Supplementary Material for IKEA Manuals at Work: 4D Grounding of Assembly Instructions on Internet Videos

Yunong Liu[1] **Weiyu Liu**[1] **Shubh Khanna**[1] **Cristobal Eyzaguirre**[1] **Manling Li**[1]
Juan Carlos Niebles[1] **Vineeth Ravi**[2] **Saumitra Mishra**[2] **Jiajun Wu**[1]

[1]Stanford University [2]J.P. Morgan AI Research
Website: https://yunongliu1.github.io/ikea-video-manual/

Contents

A Dataset Details

IKEA Video Manuals is a large-scale multimodal dataset with high-quality, spatial-temporal alignments of step-by-step instructions, 3D object representations, and real-world video demonstrations from the Internet. IKEA Video Manuals provides 34,441 annotated video frames, aligning 36 IKEA manuals with 98 assembly videos for six furniture categories. Fig. A1 shows all 3D furniture models included in the dataset. An example of the annotations associated with each frame is shown in Fig. A2. We provide details of the data and annotations associated with each frame below.

Furniture-level information

- **Category:** The category label of the furniture (e.g., Bench).
- **Name:** The furniture name (e.g., applaro).
- **Furniture IDs:** A list of IKEA product IDs for the furniture.
- **Variants:** A list of furniture variants, if applicable.
- **Furniture URLs:** A list of IKEA product page URLs for the furniture.
- **Furniture Main Image URLs:** A list of URLs for the main product images on the IKEA website.

Video-level information

- **Video URL:** The URL of the video.
- **Additional Video URLs:** A list of additional video URLs for the same furniture.
- **Title:** The title of the video.
- **Duration:** The duration of the video (in seconds).

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Figure A1: All furniture items included in the IKEA Video Manuals dataset, categorized by type–Desk, Table, Chair, Bench, and Misc.

- **Resolution:** The resolution of the video (e.g., 1920x1080).
- **FPS:** The frame rate of the video (e.g., 30).
- **People Count:** The number of people in the video.
- **Person View:** The view of the person in the video (e.g., front, side).
- **Camera Fixation:** The fixation of the camera in the video (e.g., static, moving).
- **Indoor/Outdoor Setting:** The setting of the video (e.g., indoor, outdoor).

Assembly step information

- **Step ID:** An unique ID is assigned to the assembly step.
- **Step Start:** The start time of the assembly step is shown in the video.
- **Step End:** The end time of the assembly step is shown in the video.
- **Substep ID:** The unique ID assigned to the substep within the assembly step.
- **Substep Start:** The start time of the substep in the video.
- **Substep End:** The end time of the substep in the video.

Frame-level information

- **Frame Time:** The timestamp of the frame in the video.
- **Number of Camera Changes:** The number of camera changes that have been labelled before the current frame.
- **Frame Parts:** A list of parts that are labelled in the frame (e.g., $[\{0, 2\}, \{1\}, \{3\}]$). The sub-assemblies that have been constructed in previous steps are denoted by a tuple of the part IDs.
- **Frame ID:** An unique identifier for the frame (e.g., 1584).
- **Is Keyframe:** A boolean value indicating whether the frame is a keyframe.
- **Is Frame Before Keyframe:** A boolean value indicating whether the frame is immediately before a keyframe.
- **Frame Image:** The RGB image of the frame.
- **Annotated Masks:** A list of segmentation masks for the parts in the frame.
- **Annotated Poses:** A list of 6D poses for the parts in the frame.

Manual information

- **Manual Step ID:** A unique ID assigned to the assembly step. This ID is associated with the step IDs in frame annotations.
- **Manual URLs:** A list of URLs for the assembly manual PDFs.
- **Manual ID:** An unique ID of the assembly manual from IKEA.
- **Manual Parts:** A list of part IDs is shown in the corresponding manual step.
- **Manual Connections:** List of connections between parts in the manual step.
- **PDF Page:** The page number of the manual step in the PDF.
- **Cropped Manual Image:** The cropped image of the corresponding manual step.

B Datasheets

Dataset description. The datasheet for the IKEA Video Manuals dataset, available at `https://github.com/yunongLiu1/IKEA-Manuals-at-Work/blob/main/datasheet.md`, including key aspects of data collection and annotation:

- **Consent:** The dataset is built upon two existing datasets, IKEA-Manual and IAW, which are publicly available for research purposes under the Creative Commons Attribution 4.0 International (CC-BY-4.0) license.
- **Personally Identifiable Information and Offensive Content:** The dataset does not contain any personally identifiable information or offensive content, as it focuses on furniture objects and assembly instructions.

