Aligning Step-by-Step Instructional Diagrams to Video Demonstrations Supplementary Material

Jiahao Zhang[1,*] Anoop Cherian[2] Yanbin Liu[1]
Yizhak Ben-Shabat[1,3,†] Cristian Rodriguez[4] Stephen Gould[1,‡]
[1]The Australian National University, [2]Mitsubishi Electric Research Labs
[3]Technion Israel Institute of Technology, [4]The Australian Institute for Machine Learning
[1]{first.last}@anu.edu.au [2]cherian@merl.com [3]sitzikbs@gmail.com [4]crodriguezop@gmail.com
https://davidzhang73.github.io/en/publication/zhang-cvpr-2023/

A. IAW Dataset Statistics

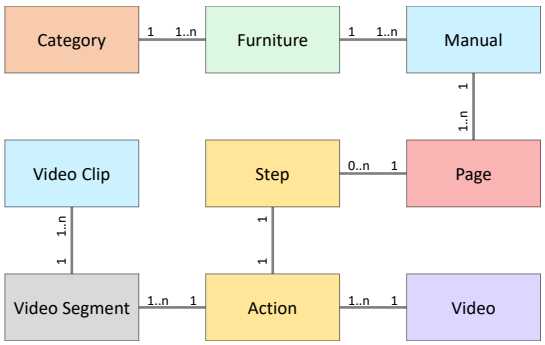


Figure 1. Entity Relationship (ER) Diagram of entities defined for the IAW dataset.

We defined several entities to describe the hierarchical structure of Ikea Assembly in the Wild (IAW) dataset as demonstrated in Fig. 1. In terms of the instructional assembly manuals, we have categories, furniture, manuals, pages and steps. As for videos, there are videos, actions, video segments (10s) and clips (64 frames 2.133s). A matching is made between a step in the manual and an action from the video, which is a one-to-one relation. The relation between step and page is many (0..n)-to-one, because a page may contain other information instead of an instructional diagram. Besides the above two, the rest are all many (1..n)-to-one relation. It is worth noting that one piece of furniture may correspond to multiple manuals, and we concatenate the manuals according to the assembly order.

In terms of the data collection, we first crawled all manuals under the category *Furniture* from Ikea official website. We manually found all related assembly videos from YouTube, split the PDF manual into pages and cropped out every individual step diagram. With the above information, we out-sourced video to diagram alignment tasks to Amazon Mechanical Turker platform. The alignments were then audited and refined carefully. Final statistics are shown in Tab. 1 and Tab. 2.

Table 1. Statistics of assembly manuals categorized by each furniture category.

Category	#furniture	#manuals	#pages	#steps
Beds	33	37	769	823
Bookcases & shelving units	55	61	961	1152
Cabinets & cupboards	28	36	851	920
Chairs	74	77	632	884
Chests of drawers & drawer units	35	50	1288	1194
Children's furniture	5	5	80	84
Furniture sets	2	2	11	20
Gaming furniture	1	1	32	24
Outdoor furniture	11	11	79	88
Sofas	28	31	500	540
TV & media furniture	11	11	321	300
Tables & desks	117	119	2129	1947
Trolleys	3	3	48	40
Wardrobes	17	17	562	552
Total	420	461	8263	8568
Per Category Median	22.5	24.0	531.0	546.0
Per Category Average	30.0	32.9	590.2	612.0

The distributions of video duration and the number of actions per video are shown in the boxplot below (Fig. 2).

As shown in Fig. 3, we manually attached four attributes for each video. When splitting the IAW dataset into train, validation and test splits, a greedy algorithm is used to balance the distribution in each split w.r.t. four attributes as shown in Fig. 4. Concretely, the greedy algorithm traverses each furniture. If a furniture contains only one video, then the video is added to train split; if two videos, then one for train and one for test; if more than two videos, one for test, one for validation and put all the rest into the train split.

*Supported by an ANU-MERL PhD scholarship agreement.
†Supported by Marie Sklodowska-Curie grant agreement No. 893465.
‡Supported by an ARC Future Fellowship (No. FT200100421).

Table 2. Statistics of assembly videos categorized by each furniture category. Annotated duration denotes the total duration of video segments that have labels.

Category	#videos	#actions	Duration	Annotated Duration
Beds	86	1554	16h:43m:00s	10h:50m:59s
Bookcases & shelving units	128	1636	21h:23m:55s	12h:42m:08s
Cabinets & cupboards	49	1082	11h:29m:12s	07h:12m:38s
Chairs	185	1478	21h:22m:27s	11h:44m:13s
Chests of drawers & drawer units	113	2891	30h:51m:36s	20h:40m:37s
Children's furniture	7	52	01h:14m:37s	00h:32m:27s
Furniture sets	3	15	00h:17m:19s	00h:02m:36s
Gaming furniture	1	31	00h:00m:56s	00h:00m:42s
Outdoor furniture	11	72	01h:25m:08s	00h:37m:57s
Sofas	83	1285	15h:08m:20s	08h:44m:44s
TV & media furniture	36	749	07h:41m:47s	05h:31m:39s
Tables & desks	266	3947	47h:07m:37s	29h:26m:20s
Trolleys	8	79	00h:51m:59s	00h:34m:14s
Wardrobes	31	812	07h:29m:12s	05h:02m:30s
Total	1007	15683	183h:07m:05s	113h:43m:48s
Per Category Median	42.5	947.0	09h:35m:29s	06h:22m:08s
Per Category Average	71.9	1120.2	13h:04m:47s	08h:07m:24s

