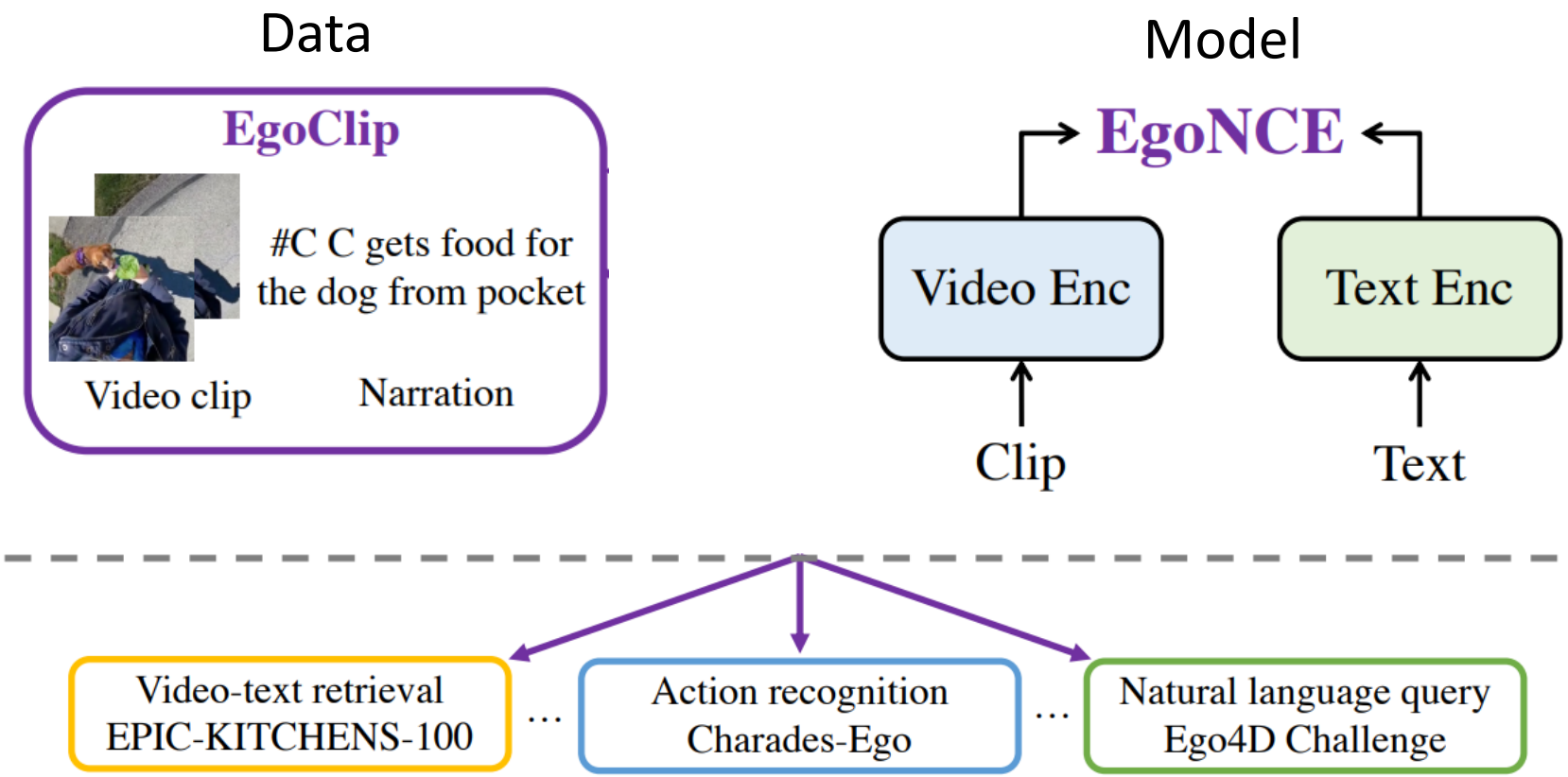
making salad

Egocentric view

Ego4Robotics

Transferable knowledge for manipulation

Literature of Ego-Visual-Text Pretraining

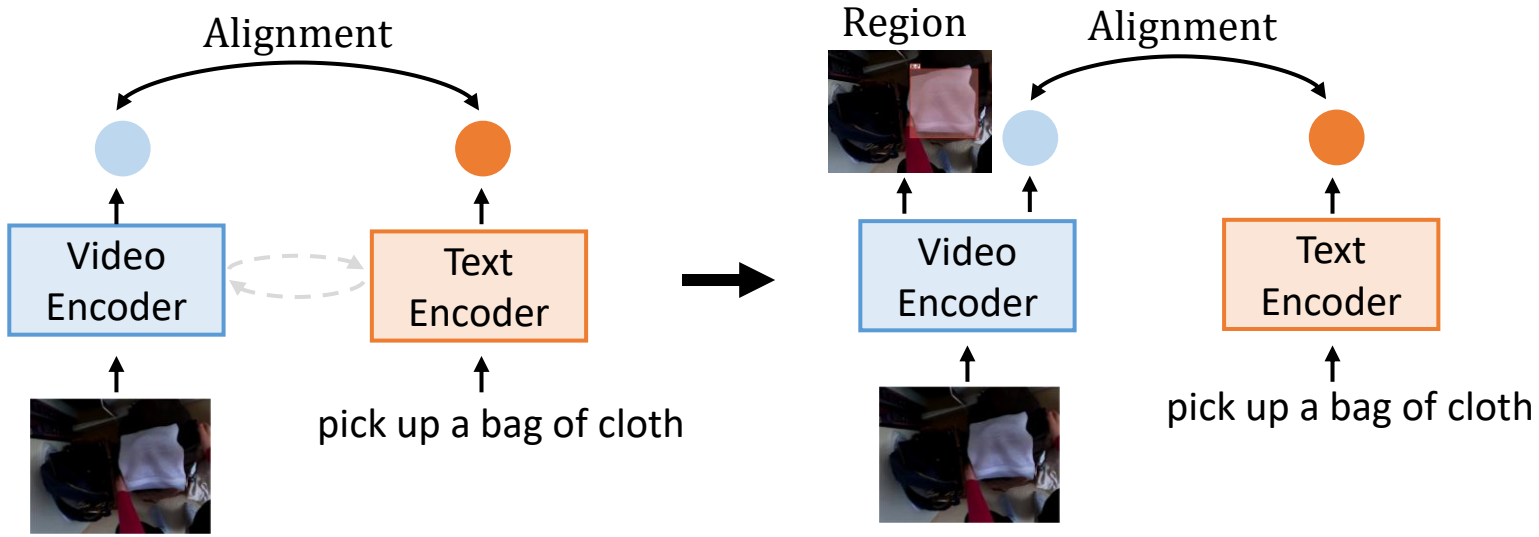


Great Success in Downstream Tasks

EgoVLP, Lin et al., [NeurIPS2022]