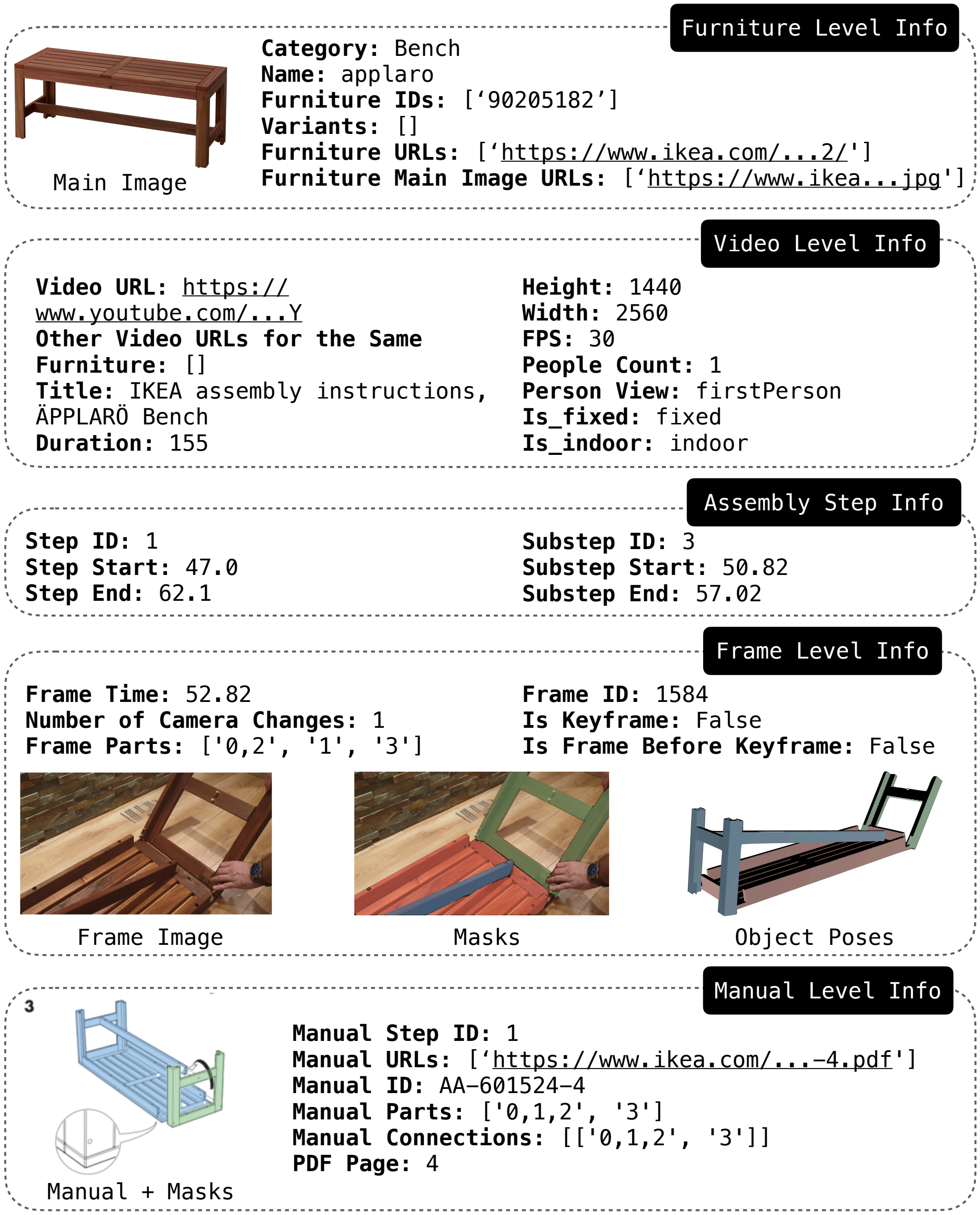

Figure A2: An example of an annotated frame in the IKEA Video Manuals dataset. The annotation and data are divided into furniture-level, video-level, assembly step-level, frame-level, and manual-level information.