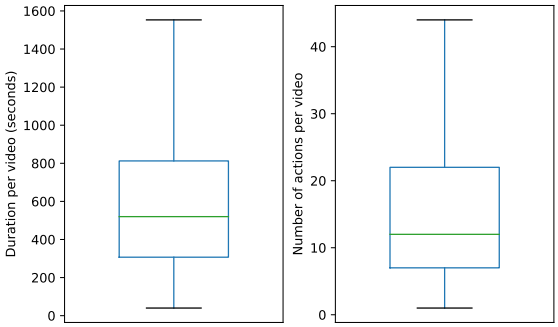
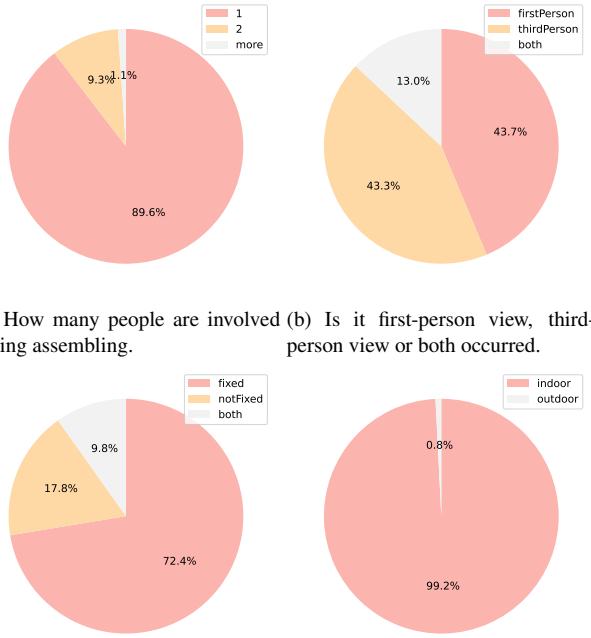


Figure 2. Distribution of video duration and the number of actions per video. Video duration: (max: 4763, min: 40, mean: 655, median: 520); Number of actions: (max: 63, min: 1, mean: 16, median: 12)

	Train	Validation	Test
#furniture	420	138	226
#videos	643	138	226
#actions	9925	2221	3537
#video segments	30876	6871	11103

Table 3. Split Statistics.

We try every possible assignments of videos in this situation, and select the one that minimises the distribution difference between the split and the entire dataset. In this way, we can ensure that there is no shared video between train and validation or test split and all manuals are fed to the model during training. The final split statistics are shown in the Tab. 3.



(a) How many people are involved during assembling.

(b) Is it first-person view, third-person view or both occurred.

(c) Is the camera fixed, not-fixed or both occurred.

(d) Is the assembly happened indoor or outdoor.

Figure 3. Proportion of the four attributes.

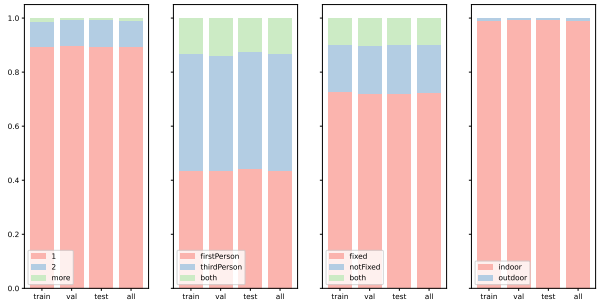


Figure 4. Balanced distribution of four attributes in each split comparing with the entire dataset (all).

We will release the dataset including video URLs and full annotations on publication of this paper. In addition, we will also make public the tool we used for annotating and the supporting infrastructures we developed for Amazon Mechanical Turker.

B. Experimental Results

In this section, we provide more experimental results both quantitatively and qualitatively to demonstrate the effectiveness of our method.

Table 4. Ablation study on SPRF. These experiments are conducted based on loss configuration B3.

Method	Video to diagram retrieval				Diagram to video retrieval					
	Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	S	P	S	P	S	P	S	P	S	P
w/o	22.28	27.70	5.983	4.639	16.97	12.95	36.36	27.25	0.548	0.357
PE [2] Add	19.10	25.72	4.317	3.248	14.49	12.71	35.04	26.93	0.544	0.356
PE [2] Concat	18.85	24.93	4.384	3.265	15.28	12.39	34.25	27.04	0.541	0.353
PRF	27.29	32.60	3.830	3.128	**21.08**	16.09	43.89	31.03	0.615	0.393
SPRF After	25.75	34.17	**3.594**	3.144	20.08	16.50	43.09	31.78	0.617	**0.394**
SPRF	**28.20**	**34.59**	3.789	**2.991**	21.02	**16.64**	**44.43**	**31.93**	**0.618**	0.393

Table 5. Ablation study for different batch sizes based on CLIP and A3 without OT.