Literature of Ego-Visual-Text Pretraining

Enhance performance with region-aware representations

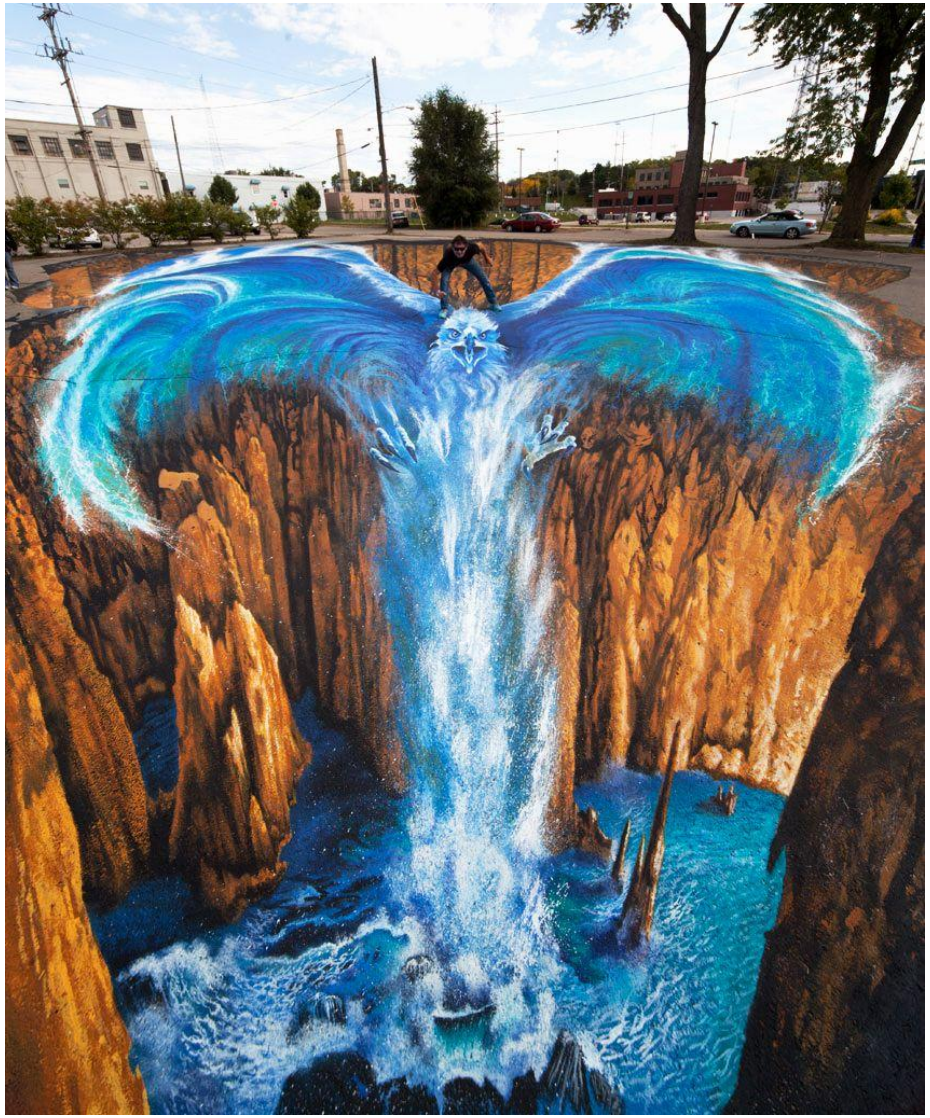


(a) Vanilla Video-Text Pretraining (*EgoVLP[NeurIPS22], EgoVLPv2[ICCV23], LaViLa[CVPR23]*)

(b) 2D Region-Aware Pretraining (*HelpingHands[ICCV23], HENASY[NeurIPS24]*)

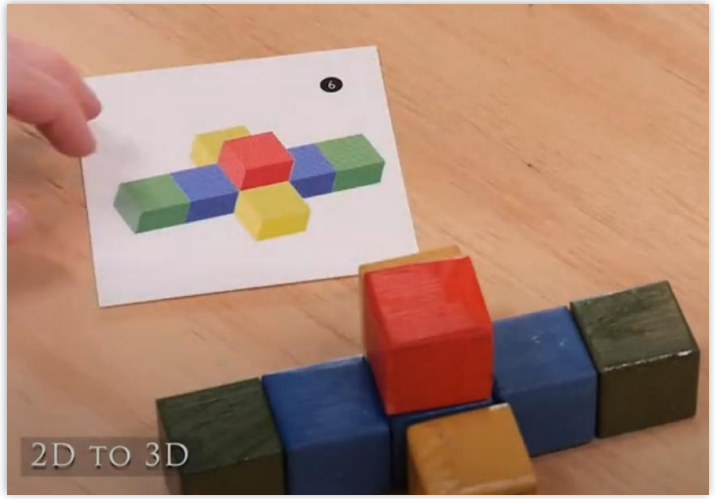
Wait! The world is not a 2D Flat!
Let’s consider how humans perceive the 3D world…

We Hallucinate 3D Information by Our Knowledge



Humans Live in a 3D World

Through embodied experiences, we humans excel in understanding 3D spatial relations, as well as interacting with the world with 2D vision.



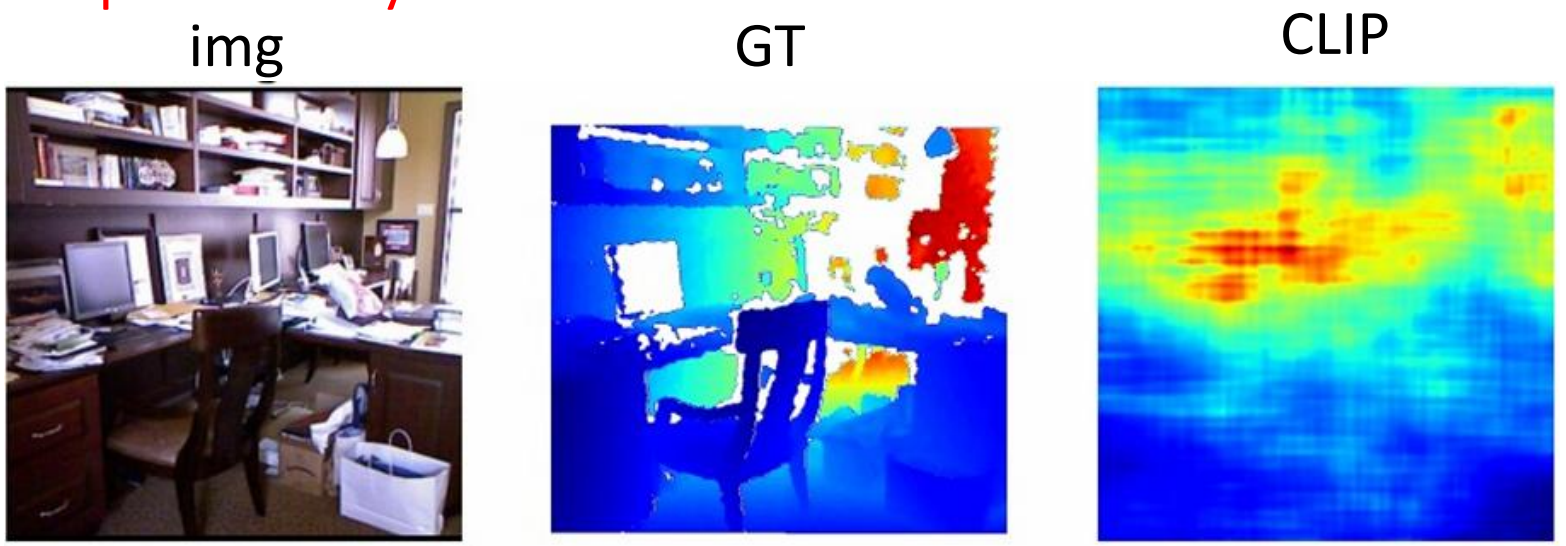
2D-to-3D Imagination

Spatial Sensing

Multi-view Consistency

But…VLPs Fall Short in 3D Perception

Current models are trained by text in advance of vision signals, strongly lacking in spatial perception ability.