- **Annotation Process and Compensation:** The data annotation process was outsourced to an annotation company. The annotators were compensated based on the work they provided, with the estimated hourly pay being above the minimum wage.

Please refer to the datasheet for more detailed information on the dataset.

Link and license. The dataset is available for public access under the CC-BY-4.0 license: `https://github.com/yunongLiu1/IKEA-Manuals-at-Work`

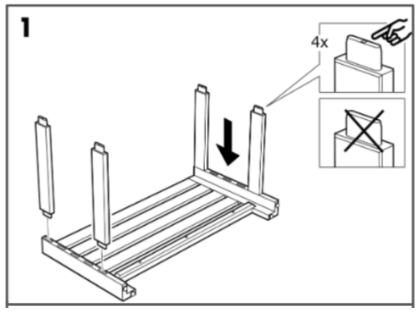

Figure A3: An example of an assembly step from the IKEA instruction manual that involves the assembly of four parts.

Maintenance. The dataset is hosted on GitHub and will be maintained by the authors. The repository can be found at: `https://github.com/yunongLiu1/IKEA-Manuals-at-Work`. The dataset has the following DOI: `https://doi.org/10.5281/zenodo.11623997`

Author statement. The authors bear all responsibility in case of violation of rights. All annotations were collected by the authors and we are releasing the dataset under CC-BY-4.0.

Format. The dataset contains videos, 3D models, manual PDFs, and annotated data (including temporal step alignments, temporal substep alignments, 2D-3D part correspondences, part segmentations, part 6D poses, and estimated camera parameters). The annotated data is stored in the `JSON` format. Other data are stored in their original formats, and uploaded in a `zip` file. Upon decompression, the dataset is organized into subdirectories for videos, 3D models, and manual PDFs. Each is organized into subdirectories for furniture categories, and further subdirectories for individual furniture items.

Croissant Metadata We will provide the structured metadata (schema.org standards) in `https://github.com/yunongLiu1/IKEA-Manuals-at-Work/metadata.json`.

C Details for Data Annotation

To create the IKEA Video Manuals dataset, we identified 36 IKEA objects from the IKEA-Manual dataset [1] that have corresponding assembly videos in the IAW dataset [2]. We matched the unique IDs of the instruction manuals to ensure correct correspondence between the datasets. We provide additional details for each of the annotation steps below.

C.1 Annotating Assembly Steps

For each assembly step annotated in the IKEA-Manual dataset [1], we identify matching video segments in the IAW dataset [2]. We manually adjust the start and end time of each video segment to include a more complete assembly process, from picking up a part to positioning and tightening. The adjustment ensures better alignment with the physical assembly actions.

C.2 Annotating Assembly Substeps

In the IKEA instruction manuals, each single step may involve the assembly of multiple parts (as shown in Fig. A3). We provide a more fine-grained assembly process by introducing *substeps*. A substep is labelled when 1) a new part appears in the video or 2) a new sub-assembly is created through positioning and/or fastening of parts. On average, each assembly step contains 7.59 substeps. In total, the IKEA Video Manuals dataset contains 1120 substeps.

C.3 Annotating Part Identities

In our dataset, each part of the 3D furniture model is assigned a unique ID consistent with the IKEA-Manual dataset. However, locating inividual 3D furniture part in the frame can be challenging due to several ambiguities, as illustrated in Fig. A4:

(a) Wrongly assembled parts that are initially placed incorrectly and later relocated, causing confusion for the annotator (Fig. A4a).

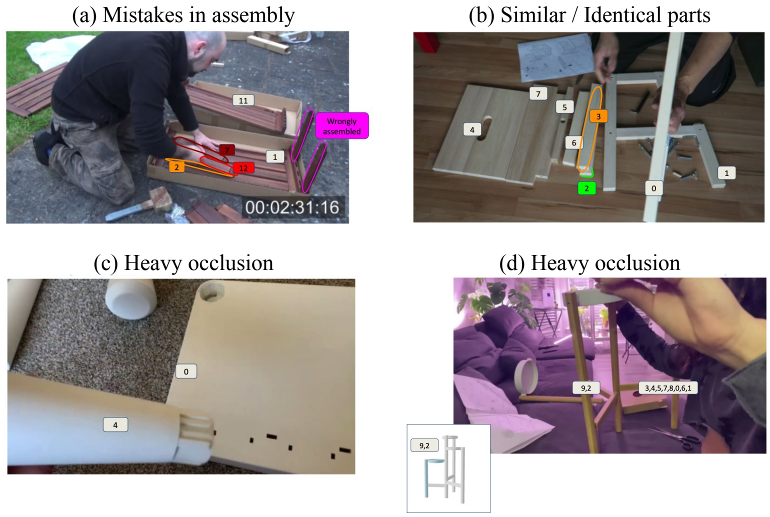

Figure A4: Examples of ambiguities in annotating part identities in assembly videos: (a) Wrongly assembled parts later relocated, (b) Similar-looking parts that are difficult to distinguish, (c-d) Heavily occluded parts and boundaries between parts.