Method	Batch Size	Video to diagram retrieval				Diagram to video retrieval					
		Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
		S	P	S	P	S	P	S	P	S	P
CLIP	64	22.59	23.10	4.011	3.976	19.74	12.07	42.43	27.43	0.611	0.386
	128	22.08	23.67	4.186	3.870	18.71	12.17	41.64	27.12	0.606	0.387
	256	19.61	19.05	4.274	4.180	16.94	10.25	38.67	23.45	0.590	0.373
A3	64	22.18	23.28	4.097	3.972	19.48	12.63	42.58	27.13	0.610	0.387
	128	21.71	22.84	3.999	3.956	19.73	12.74	40.97	27.42	0.601	0.383
	256	20.58	19.34	4.036	4.090	17.08	10.13	39.89	24.64	0.583	0.371

B.1. Quantitative Results

Sinusoidal Progress Rate Feature (SPRF). Firstly, we conducted ablation experiments on Positional Embedding (PE) proposed by [1]. The original PE is used to represent the order information for words in a sentence of arbitrary length. In our task, we need to align two different progress rates with different scales, so we manually set length for PE to be 100, and sample positional embeddings from it. We tried to replace the proposed SPRF with either adding (denoted as "PE Add") the positional embedding or concatenating (denoted as "PE Concat") to the feature. As shown in the Tab. 4, both variants of position embedding are inferior to our baseline. Secondly, we tried to modify the overall architecture so that the SPRF locates after the final linear layer and before the loss (denoted as "SPRF After"), to make the progress rate information applied directly on the contrastive loss. This modification, however, failed to outperform the proposed architecture. We conjecture that it is because the vision and progress rate features can be better fused through a fully connected layer. Besides that, we removed sinusoidal transform (denoted as "PRF") and also found the performance dropped. Converting progress rate feature into sinusoidal space is intuitive because we are using cosine similarity for distance calculation.

Batch Size. Batch size is important for contrastive learning, since the larger batch size leads to more negative samples, hence the stronger supervision signal. However, in this task, it failed to increase performance when we enlarge the batch size. We suspect it is due to the fact that there are semantic collisions or duplications (Fig. 5) because there are multiple segments for each action. As shown in our ablation study (Tab. 5), all the metrics of both CLIP and A3 drops as batch size increases. But A3 has a relatively slower rate of decline compared with CLIP. Due to the GPU memory constraints, we didn't report results for batch size >256.

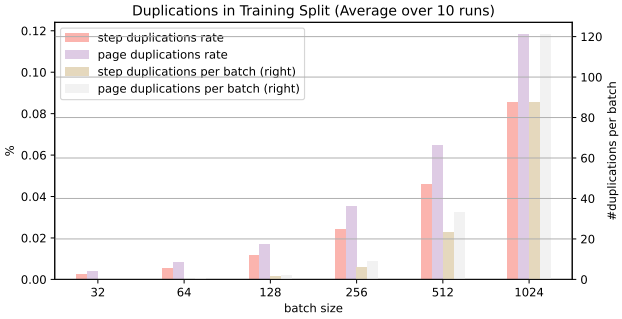


Figure 5. Duplication statistics in a batch w.r.t different batch sizes.

Post Process. We use optimal transport for post-process and the results in Tab. 6 show that OT improves performance for most cases. In particular, for the loss combina-

Table 6. Ablation study results on different loss combinations. All the results in this table are obtained **after** applying the optimal transport post-processing. Optimal transport is generally effective for most model variants.

Exp.	Loss A		Loss B		Loss C		Video to diagram retrieval				Diagram to video retrieval					
							Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	S	P	S	P	S	P	S	P	S	P	S	P	S	P	S	P
CosSim†							17.47	9.97	3.837	4.802	15.70	6.50	39.78	18.91	0.574	0.356
CLIP†							20.52	19.36	4.175	4.104	18.72	10.71	39.35	22.23	0.553	0.352
A1†	✓						20.56	14.44	4.323	4.892	18.63	8.42	39.81	18.84	0.547	0.330
A2†		✓					19.17	18.29	4.203	4.420	17.35	10.13	37.30	20.71	0.534	0.342
A3†	✓	✓					20.18	17.90	4.236	4.574	17.48	9.51	38.67	20.62	0.538	0.341
B1			✓				29.54	21.08	3.563	4.134	24.25	11.48	46.64	24.81	0.607	0.369
B2				✓			25.99	**36.74**	3.528	**2.791**	22.79	18.33	45.01	31.18	0.596	0.393
B3			✓	✓			29.74	36.40	3.605	2.880	24.22	17.89	46.71	29.96	0.598	0.389
C1	✓		✓				29.29	19.61	3.754	4.402	23.59	10.82	45.77	23.43	0.590	0.355
C2		✓		✓			25.67	36.22	3.588	2.890	22.11	18.12	44.67	30.30	0.589	0.393
C3	✓	✓	✓	✓			30.37	35.49	3.606	3.022	24.04	18.02	46.31	29.44	0.593	0.389
D1			✓	✓	✓	✓	**31.61**	36.71	**3.458**	2.816	**26.62**	18.28	**49.11**	**32.28**	**0.626**	**0.401**
D2	✓	✓	✓	✓	✓	✓	30.66	36.12	3.539	2.939	25.31	**18.44**	48.86	31.32	0.620	0.396

tion D1, it outperforms the one without OT (Tab. 2 in the main paper) by almost 3%. And it is worth noting that compared with B3, Loss C plays a positive role in terms of the performance.