Probe3D, Mohamed et al., CVPR2024

Direction estimation

Q: Pretend that you are standing facing the stove as shown in this image. At what direction (in degrees) is the storage chest relative to you?

Choices:

A) -49 B) 11
C) -10 D) 41

Observation space: Ego image

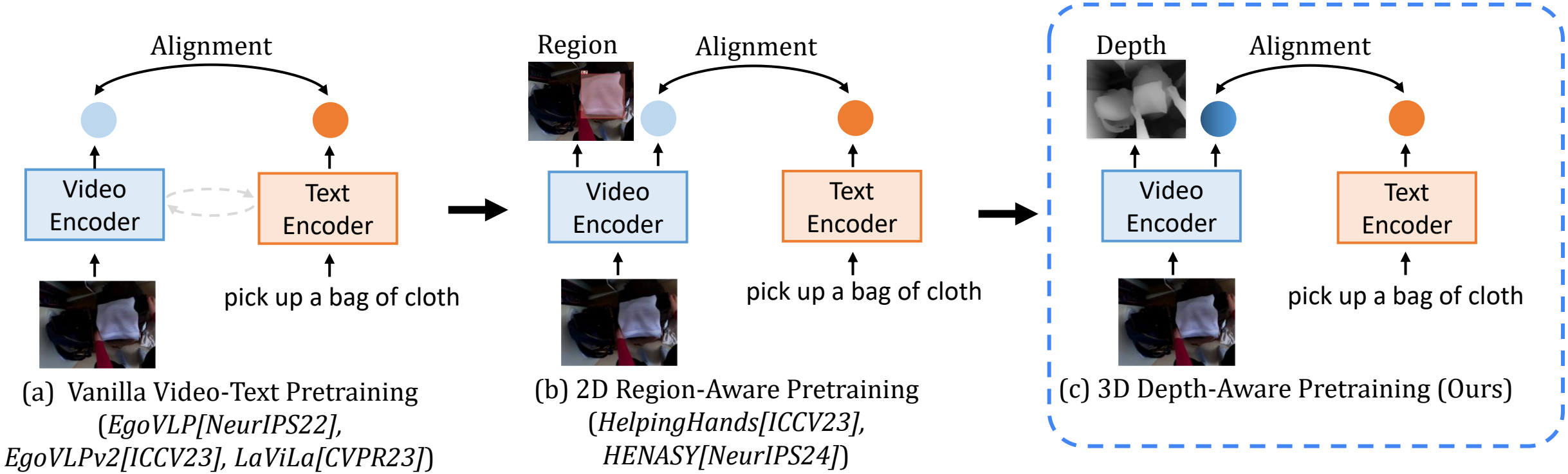
Method	Direction estimation	Distance estimation	Map sketching	Route retracing	Novel shortcuts	Average
Human	82.8	83.2	96.6	-	-	-
GPT-4o	32.0 ±4.1	36.5 ±5.0	33.3 ±4.1	6.6 ±3.6	6.4 ±1.0	23.0
GPT-4v	29.7 ±0.3	31.9 ±2.7	20.0 ±11.8	1.6 ±1.2	3.9 ±0.9	17.4
Chance	25.0	25.0	25.0	0.0	0.0	15.0

SPACE, Santhosh et al., ICLR2025

How to develop 3D-aware 2D VLP?

Towards 3D-Aware EgoVLP

Let's define 3D as Depth Maps, then
Pretrain video encoder with both depth and text



(a) Vanilla Video-Text Pretraining (*EgoVLP[NeurIPS22], EgoVLPv2[ICCV23], LaViLa[CVPR23]*)

(b) 2D Region-Aware Pretraining (*HelpingHands[ICCV23], HENASY[NeurIPS24]*)

(c) 3D Depth-Aware Pretraining (Ours)

Challenges for Developing 3D-aware EgoVLP

Architecture for Joint Learning from Heterogeneous Supervision

- Dense pixel-level knowledge

Depth (Geometry, Relation)

- Sparse non-pixel-level knowledge

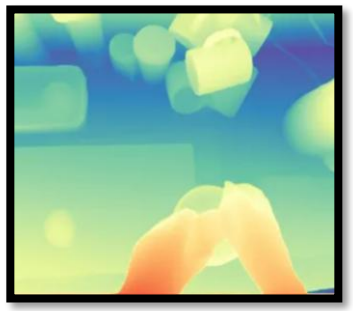
Text A person squeezes the lemon

→ Main Idea: Simplify the learning difficulty for depth estimation

Data Scarcity for Million-level (video, Depth, Text)

