Aligning Step-by-Step Instructional Diagrams to Video Demonstrations

Jiahao Zhang[1,*] Anoop Cherian[2] Yanbin Liu[1] Yizhak Ben-Shabat[1,3,†] Cristian Rodriguez[4] Stephen Gould[1,‡]

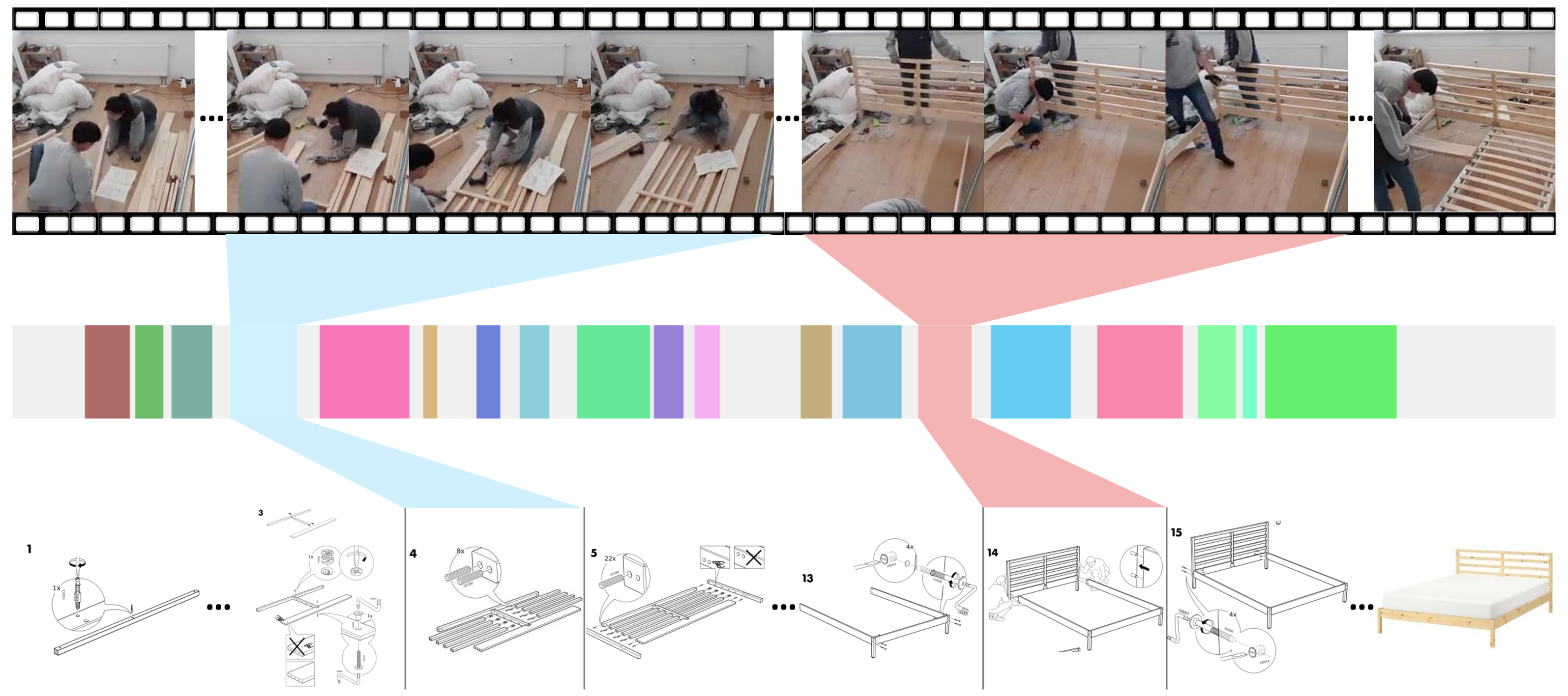
[1]The Australian National University [2]Mitsubishi Electric Research Labs [3]Technion Israel Institute of Technology [4]The Australian Institute for Machine Learning

[1]{first.last}@anu.edu.au [2]cherian@merl.com [3]sitzikbs@gmail.com [4]crodriguezop@gmail.com

Project Website

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

1. Introduction



▶ Problem Definition

Given a **video** sequence of a human demonstrating a furniture assembly (e.g., a DIY video) and also given a sequence of instruction **diagrams** pictorially demonstrating the assembly steps (as is common in Ikea instruction manuals), we consider the problem of **aligning** the instruction diagrams and the temporal locations of the corresponding human actions in the video.