B.2. Qualitative Results

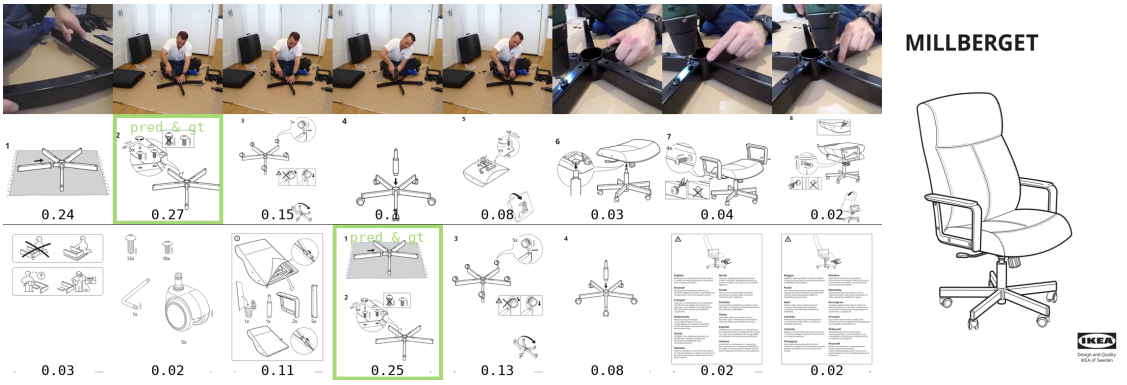
In this section, we show sixteen selected examples: eight for successful alignments and another eight for failure cases.

Moreover, three video examples are attached in the zip file demonstrating the result of our work. We encourage readers to watch the attached videos for a better understanding of our method and dataset.

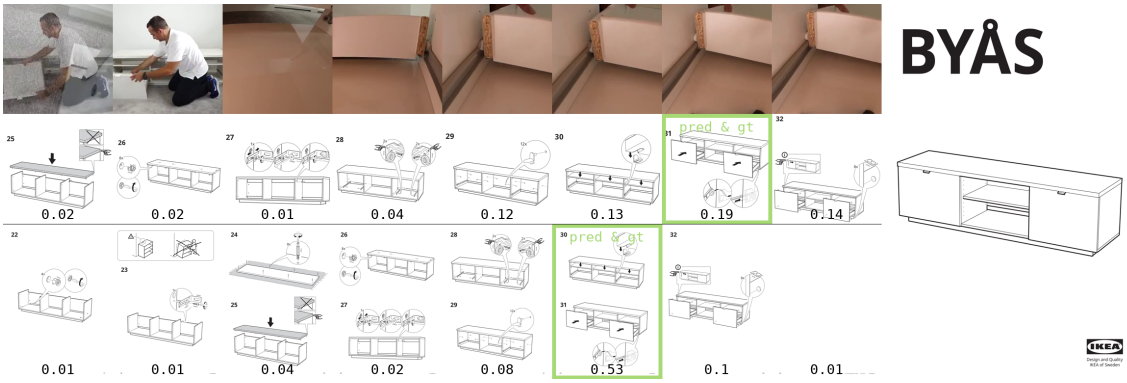
References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3

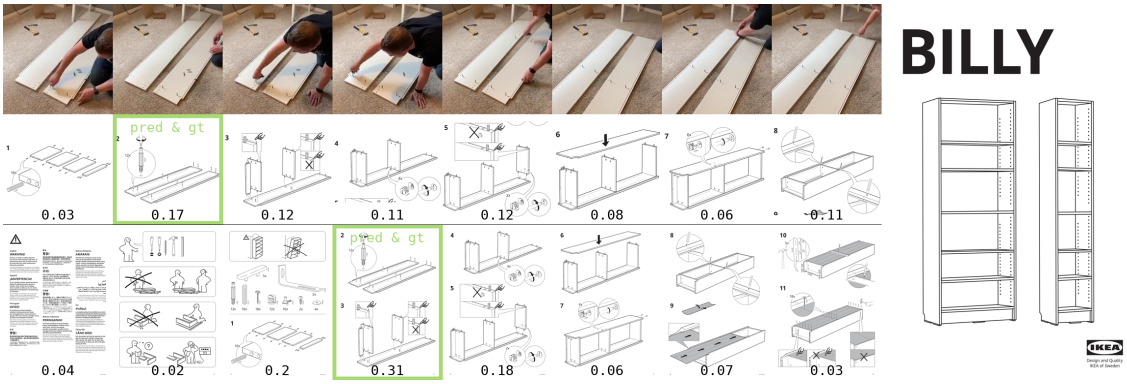
[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3