- Lack of paired (visual, depth) data

Massive



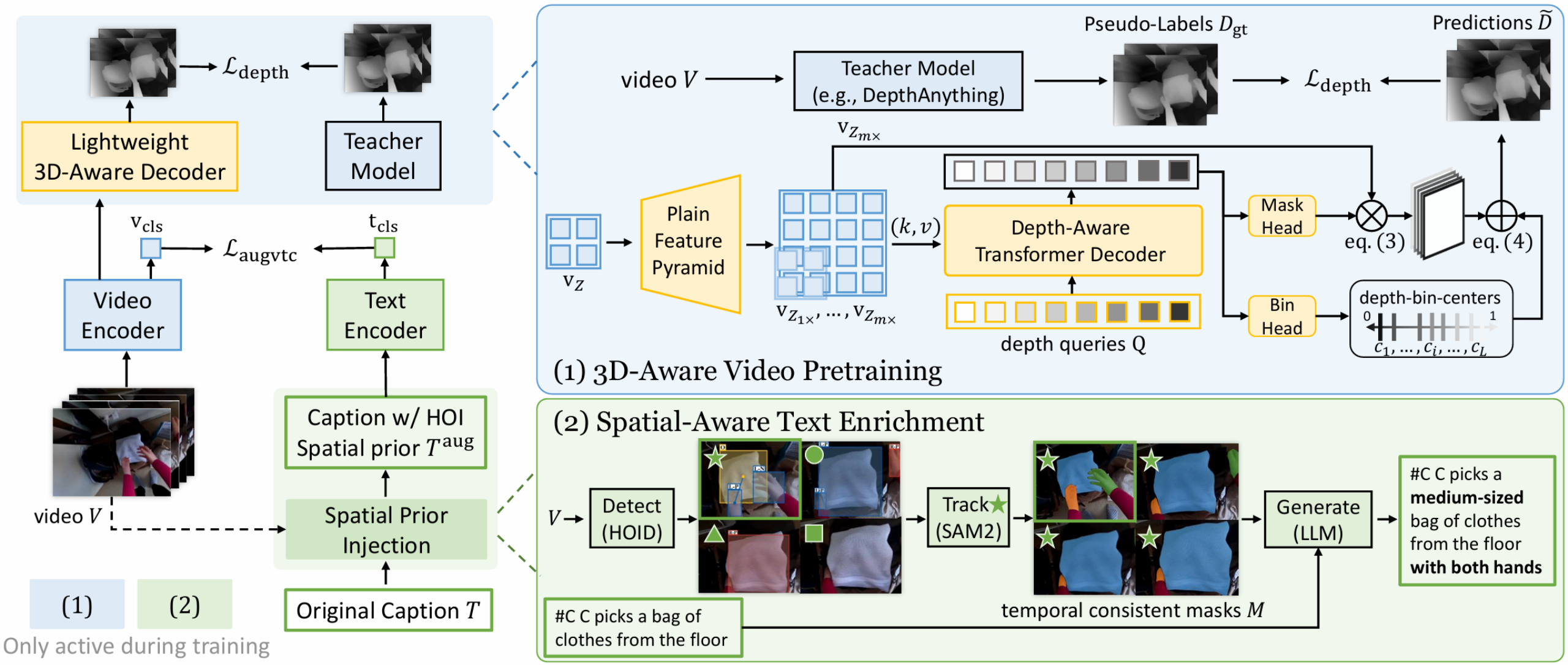
Few

- Short text descriptions, may not contribute to spatial awareness

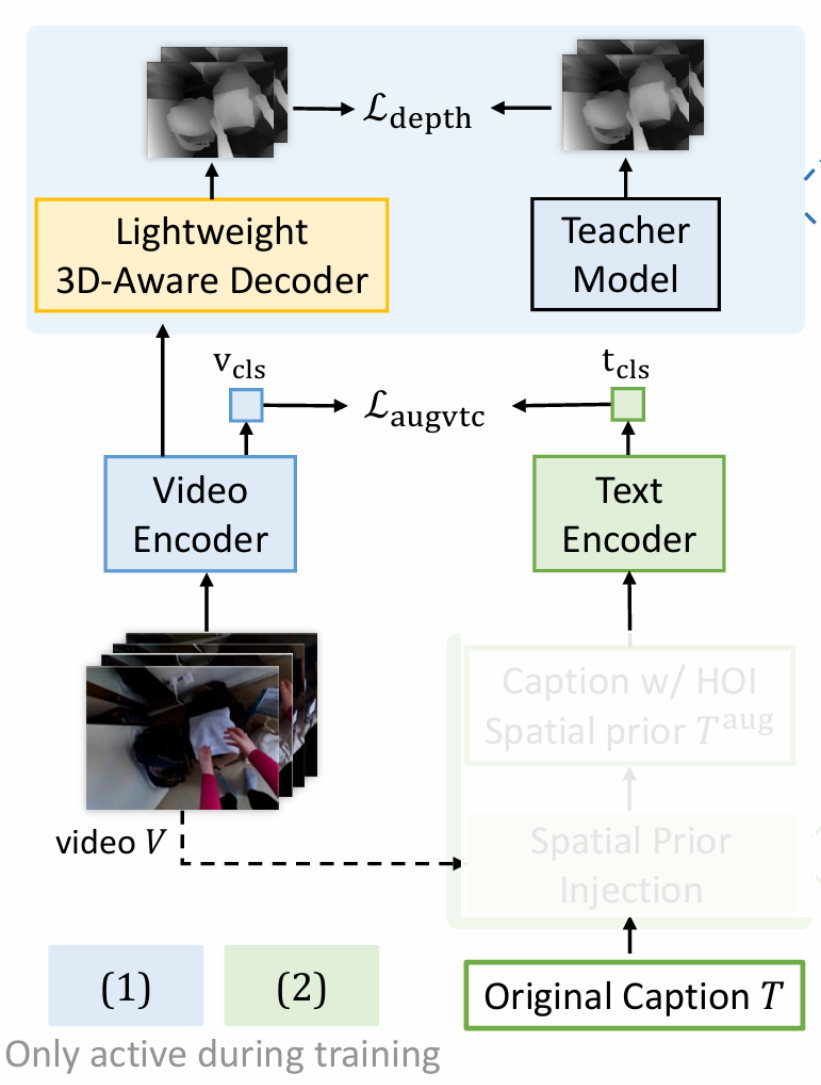
Raw Text A person squeezes the lemon

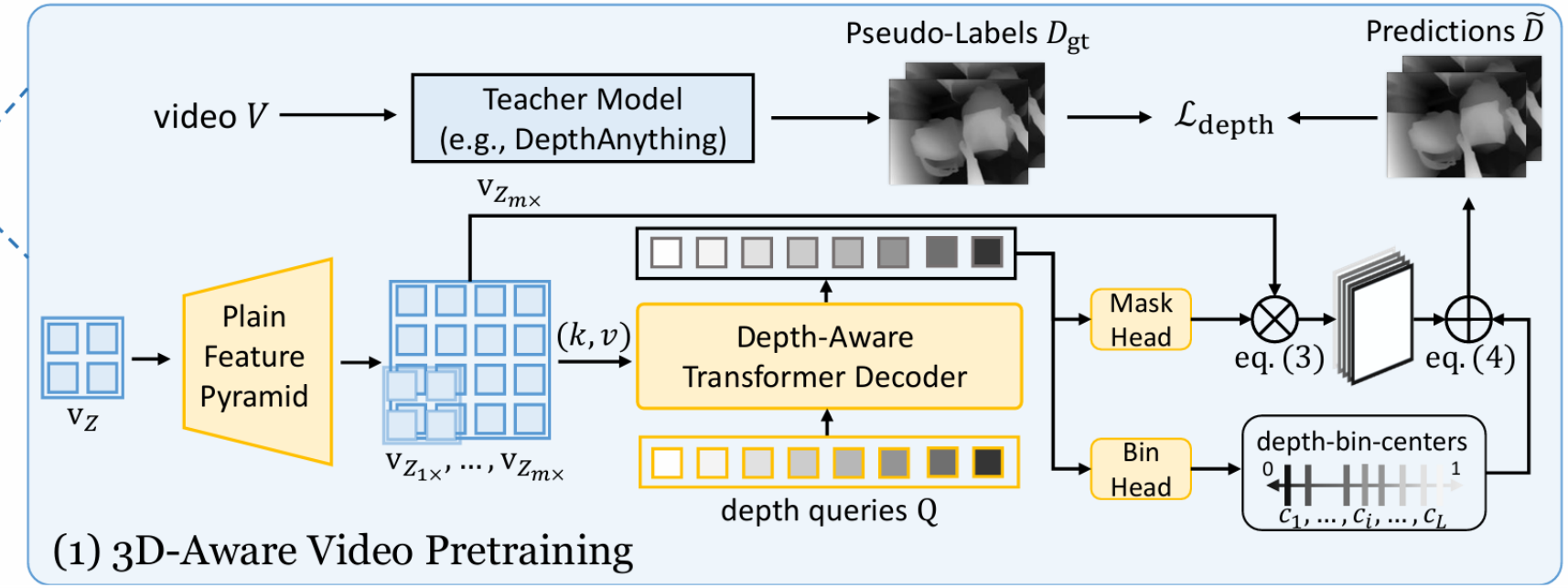
→ Main Idea: Data generation via visual foundation models