Figure A5: An example of refining the part segmentation using the brush tool. The initial segmentation is generated by the Segment Anything Model (SAM) model.

(b) Similar or identical-looking parts, such as chair legs, that are difficult to distinguish and label accurately (Fig. A4b).

(c) Parts that are heavily occluded, making it challenging to recognize the parts and their boundaries (Fig. A4c-d).

To ensure accurate part tracking throughout the video, we manually label the parts in the first frame of each substep after watching the entire video. This annotation is crucial for maintaining consistency when annotators only see individual frames in subsequent annotations. This approach ensures consistent part identities throughout the video, addressing challenges posed by heavy occlusions, similar-looking parts, and assembly mistakes.

C.4 Annotating Segmentation Mask

To efficiently generate segmentation masks for furniture parts in video frames, we utilize the Segment Anything Model (SAM). When SAM fails to generate accurate masks (e.g., Fig. A5), we use a brush tool built into our annotation interface to refine the masks manually.

C.5 Annotating 2D-3D Keypoints

To establish correspondence between the 3D parts and their 2D projections in the video frames, we annotate keypoints on both the 3D parts and the 2D images. Our annotation interface is shown

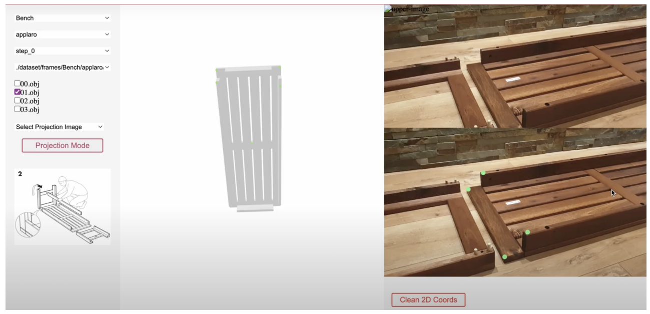

Figure A6: The annotation interface for labelling keypoint correspondences between 3D models and 2D video frames.

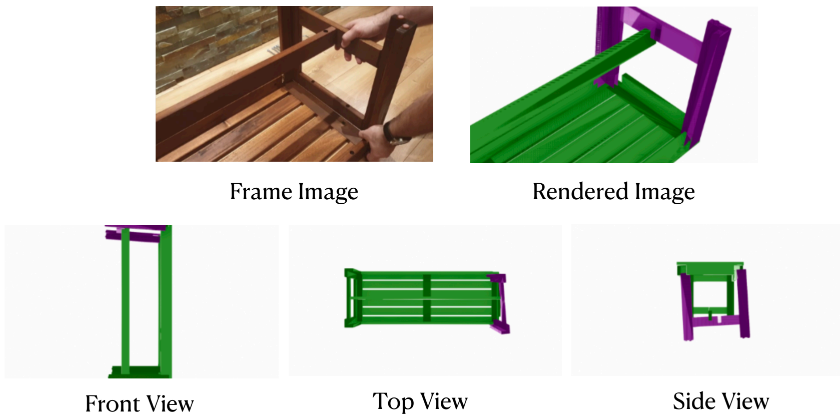

Figure A7: An example of an inaccurate part pose estimated from the 2D-3D keypoints. Despite that the 2D projections of the 3D models overlap with the parts in the video frame (top row), the 3D poses of the parts can be found incorrect when visualized in 3D and viewed from other angles (bottom row).

in Fig. A6. The annotation interface computes the part poses and camera parameters using the Perspective-n-Point (PnP) algorithm and visualizes the 2D projection in real-time. Based on the visualization, the annotator can interactively refine the keypoint annotations to maximize the overlap between the 2D projection and the part seen in the 2D image.