▶ Two-way Retrieval Task

Video-to-Diagram retrieval: $j^\star = \operatorname*{argmax}_{j=1,\ldots,M} f_{sim}(\mathbf{f}^V, \mathbf{f}^I_j)$

Diagram-to-Video retrieval: $i^\star = \operatorname*{argmax}_{i=1,\ldots,N} f_{sim}(\mathbf{f}^V_i, \mathbf{f}^I)$

▶ Motivation

1. Help assemblers to locate steps in online instructional videos.
2. **A picture is worth a thousand words**, which better describes assembly.
3. Most DIY assembly videos do **NOT** have subtitles nor narratives and manually labeled language description can be ambiguous.

2. Contributions

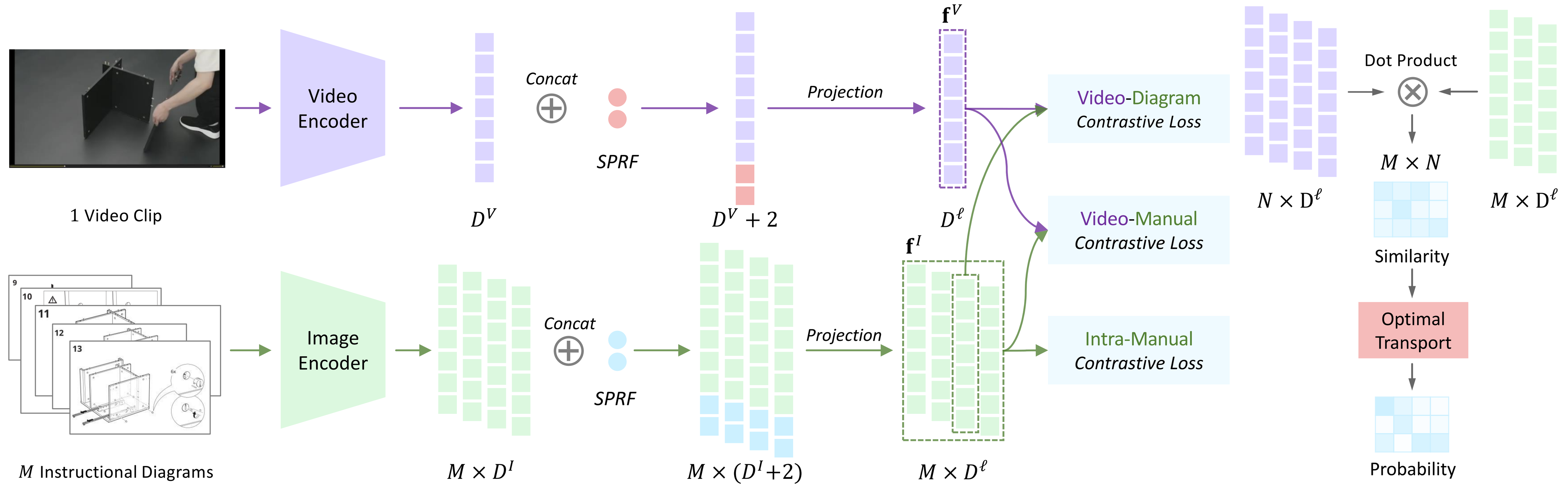
- A **novel task** of multimodal alignment between instruction videos and abstract diagrams of assembly steps.
- **Three new losses** to take into account the many-to-one mapping of video clips to images, **prior knowledge** of the assembly task, and the usage of **optimal transport** as post-processing.
- We introduce an annotated high-quality **dataset (Ikea Assembly in the Wild)** for studying our retrieval and alignment tasks.

Acknowledgements

* 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.