EgoDTM: Egocentric Depth- and Text-Aware Model



Lightweight 3D-Aware Decoder Design

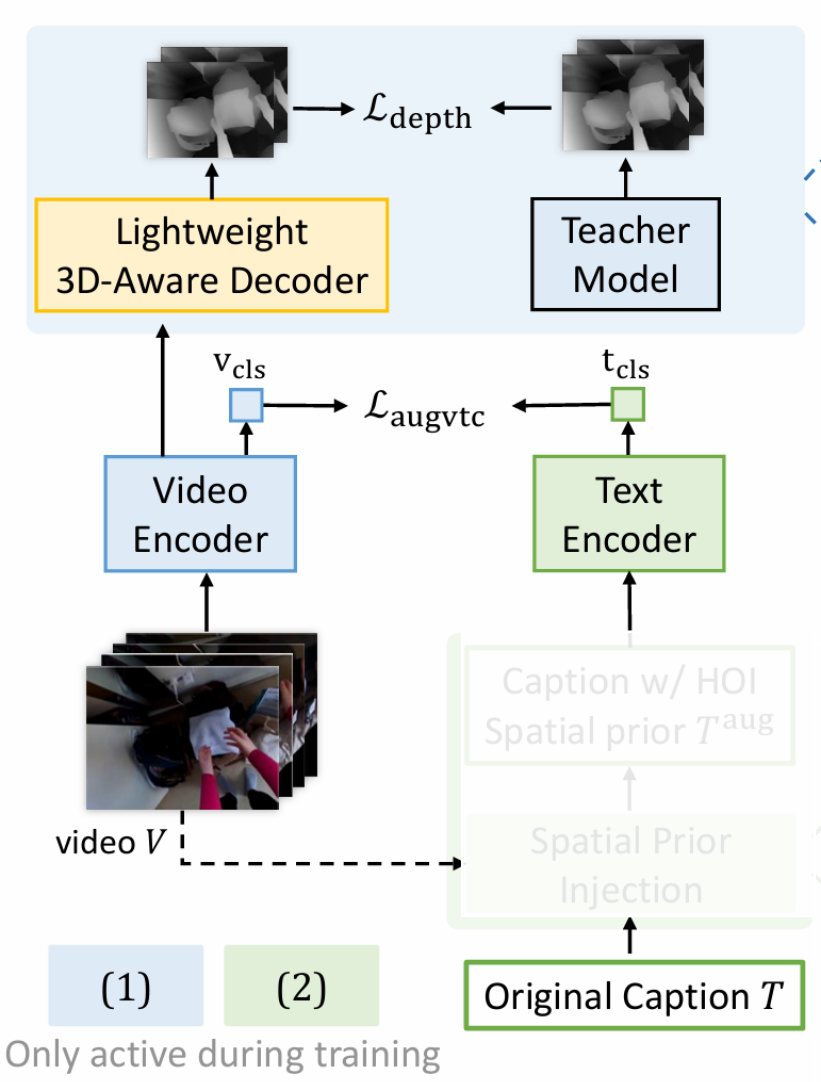


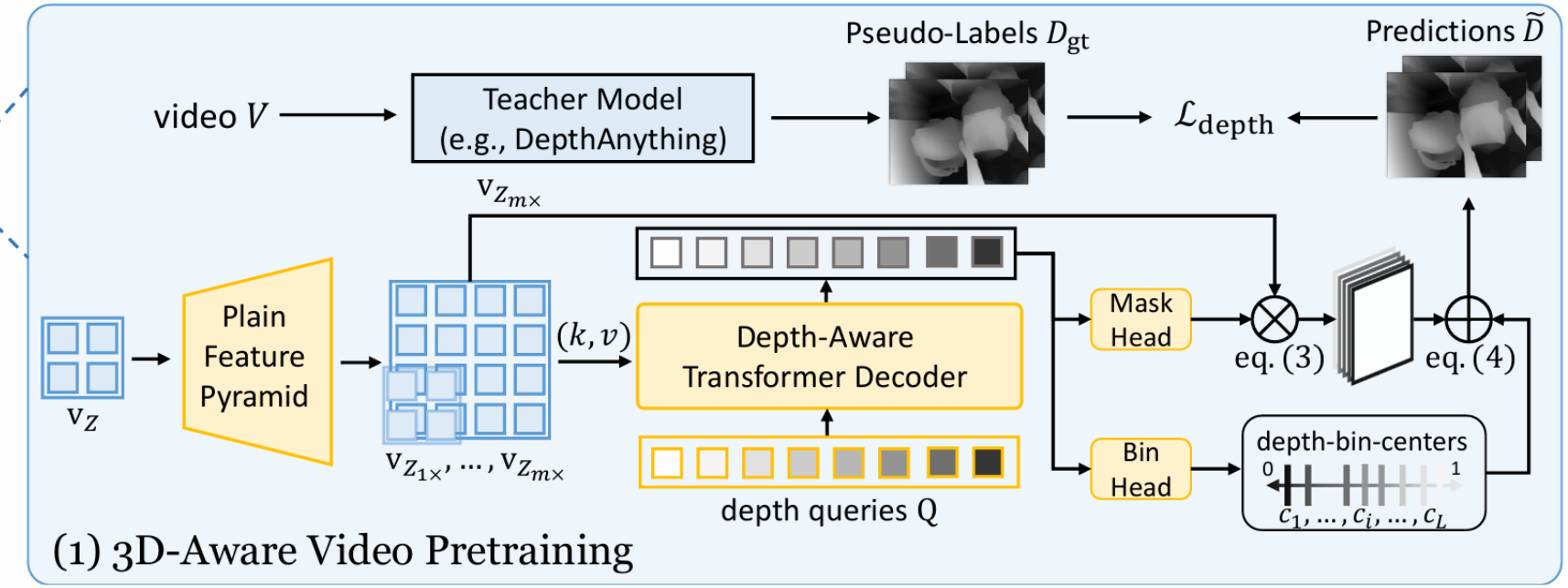


Target: 3D-aware representation v_Z for **low-resolution** depth estimation

- **PlainFPN**: got multi-scale features for better pixel-level prediction
- **Depth-Aware Transformer Decoder & Mask Head & Bin Head**: discretize depth prediction in previous SOTA fashion.
- **Extendable** for multi-task learning such as segmentation and detection.