C.6 Annotating Camera Changes

A prerequisite for achieving spatially and temporally accurate pose annotation is a correct estimation of camera parameters from the video. Many videos in the dataset include changes in camera viewpoints, camera movements, and adjustments of focal lengths. These changes can potentially lead to different camera parameters. To account for these factors, we manually annotate camera changes in the IKEA Video Manuals dataset. By annotating all frames when a camera change occurs, we can estimate the intrinsic parameters for each video segment between two camera changes, assuming that the intrinsic parameters remain consistent within the segment.

C.7 Pose Refinement

While initial estimates of the part poses can be obtained from the annotated 2D and 3D keypoints, these estimates are often inaccurate, particularly in terms of the relative positions and orientations of the parts in 3D space. Fig. A7 shows an example where the 2D projection of a part appears correct, but when viewed in 3D, the part is positioned incorrectly relative to other parts. To address this issue, we developed an interface that allows annotators to refine the initial estimate by rotating and translating each part in 3D space. The annotators can view 3D parts from different viewpoints to ensure that the relative poses of the parts are correct and consistent with the assembly process seen in the videos.

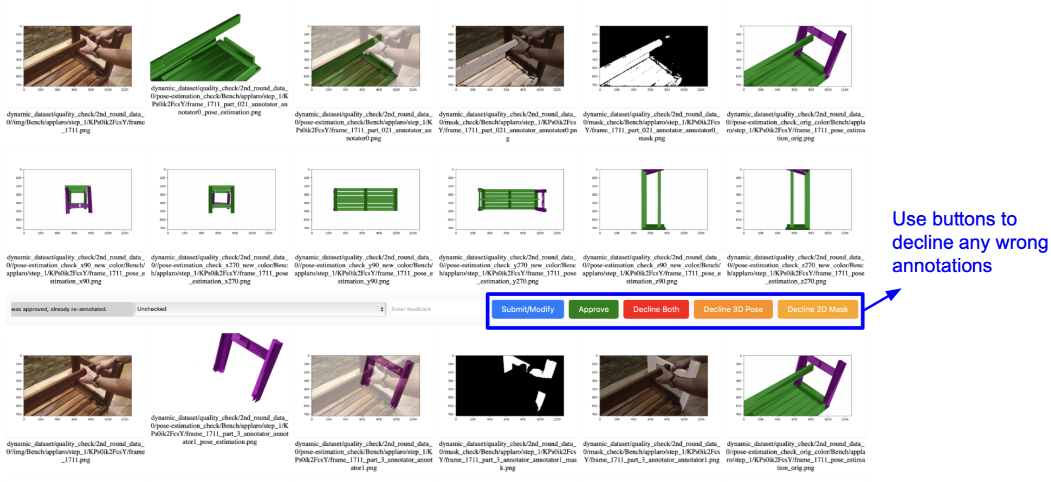

Figure A8: The verification interface for assessing the quality of mask and pose annotations. The interface visualizes the original video frame, mask overlay, pose overlay, and 3D parts in estimated poses from different viewpoints.

We take an additional step in the pose refinement process to help maintain the temporal smoothness of the part trajectories. During the pose refinement process, the initial poses of the parts in a frame are set to the refined poses of the corresponding parts from the previous frame. This initialization strategy helps to reduce large changes in the annotated part poses between neighboring frames, resulting in more coherent and realistic pose trajectories.

D Details for Quality Control

To ensure the accuracy and consistency of the annotations in the IKEA Video Manuals dataset, we analyze the common errors in the annotations and perform extensive verifications.

D.1 Common Errors

For mask annotations, common errors include incorrect part segmentation, missing parts, and noisy masks. Incorrect part segmentation occurs when annotators misidentify the boundaries of a part due to similar colours or complex shapes. Missing parts occur when certain parts are not segmented, often due to occlusions. Noisy masks often occur when the SAM model fails to generate accurate masks, leading to incomplete or inaccurate segmentation.

For pose annotations, common errors include incorrect part identification, incorrect relative poses, and interpenetrations. Incorrect part identification occurs when the annotators annotate an incorrect part, leading to an incorrect pose. Incorrect relative poses occur when the estimated pose does not accurately reflect the actual position and orientation of the part relative to other parts in 3D space. Interpenetrations occur when parts intersect or overlap in 3D space, leading to unrealistic poses.