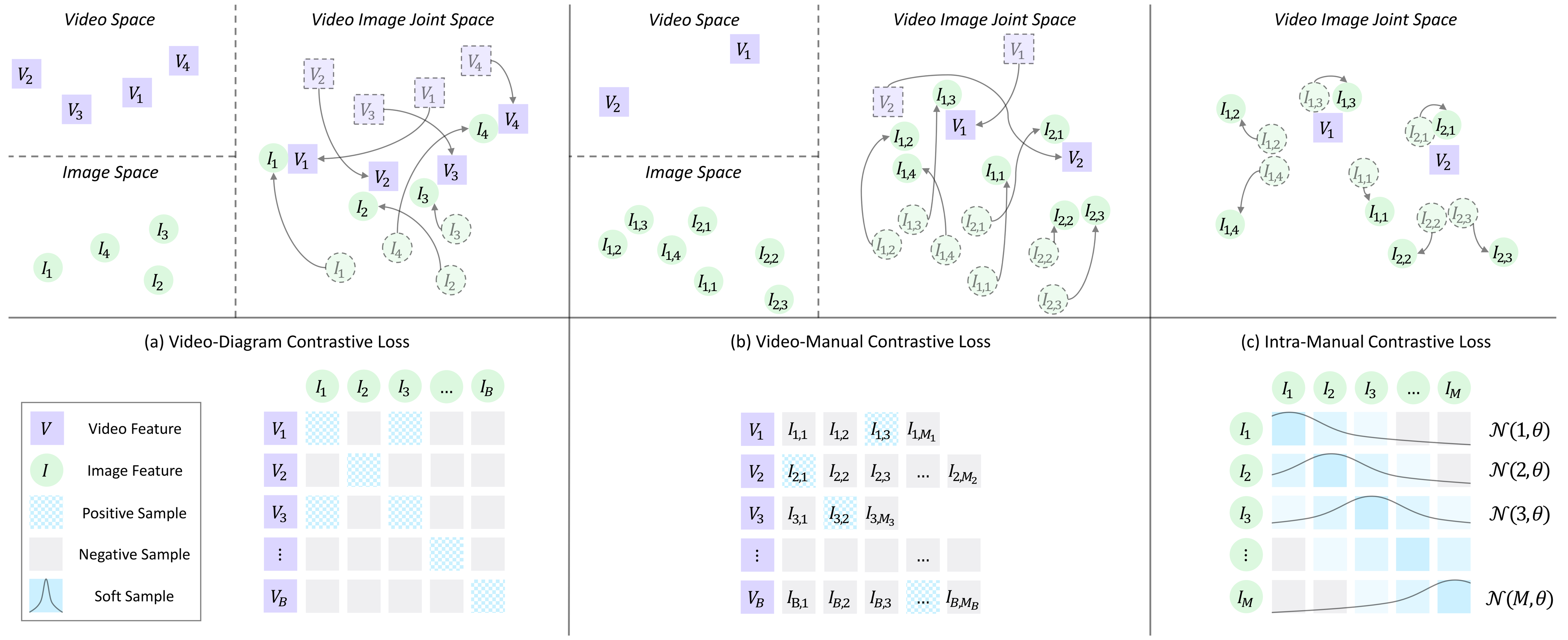
3. Methods - Contrastive Learning Based Video & Instruction Diagram Alignment

▶ Overall Pipeline



(a) Training Stage. (b) Inference Stage.

▶ Three Task-Specific Designed Contrastive Losses



(a) Video-Diagram Contrastive Loss (b) Video-Manual Contrastive Loss (c) Intra-Manual Contrastive Loss

▶ Sinusoidal Progress Rate Feature (SPRF)

- There is a positive correlation between the progress of video and step index.

$$r^V = \frac{t_{start} + t_{end}}{2 t_{duration}}, \mathrm{SPRF}^V = (\sin(\pi r^V), \cos(\pi r^V)); r^I = \frac{j}{M}, \mathrm{SPRF}^I = (\sin(\pi r^I), \cos(\pi r^I))$$

▶ Optimal Transport (OT) for Post-Processing

1. Calculate the cost matrix.

$\mathbf{f}^V$: Video Feature
$\mathbf{f}^I$: Diagram Feature
f_{sim}: Similarity function
s: Similarity Matrix
c: Cost Matrix
α: Accentuation Factor