Lightweight 3D-Aware Decoder Design

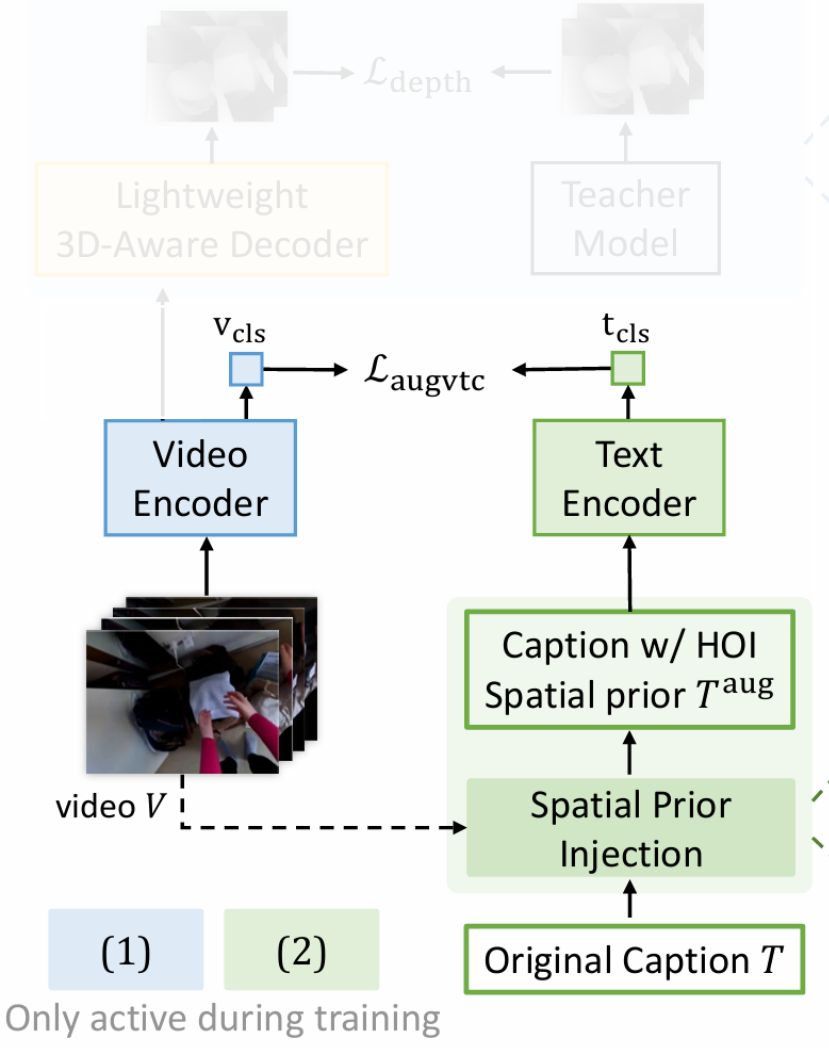




Target: 3D-aware representation v_Z for **low-resolution** depth estimation

- We got inspired from *SimpleFPN [ECCV22, He et al.]*. *AdaBins [CVPR21, Shariq et al.]*, *Mask2Former [CVPR22, Cheng et al.]*
- *See our full paper for details and motivation*

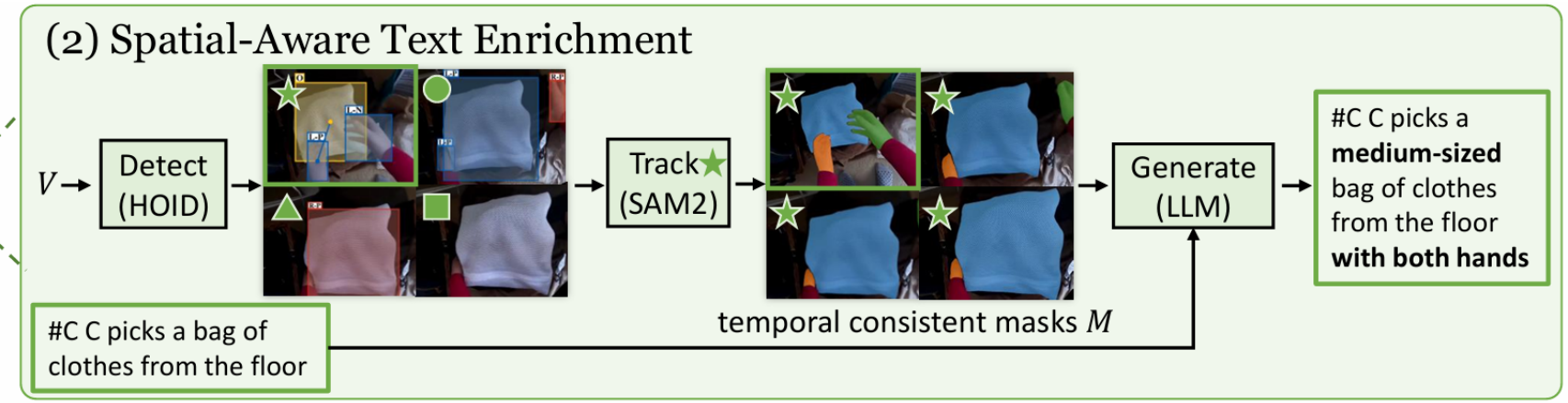
Detect-Track-Generate Pipeline



Target: from original text to spatially-informed text

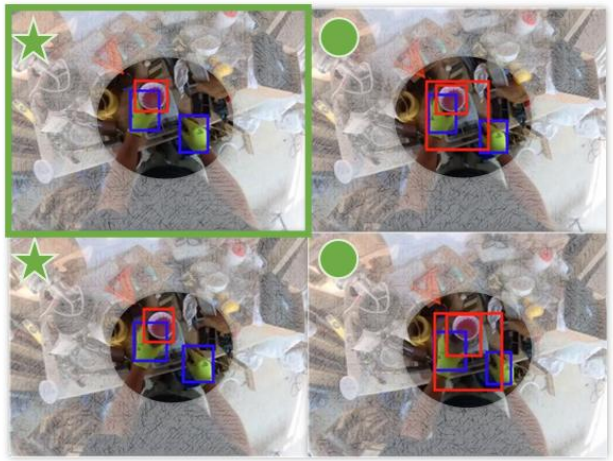
- **Detect-Track**: address the problem of Inconsistent HOI bounding boxes:

- **Generate**: leverage LLM to inject spatial information into text

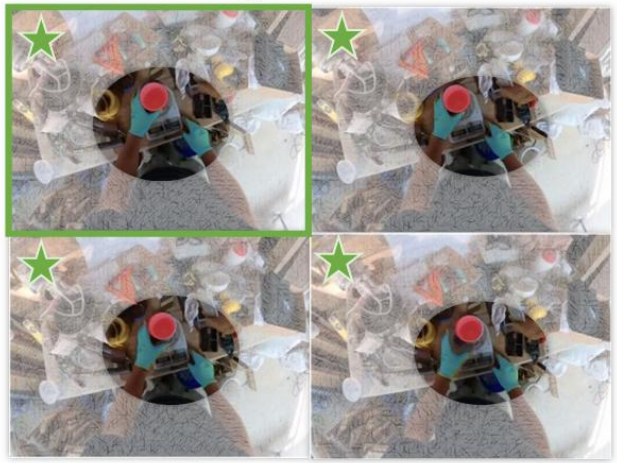


Case of Our Robust Data Generation

Original Video



Inconsistent HOI Boxes



Temporal Consistent HOI Masks

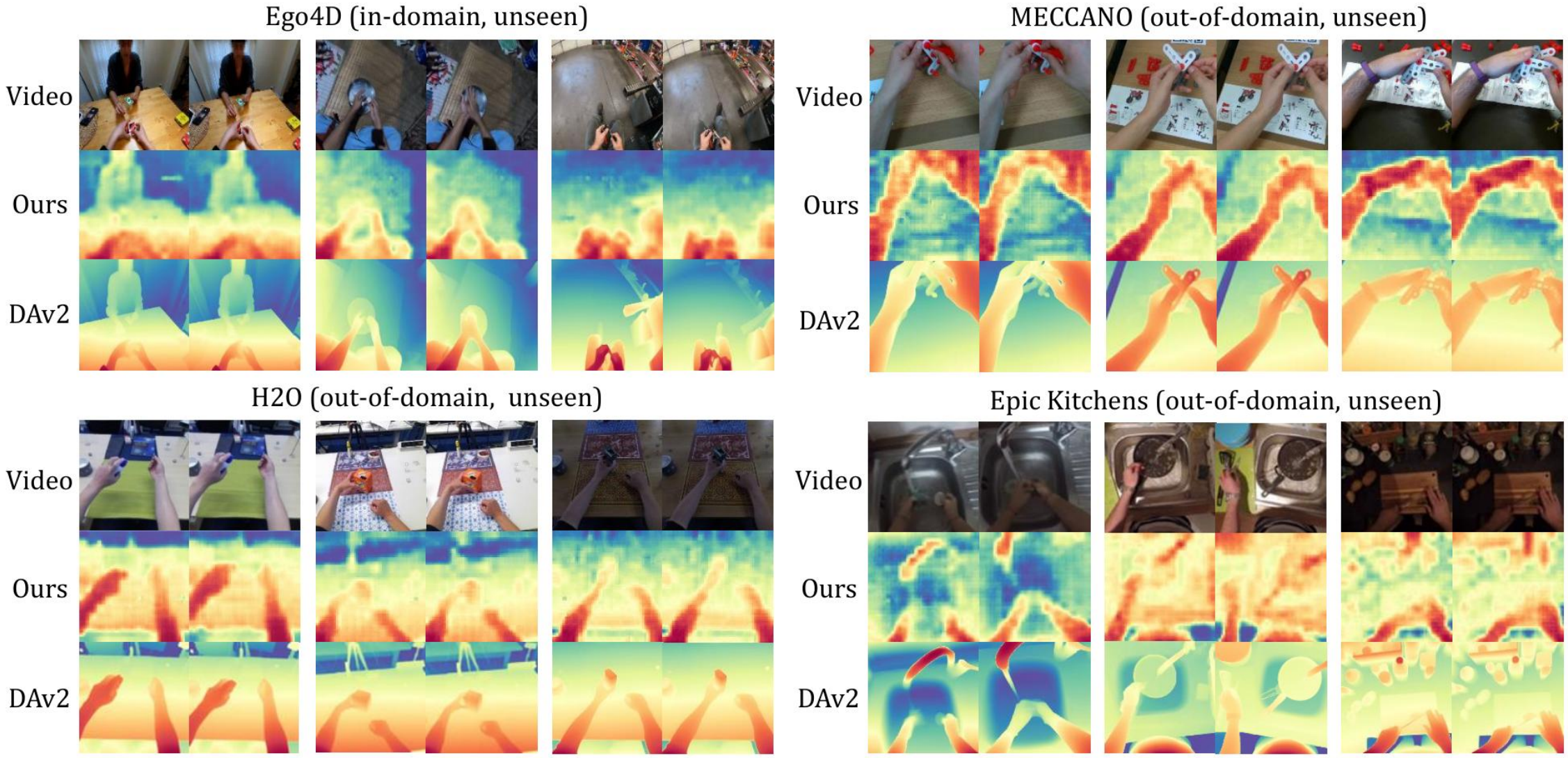
Original Captions:
#C C adjusts a mug on a weighing plate

Spatial-Aware Captions:
#C C adjusts a small mug on a weighing plate with the left hand, it slightly to the right and downward, while the right hand maintains contact with the stationary object

Spatial-Aware Captions

- Small objects
- Noisy background
- Consistent HOI masks

Generalization of the 3D-Aware Decoder



Evaluation Setups

■ Main Results

Short Video Understanding: **video-text retrieval** (Epic-Kitchens-100, EgoMCQ), **action recognition** (EGTEA, Epic-Kitchens-100-CLS)

Long Video Understanding: **natural language query** (EgoNLQ), **moment query** (EgoMQ)

3D-Aware Tasks: **robot manipulation** (Franka Kitchen), **Depth estimation** (H2O)

■ Pretraining Datset

2M (Video, Text) data + 2M (Video, Enriched Text, Depth) data

■ Evaluation Setting

Zero-shot: Directly apply evaluation after pretraining

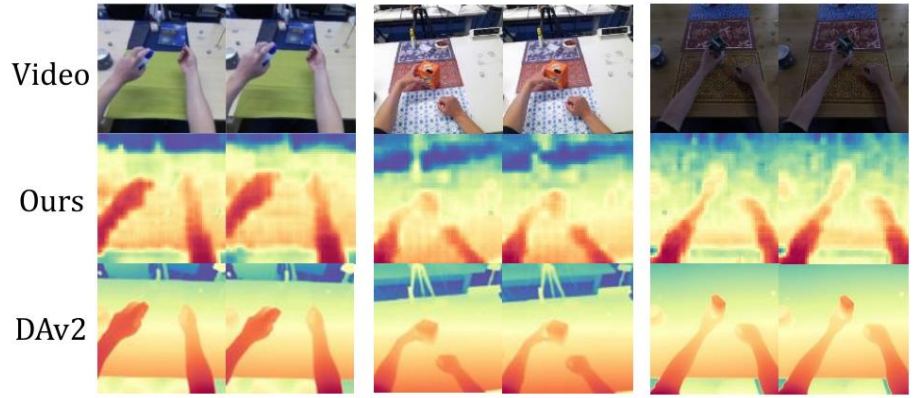
Fine-tune: Fine-tune on downstream tasks

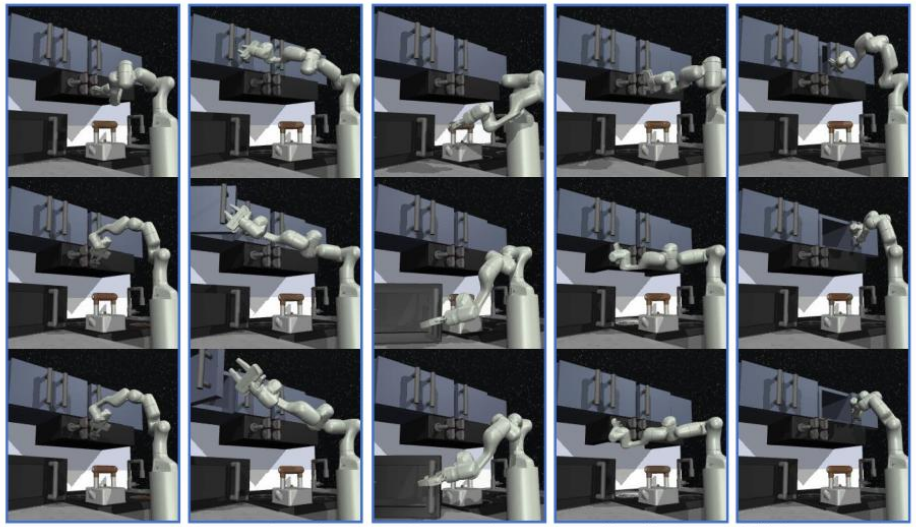
Superior Performance on Short Video Understanding

