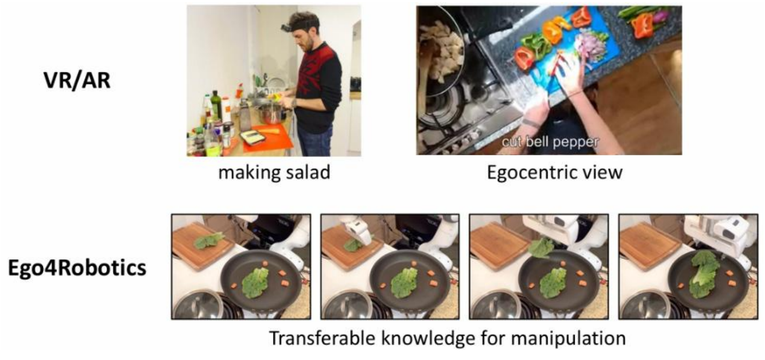
# EgoDTM: Towards 3D-Aware Egocentric Video-Language Pretraining

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

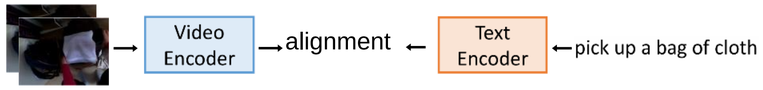
NeurIPS

## Background: Egocentric Video-Language Pretraining

- Egocentric videos: how humans see the world

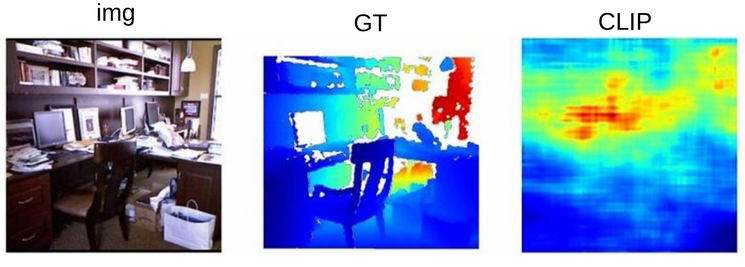



- Prevailing paradigm: egocentric video-language pretraining

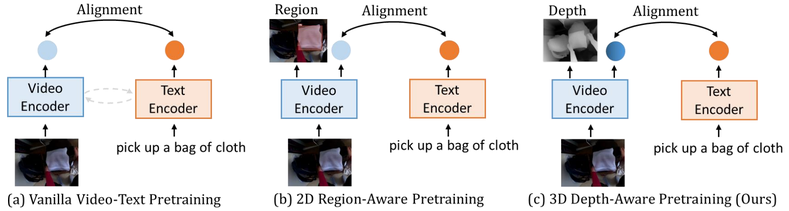



## Motivation: Develop 3D-Aware Video-Language Model

- Define **3D** as **Depth Maps**. VLMs are short of 3D perception




- Compare our 3D-awareness v.s. current 2D-aware approaches

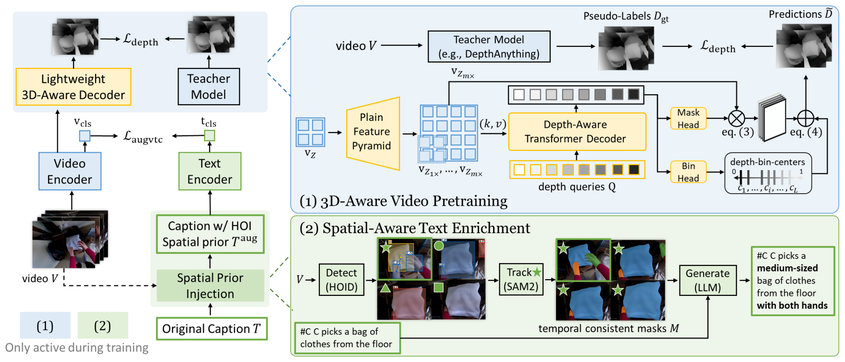



## Challenges: Joint Learning Architecture & Data Scarcity

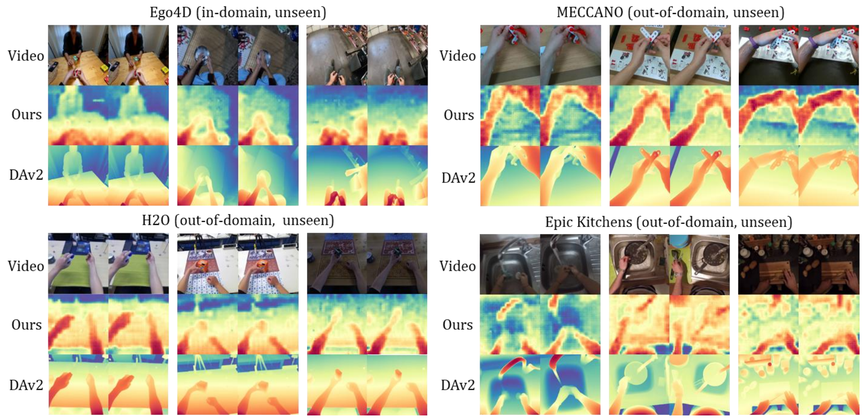
- Architectecture for learning from heterogenous supervision of (depth, text)
- Data Scarcity for million-level (video, depth, text)
  - *Depth: dense pixel-level knowledge, expensive & small-scale*
  - *Text: sparce non-pixel-level knowledges, short & no spatial-awareness*

## Method: 3D-Aware Egocentric Video-Language Pretrain

- (1) 3D-aware video pretraining: lightweight 3D-aware decoder design
- (2) Spatial-aware text enrichment: robust detect-track-generate pipeline




- Generalization of 3D-aware decoder




## Experiments

- Video-text retrieval & Action recognition

| Method | Epic-Kitchens-100-MIR mAP (%) V→T | T→V | Avg. | nDCG (%) V→T | T→V | Avg. | EGTEA mean | top1 | EgoMCQ 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** |

- Depth estimation on H2O

| Method | Scale-Aware Metrics $\delta_1\uparrow$ | $\delta_2\uparrow$ | $\delta_3\uparrow$ | RMSE↓ | Scale-Invariant Metrics $\delta_1\uparrow$ | $\delta_2\uparrow$ | $\delta_3\uparrow$ | RMSE↓ |
|---|---|---|---|---|---|---|---|---|
| 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** |

- Natural language query & moment query on Ego4D

| 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** |

| 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** |

- Robot manipulation on Franka Kitchen (simulation)

| 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%** |

- Ablation Study

| 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 ↓ |
|---|---|---|---|---|---|---|
| $\mathcal{L}_{vtc}$ | 29.7 / 30.7 | 94.2 / 60.2 | 12.847 / 30.037 | 6.14 | 6.97 | 0.572 / 0.495 |
| $\mathcal{L}_{vtc} + \mathcal{L}_{depth}$ | 31.3 / 31.2 | 94.2 / **62.6** | 15.412 / 32.995 | 5.98 | 7.52 | 0.5637 / **0.464** |
| $\mathcal{L}_{augvtc}$ | 31.3 / 32.2 | 94 / 61.6 | **16.508** / 32.851 | **6.53** | 6.14 | 0.550 / 0.489 |
| $\mathcal{L}_{augvtc} + \mathcal{L}_{depth}$ | **33.1 / 33.1** | **94.6 / 62.6** | 15.898 / **33.895** | 6.17 | **8.87** | **0.539** / 0.481 |