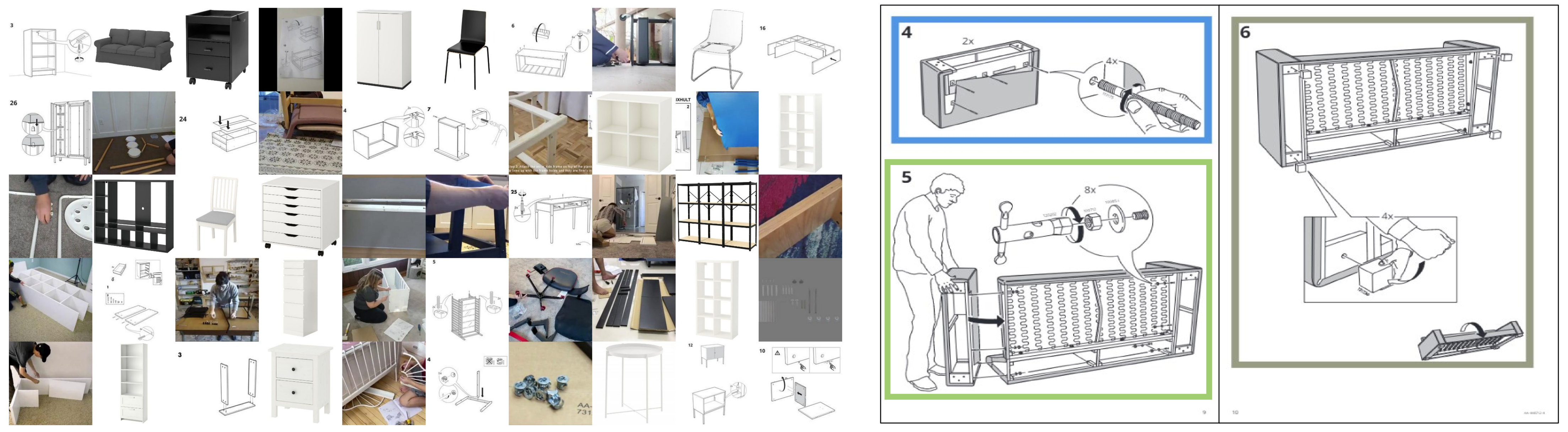
$$s_{ij} = f_{sim}(\mathbf{f}^V_i, \mathbf{f}^I_j)$$
$$\bar{s} = \max_{ij} s_{ij}$$
$$\underline{s} = \min_{ij} s_{ij}$$
$$c_{ij} = \frac{s_{ij}^\alpha - \underline{s}^\alpha}{\bar{s}^\alpha - \underline{s}^\alpha}$$

2. Entropy regularized OT problem.

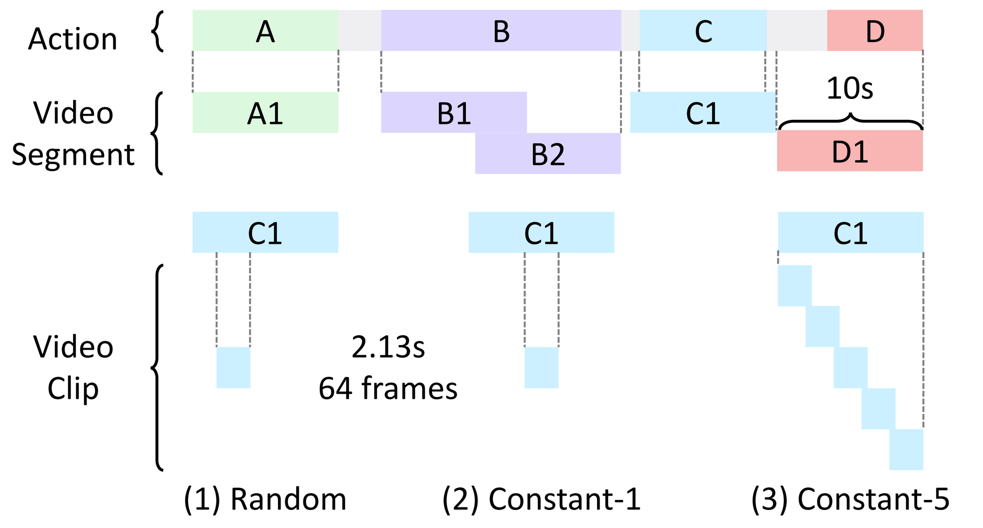
$$\text{minimize} \sum_{i=1}^{N} \sum_{j=1}^{M} T_{ij} c_{ij} - \epsilon H(T)$$
$$\text{subject to} \sum_{i=1}^{M} T_{ij} = \frac{1}{N}, for\ j = 1, \ldots, N$$
$$\sum_{j=1}^{N} T_{ij} = \frac{1}{M}, for\ i = 1, \ldots, M$$

Using Sinkhorn-Knopp algorithm to get the optimal transport plan $T^\star$, which is regarded as the final alignment probability.