D.2 Extensive Verification

We conduct extensive verification to ensure the high quality of the mask and pose annotations. The verification interface (as shown in Fig. A8) displays the original video frame, the video frame overlaid with segmentation masks, the video frame overlaid with 2D projections based on estimated poses, and the 3D parts in the estimated poses from different viewpoints. In particular, for mask annotations, we verify if the 2D mask corresponds to the correct part, covers the entire part, does not contain pixels of other parts, and is free of noise due to limitations of the Segment Anything Model (SAM). For pose annotations, we verify if the pose annotation corresponds to the correct part, the 2D projection aligns with the part in the frame image, and the 3D parts have correct relative poses. We automatically filter out pose annotations with interpenetrations between parts.

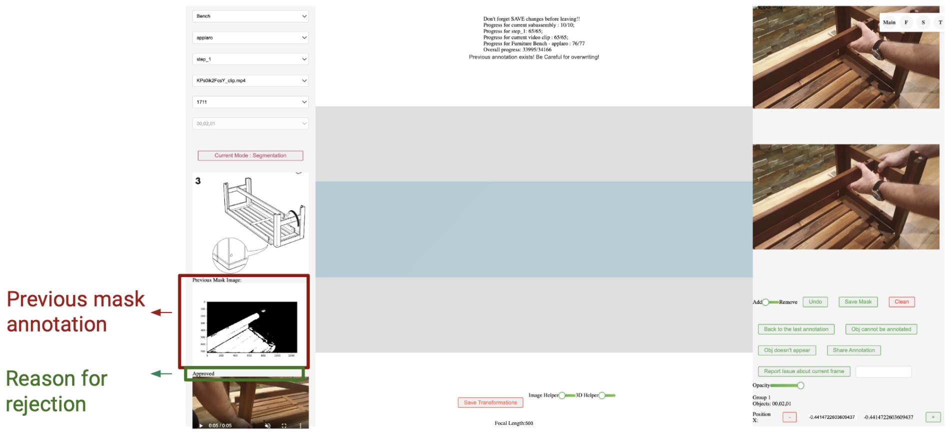

Figure A9: The interface for correcting and refining mask annotations based on feedback. The interface provides tools for manual refinement of the segmentation masks including a brush and an eraser.

E Annotation Interfaces and Instructions

This section provides details of the instructions given to annotators. Instructions for using these interfaces are mainly provided through demonstration videos, which are included in the project's GitHub repository (`https://github.com/yunongLiu1/IKEA-Manuals-at-Work`) for reference.

E.1 Segmentation Mask and 2D-3D Points Correspondence Annotation Interface (Fig. A6)

The Segmentation Mask and 2D-3D Points Correspondence Annotation Interface allow annotators to generate segmentation masks and establish correspondences between 3D models and 2D video frames. Annotators can switch between the two annotation modes using a dedicated button in the interface. The following steps outline the annotation process:

1. Select the appropriate category, subcategory, object, and step for the video you want to annotate.
2. In the Segmentation Mask mode:
   - Select points that best represent the overall shape and area of the part to ensure optimal performance of the Segment Anything Model (SAM).
   - Use the provided tools, such as a brush or eraser, to refine the mask based on the feedback provided.
3. In the 2D-3D Points Correspondence mode:
   - Select corresponding points on the 3D model and the 2D video frame that represent key features or edges of the furniture parts.
   - Review the rendered image and adjust the selected points if necessary to improve the alignment between the 3D model and the 2D video frame.
4. Navigate frames using the 'Next Frame' button and review the predicted points from the TAPIR model, modifying any unsatisfactory points.
5. Review the segmented images and estimated poses for accuracy and consistency, and submit the annotations.

E.2 Mask Re-Annotation Interface (Fig. A9)

1. **Review Previous Mask:**
   - The interface will display the previously annotated mask in the bottom left corner of the screen, along with the reason for the decline. Reviewing the previous mask and the reason for the decline helps the annotator understand the required corrections.

