IKEA Manuals at Work: *4D Grounding of Assembly Instructions on Internet Videos*

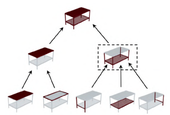
Yunong Liu[1], Cristobal Eyzaguirre[1], Manling Li[1], Shubh Khanna[1], Juan Carlos Niebles[1], Vineeth Ravi[2], Saumitra Mishra[2], Weiyu Liu*[1], Jiajun Wu*[1] (*Equal advising)
I'm looking for 2025Fall PhD/Internship
[1]Stanford University [2]J.P. Morgan AI Research

NEURAL INFORMATION PROCESSING SYSTEMS

Why it matters — Teaching AI to interpret assembly instructions

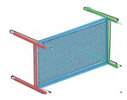
- Autonomous 3D shape assembly requires task planning, visual grounding, and control.
- Utilizing instructions, including manuals and how-to videos, requires grounding them in 3D over time.

Task Planning



extract dependencies and ordering of the subtasks

Visual Grounding

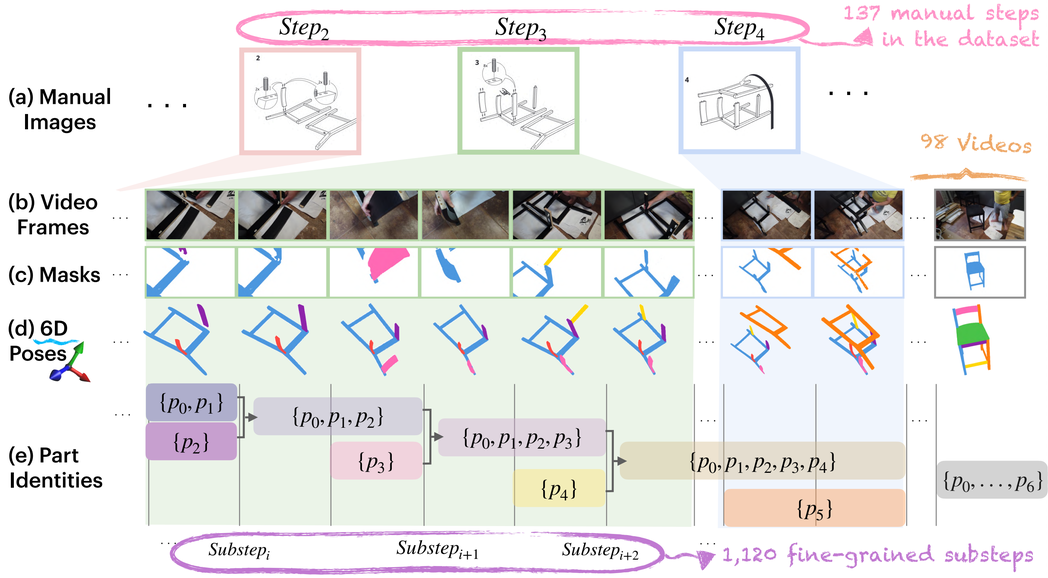


register parts with perceptual input to identity part geometry and pose

Control

compute motion that adheres to physical constraints

Our Dataset: IKEA Manuals at Work



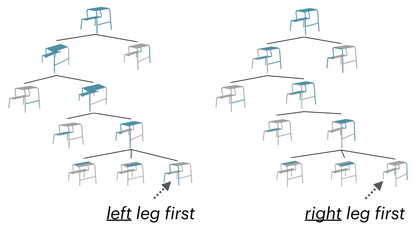
- The first multimodal dataset with **extensive 4D groundings** of assembly in internet videos.
- Reveal how furniture parts connect in 2D and 3D with parts' **6-DoF poses** and **segmentation masks.**
- Provide **fine-grained temporal alignments** (steps→substeps→frames) from instruction manuals to internet videos to 3D models.

A Diverse Collection of 36 IKEA Furniture Pieces

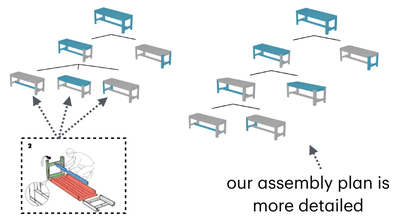
- 36 furniture items provide diversity in 3D geometry, structure, material, and dimension.