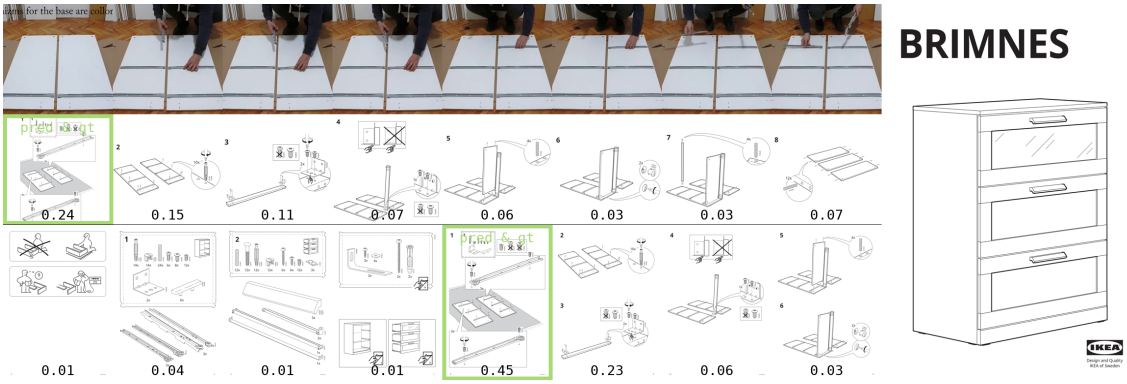
(a) Alignment between a clip from YouTube video 36T-ytb8EhM and an Ikea furniture manual 00339416.



(b) Alignment between a clip from YouTube video aPJF36W93wE and an Ikea furniture manual 00352573.

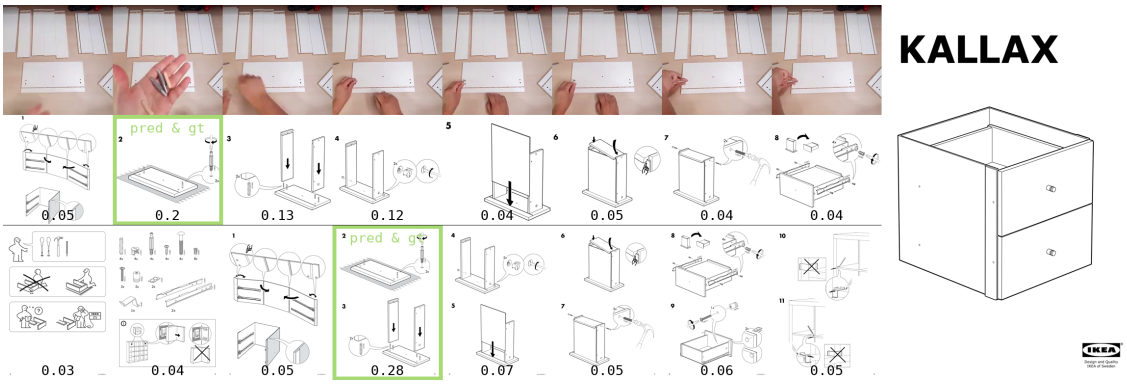


(c) Alignment between a clip from YouTube video Lji-ZnRLPBQ and an Ikea furniture manual 10351568.

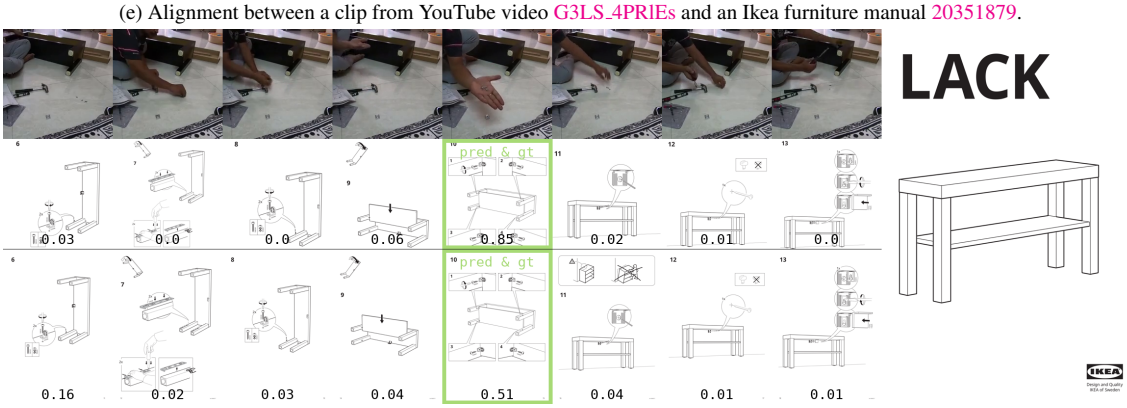


(d) Alignment between a clip from YouTube video n7ZlhsVKnuY and an Ikea furniture manual 10355415.