4. Dataset – Ikea Assembly in the Wild (IAW)



- **420** Ikea furniture from 14 categories.
- **1005** YouTube videos, each with 4 extra attributes.
- $\approx$ **183** hours in total, $\approx$ **11** min in average.
- **461** Ikea furniture assembly manuals.
- **8263** manually-cropped assembly step diagrams.
- **15649** pairs of aligned video clips and steps.
- $\approx$ **114** hours of video ($\approx$ 61%) are aligned.
- Powered by Amazon Mechanical Turk and Vidat.



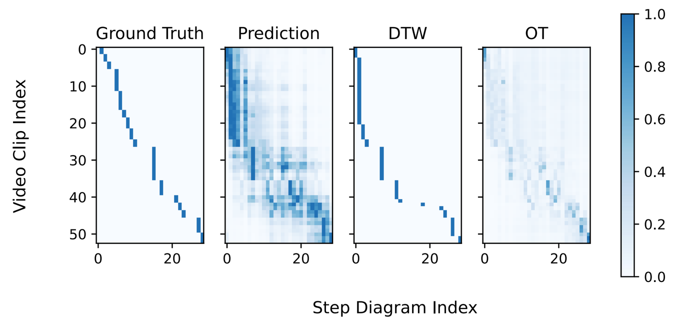
Video clip sampler.

5. Results

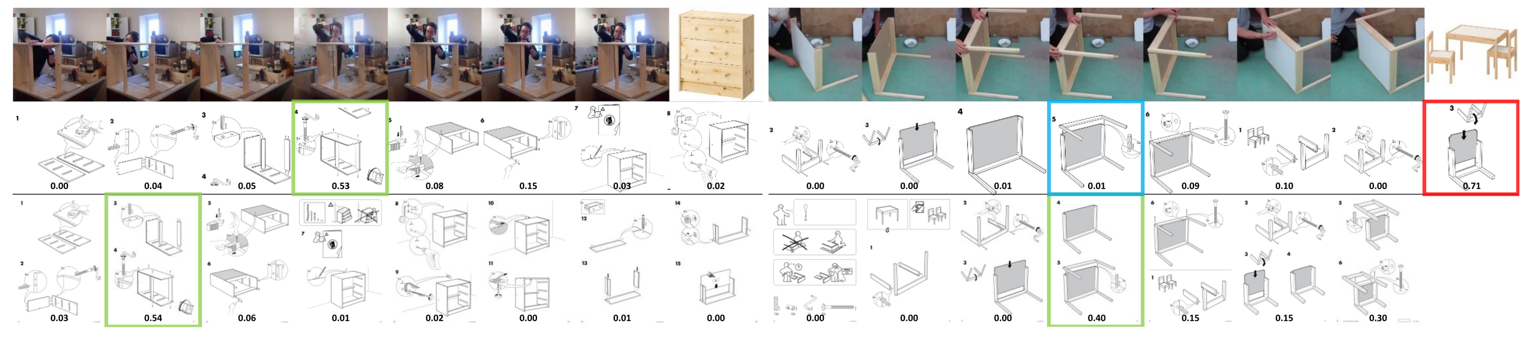
▶ Quantitative

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
Random	5.664	5.107	9.334	8.131	6.576	3.393	19.90	10.16	0.375	0.244
CosSim	11.89	11.06	4.360	4.368	12.43	6.780	32.90	20.93	0.561	0.336
CLIP	19.61	19.05	4.274	4.180	16.94	10.25	38.67	23.45	0.590	0.373
Ours	28.62	34.55	3.734	2.928	22.30	16.48	45.00	32.20	0.617	0.390†
w/o SPRF	21.73	27.08	6.018	4.485	16.90	13.17	36.07	26.70	0.558	0.357
w/ DTW	31.45	36.20	**3.382**	**2.752**	23.20	17.32	32.45	17.55	0.467	0.310
w/ OT	**31.61**	**36.71**	3.458	2.816	**26.62**	**18.28**	**49.11**	**32.28**	**0.626**	**0.401**

▶ Qualitative



Comparing DTW and OT.



Success example. Failure example.