Detailed Step-by-Step Instructions

Plan Variations from Internet Videos



IKEA-Manual vs IKEA Video Manual



- 25% of furniture objects have more than one unique assembly sequences.

Complexities in Real-World Videos

Camera Changes
zooms, movements, switches between views

Similar Parts
ambiguous assembly and disassembly steps.

Heavy Occlusion
occlusion with hands and other parts

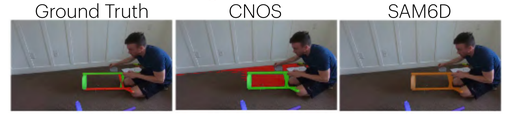
Complex Backgrounds
e.g., multi-person and outdoor assembly.

- Instruction videos from the internet bring challenges in visual grounding.

Applications

Break down 4D grounding of videos into sub-tasks.

Part-conditioned Segmentation



Challenges from visual complexities.

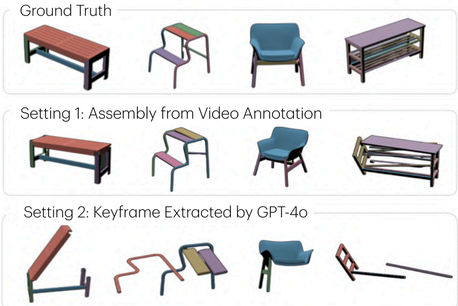
Part conditioned Pose Estimation

Inaccurate estimated depth can weaken the results.

Mask Tracking

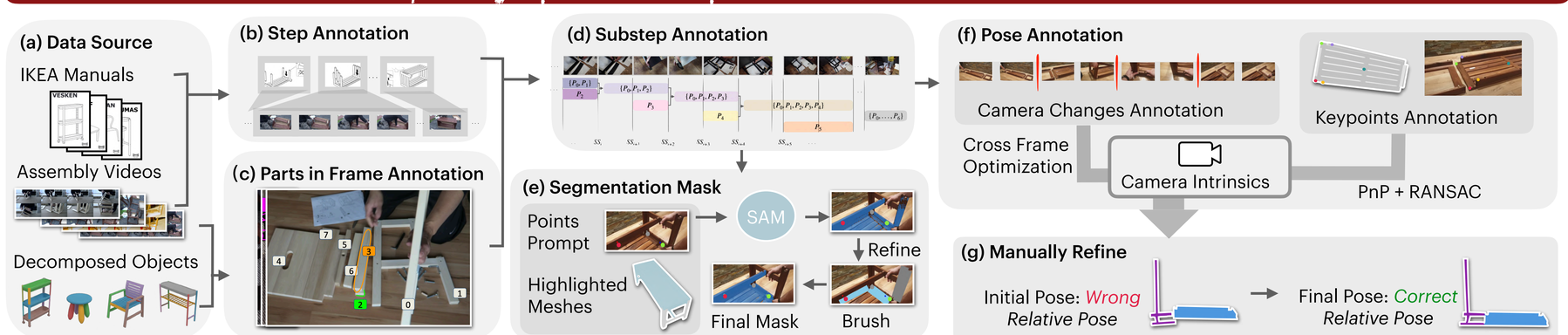
Harder than existing benchmarks for VOS.

Shape Assembly with Instruction Videos

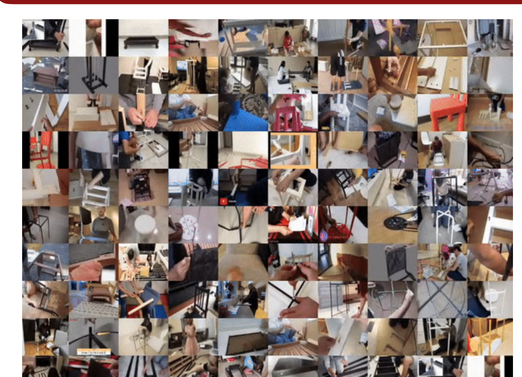


VLMs face challenges in robustly detecting steps and identifying involved parts in videos.

How we got the high-quality spatial-temporal aligned annotations? A semi-automatic pipeline



More details on the website!



We are **the first dataset to provide 6D pose annotations on internet videos** for furniture assembly

Scan for more details!