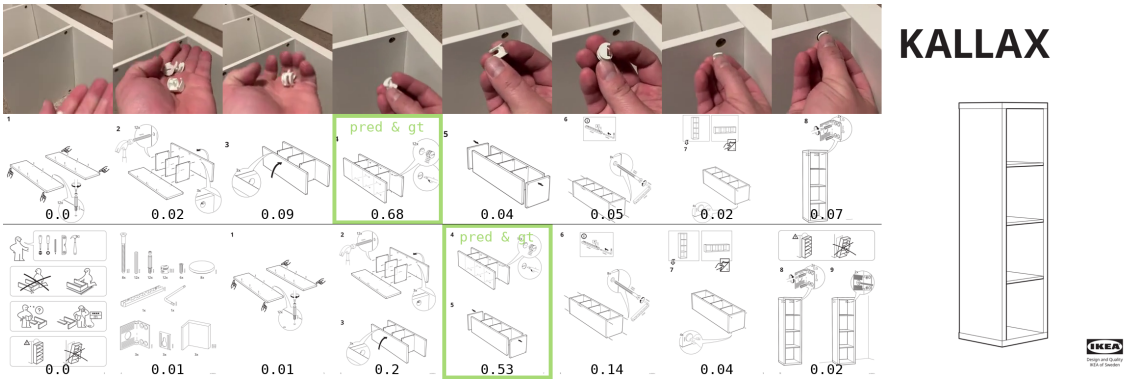
Figure 6. Eight success examples (4/8).



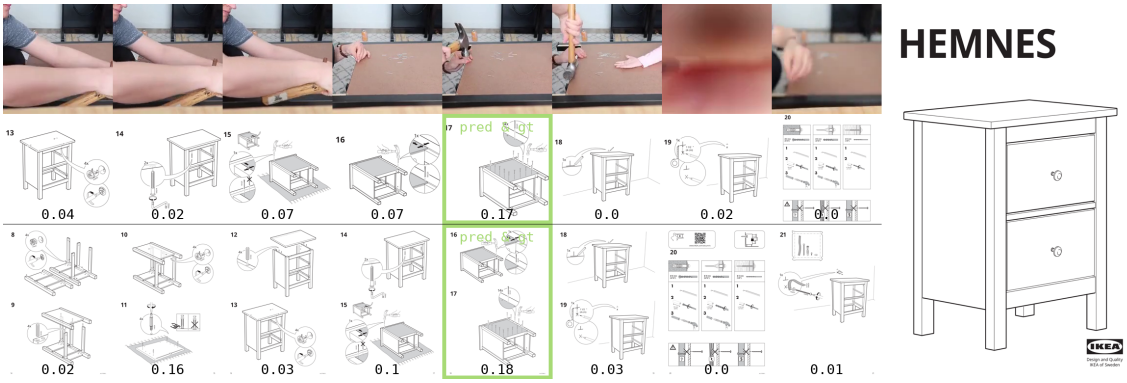
(e) Alignment between a clip from YouTube video G3LS_4PRlEs and an Ikea furniture manual 20351879.



(f) Alignment between a clip from YouTube video mr92aJSM2hM and an Ikea furniture manual 30353566.

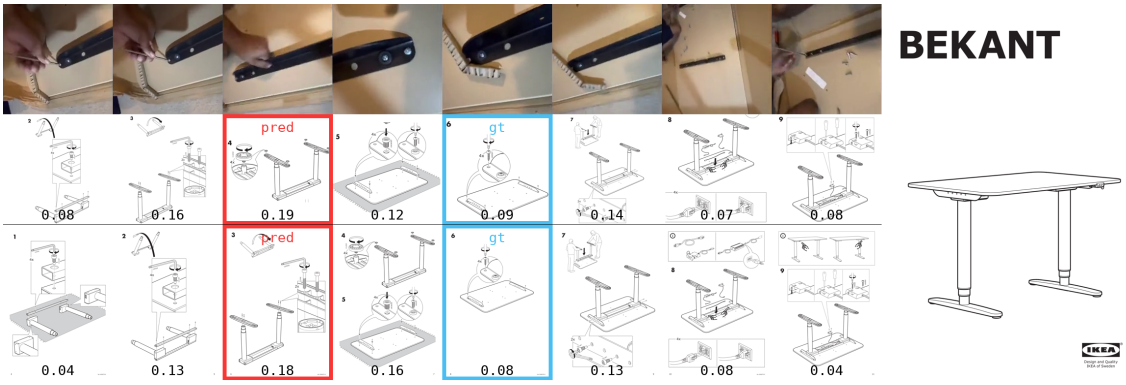


(g) Alignment between a clip from YouTube video jthbbffqPlg and an Ikea furniture manual 40351883.

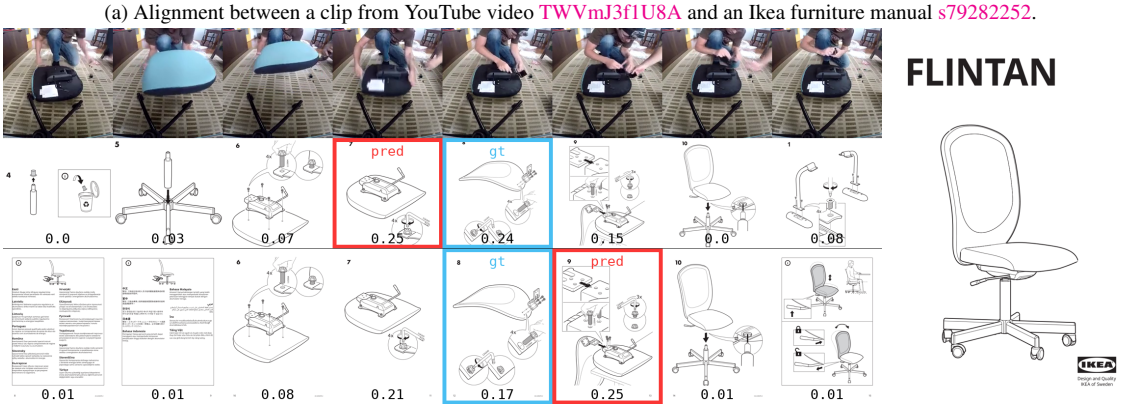


(h) Alignment between a clip from YouTube video n7ZlhsVKnuY and an Ikea furniture manual 50355687.