Method	Epic-Kitchens-100-MIR						EGTEA		EgoMCQ	
	mAP (%)			nDCG (%)						
	V→ T	T→ V	Avg.	V→ T	T→ V	Avg.	mean	top1	Inter	Intra
EgoVLP [36]	26.0	20.6	23.3	28.8	27.0	27.9	-	-	90.6	57.2
EgoVLPv2 [51]	35.1	26.6	30.8	33.7	30.4	32.0	30.9	35.1	91.0	60.9
LaViLa [87]	35.1	26.6	30.8	33.7	30.4	32.0	30.9	35.1	93.6	59.1
AVION [86]	37.1	28.7	32.9	34.4	31.0	32.7	38.6	42.3	94.4	62.1
HelpingHands* [80]	35.6	26.8	31.2	34.7	31.7	33.2	29.4	35.3	93.2	58.8
HENASY* [47]	35.5	27.1	31.3	34.6	31.7	33.2	29.6	35.9	94.1	61.3
EgoDTM (ours)	**37.9**	**29.1**	**33.5**	**34.8**	**31.9**	**33.4**	**40.2**	**43.2**	**94.6**	**63.6**

Improvement of VLP's 3D-Awareness

The egocentric video encoder functions as visual feature extractor, then process by task-specific methods.





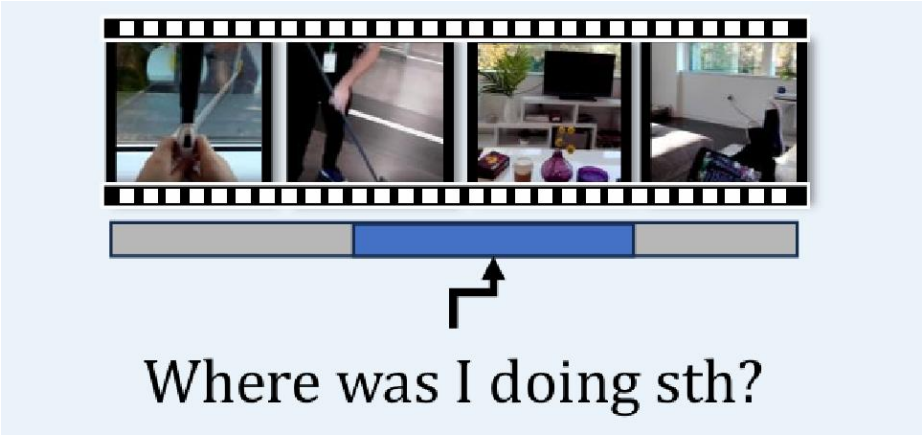
Depth Estimation

Method	Scale-Aware Metrics				Scale-Invariant Metrics			
	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	RMSE$\downarrow$	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$	RMSE$\downarrow$
ConvNext [40]	0.721	0.965	0.991	0.644	0.727	0.969	0.996	0.593
CLIP [52]	0.795	**0.966**	0.988	0.624	0.811	0.976	0.994	0.559
EgoVLP [36]	0.778	0.954	0.989	0.610	**0.853**	**0.977**	0.996	0.497
LaViLa [87]	0.801	0.954	0.987	0.598	0.811	0.964	0.993	0.552
AVION [86]	0.786	0.960	0.991	0.606	0.812	0.969	0.996	0.543
EgoDTM (ours)	**0.826**	0.964	**0.993**	**0.539**	0.848	**0.977**	**0.998**	**0.481**

Robot Manipulation

Method	TK	OD	OM	FS	SD	Average
R3M [45]	53.3%	50.7%	59.3%	86.3%	97.7%	69.4%
MPI [26]	83.3%	54%	44.5%	93.5%	100%	75%
ResNet [24]	28%	18%	26.7%	50%	75.5%	39.7%
CLIP [52]	26.3%	13%	24.7%	41.7%	86.3%	38.4%
LaViLa [87]	48%	26%	22.5%	69%	**94.5%**	52%
EgoDTM (ours)	**56%**	**28%**	**35.5%**	**81%**	92.5%	**58.6%**

Competitive on Long Video Understanding



The egocentric video encoder functions as visual feature extractor, then process by task-specific methods.

Natural language query: free-form text query

Method	R1@0.3	R5@0.3	R1@0.5	R5@0.5
EgoVLP [36]	6.32	13.84	3.41	8.80
LaViLa [87]	7.12	14.82	3.87	9.55
AVION [86]	7.33	14.89	4.31	9.53
EgoDTM (ours)	**8.13**	**16.11**	**4.83**	**10.30**

Moment query: action category query

Method	R1@0.3	R5@0.3	R1@0.5	R5@0.5
EgoVLP [36]	30.44	46.66	22.41	35.75
LaViLa [87]	32.9	48.68	**24.12**	37.59
AVION [86]	32.17	47.3	23.11	36.3
EgoDTM (ours)	**32.92**	**50.08**	23.94	**39.15**

Ablations on Major Components

- Generally improves over downstream benchmarks
- Text-depth joint learning strategy still have room to explore

Metrics	EK100MIR mAP / nDCG ↑	EgoMCQ inter / intra acc ↑	EK100CLS top-1 / top-5 acc ↑	EgoNLQ mIoU ↑	EgoMQ mAP ↑	DE scale-aware RMSE / scale-invariant RMSE ↓
$\mathcal{L}_{\text{vtc}}$	29.7 / 30.7	94.2 / 60.2	12.847 / 30.037	6.14	6.97	0.572 / 0.495
$\mathcal{L}_{\text{vtc}} + \mathcal{L}_{\text{depth}}$	31.3 / 31.2	94.2 / **62.6**	15.412 / 32.995	5.98	7.52	0.5637 / **0.464**
$\mathcal{L}_{\text{augvtc}}$	31.3 / 32.2	94 / 61.6	**16.508** / 32.851	**6.53**	6.14	0.550 / 0.489
$\mathcal{L}_{\text{augvtc}} + \mathcal{L}_{\text{depth}}$	**33.1 / 33.1**	**94.6 / 62.6**	15.898 / **33.895**	6.17	**8.87**	**0.539** / 0.481

Conclusion

- We introduce **EgoDTM**, a 3D-aware egocentric video-language model learned from 3D-aware video-language pretraining
- We develop a **lightweight 3D-aware decoder** for depth estimation and a **data construction pipeline** to enrich captions with spatial infor mation.
- Extensive experiments demonstrate that EgoDTM significantly enhances performance on **video understanding** tasks like video-text matching, and **3D understanding** tasks like robot manipulation.

Discussion

- Broader Generalization: While EgoDTM demonstrates strong performance in egocentric hand-object interaction scenarios, its generalization to broader *indoor scenarios* remains limited.
- 3D-Aware MLLM: Further exploration may include integrating 3D-aware visual encoders into *multimodal large language models (MLLM)* to enhance spatial awareness.
- More 3D Geometry pretraining: Moreover, pretraining large-scale spatial-aware egocentric models with richer 3D signals such as *camera parameters* and *point maps*, as explored in *VGGT*, remains a promising yet challenging direction.