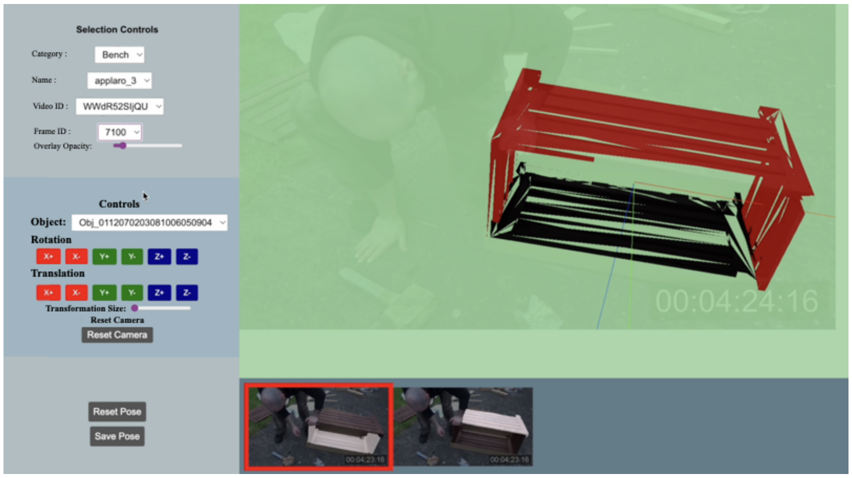

Figure A10: The interface for manually adjusting part poses in 3D. The interface supports adjusting the 3D position and orientation of each part. The interface also provides visualization of the 3D parts from different perspectives.

- Reasons for the decline include "mask was annotated to the wrong part," "mask did not include the whole part/include other parts," or "noisy mask, which is normally caused by the limitation of SAM and can be solved by using the brush." These specific reasons guide the annotator in refining the mask.

2. **Refine Mask:**
   - Use the provided tools, such as a brush or eraser, to refine the mask based on the feedback provided. These tools allow precise modifications to the mask.
   - Ensure the refined mask accurately captures the entire part while excluding any neighbouring parts or background. An accurate and complete mask is essential for downstream tasks.
   - Pay attention to the edges and boundaries of the part to create a clean and precise mask. Well-defined edges improve the quality and usability of the mask.
3. **Additional Buttons:** (Same as in the Segmentation Mask Annotation Interface)
4. **Review and Submit:** Review the refined mask for accuracy and completeness, ensuring it addresses the reason for the decline. This final review step verifies that the necessary corrections have been made. Submit the updated mask using the provided submission functionality to save the work and proceed to the next task.

E.3 Pose Refinement Interface (Figure A10)

The Pose Refinement Interface enables annotators to refine the initial poses estimated from the previous annotation. The following steps outline the pose refinement process:

1. Review the initial poses of all parts in the frame, estimated from the previous annotation.
2. Use the provided controls to adjust the position and orientation of each part in the camera frame.
3. Ensure that the relative positions and orientations of the parts are consistent with the assembly process.
4. Review the refined poses for accuracy and submit the updated poses.

By following these instructions and leveraging the provided video demonstrations, annotators can effectively use the annotation interfaces to generate high-quality segmentation masks, 2D-3D point correspondences, and refined part poses for the IKEA Video Manuals dataset.

F Error Bar

To assess the variability of our model's performance, we run experiments on a subset of the data with 3 different random seeds for both the segmentation and pose estimation tasks. For part-conditioned segmentation, the standard deviation of the IoU metric is 0.01 for the CNOS method and 0.03 for the SAM-6D method. When considering the Top-5 IoU, the standard deviations are 0.08 and 0.09 respectively. For part-conditioned 6D pose estimation, the SAM-6D method has a standard deviation of 0.12 for the ADD score and 0.08 for the ADD-S score. The MegaPose method has standard deviations of 0.09 and 0.05 for ADD and ADD-S. These results indicate that the performance of these models on our dataset remains relatively consistent overall.

G Compute Resources

The computational resources used for this project were computing nodes from the Stanford SC computational cluster. We used around 40 jobs lasting 7 days for running segmentation experiments using SAM-6D and CNOS, and pose estimation experiments using SAM-6D and MegaPose. The jobs were assigned to nodes equipped with different NVIDIA GPU models, including 2080 Ti, Titan RTX, 3090, A40, A5000, A6000, and L40S, based on availability.

References

[1] Wang, R., Zhang, Y., Mao, J., Zhang, R., Cheng, C.Y., Wu, J.: Ikea-manual: Seeing shape assembly step by step. Advances in Neural Information Processing Systems **35**, 28428–28440 (2022)

[2] Zhang, J., Cherian, A., Liu, Y., Ben-Shabat, Y., Rodriguez, C., Gould, S.: Aligning step-by-step instructional diagrams to video demonstrations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023)