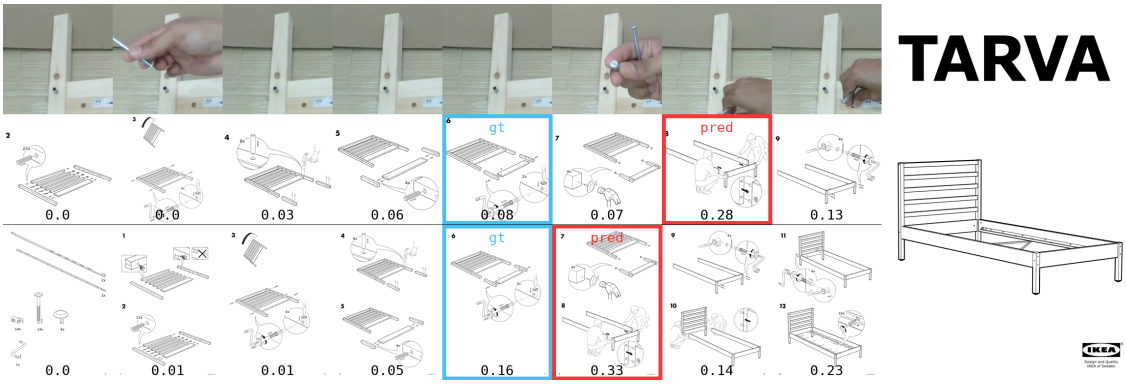
Figure 6. Eight success examples (8/8).



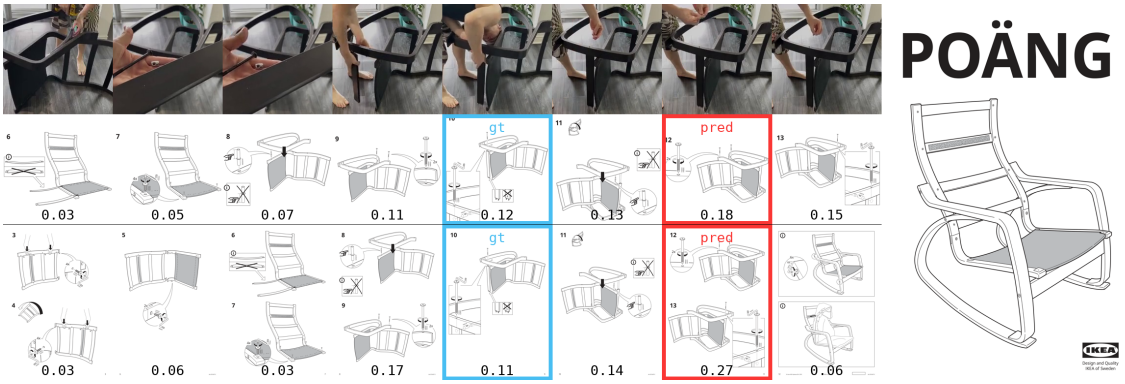
(a) Alignment between a clip from YouTube video TWVmJ3f1U8A and an Ikea furniture manual s79282252.



(b) Alignment between a clip from YouTube video CM-n5RaTtOw and an Ikea furniture manual s69424469.

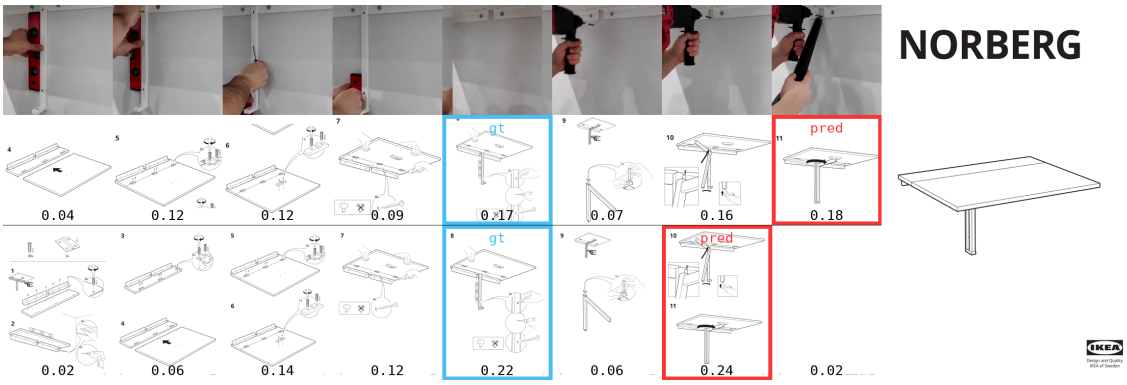


(c) Alignment between a clip from YouTube video u-2oTXnp48c and an Ikea furniture manual s09009572.

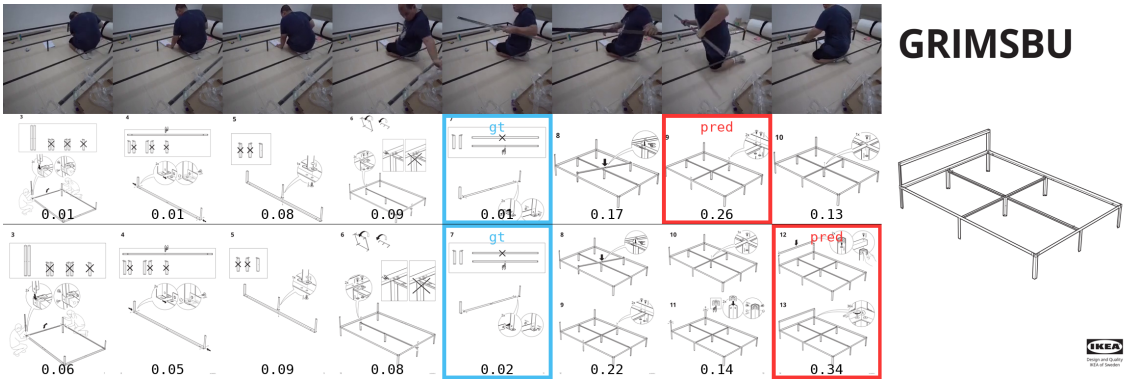


(d) Alignment between a clip from YouTube video NXNxUE6XsF4 and an Ikea furniture manual s19395841.

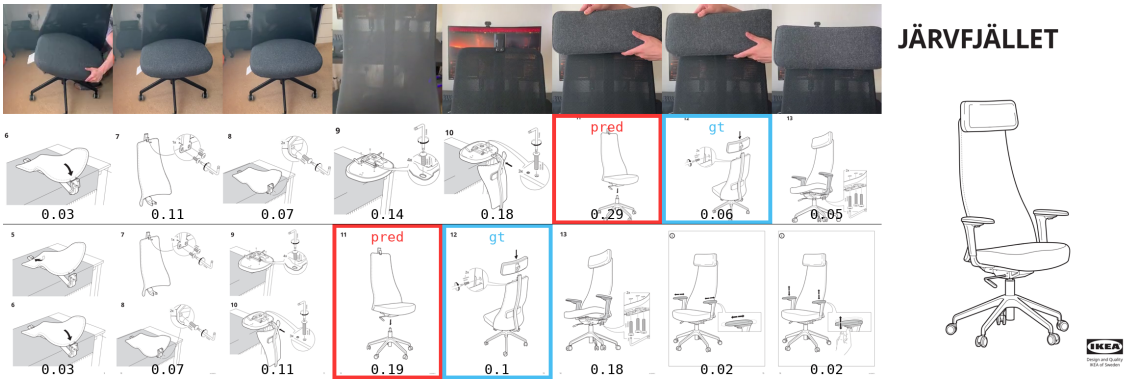
Figure 7. Eight failure cases (4/8).



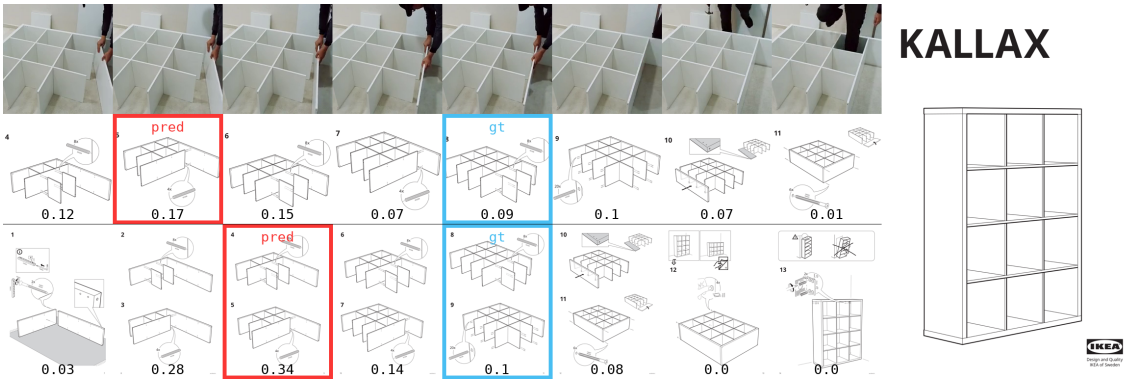
(e) Alignment between a clip from YouTube video IUI8AQYTB0k and an Ikea furniture manual 90365793.



(f) Alignment between a clip from YouTube video hFA07VckvOk and an Ikea furniture manual 80458759.



(g) Alignment between a clip from YouTube video DSD9jawiz7o and an Ikea furniture manual 80494540.



(h) Alignment between a clip from YouTube video T6sQN6FtaDQ and an Ikea furniture manual 60409939.

Figure 7. Eight failure cases (8/8).