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

Project: https://academic.davidz.cn/en/publication/zhang-cvpr-2023/

Dataset: https://iaw.davidz.cn/

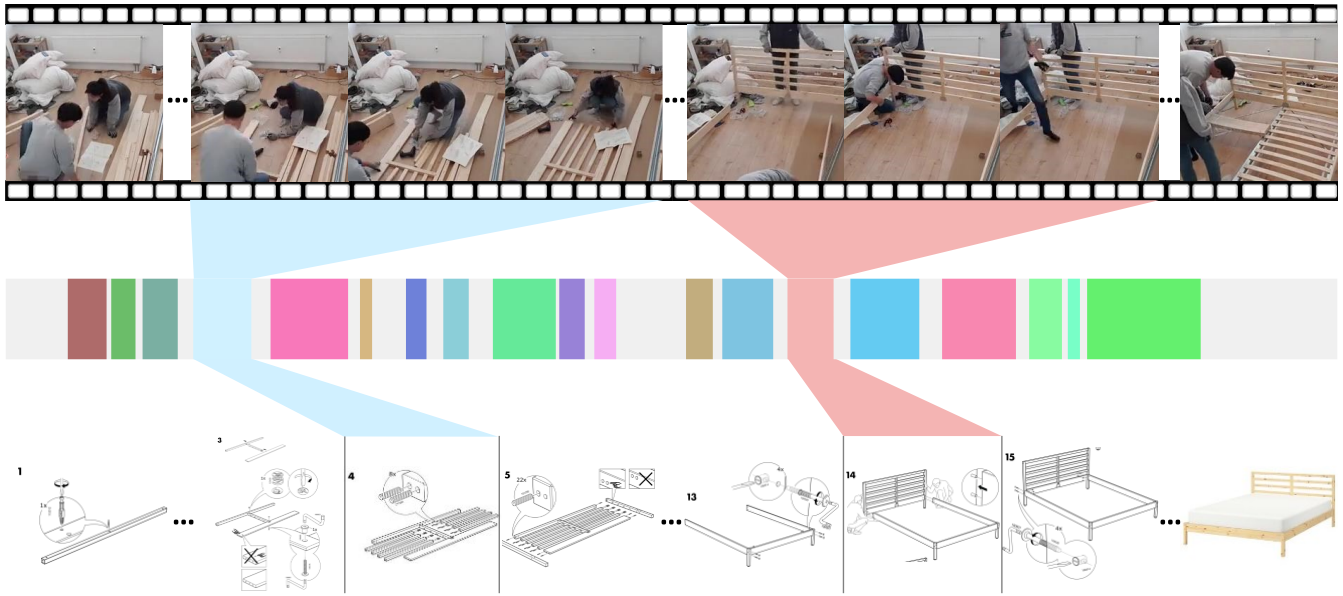
Poster: TUE-AM-237

Project

Dataset

* 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.

Overview - Task



N clips $\{V_i\}_{i=1}^{N}$ in a video

M instructional diagrams $\{I_j\}_{j=1}^{M}$ in a manual

Figure 1. An illustration of video-diagram alignment between a YouTube video (top) He0pCeCTJQM and an Ikea furniture manual (bottom) s49069795.

Overview - Task

Aligning N clips and M diagrams can be interpreted as a two-way retrieval task:

1. Video-to-Diagram retrieval, given a clip V, find the $j^\star$-th diagram $s.t.$

$$j^\star = \underset{j=1,\ldots,M}{\operatorname{argmax}} f_{sim}(\mathbf{f}^V, \mathbf{f}_j^I)$$

2. Diagram-to-Video retrieval, given a diagram I, find the $i^\star$-th video clip $s.t.$

$$i^\star = \underset{i=1,\ldots,N}{\operatorname{argmax}} f_{sim}(\mathbf{f}_i^V, \mathbf{f}^I)$$

f_{sim}: (cosine) similarity function
$\mathbf{f}^V$: video clip feature
$\mathbf{f}^I$: Image (diagram) feature

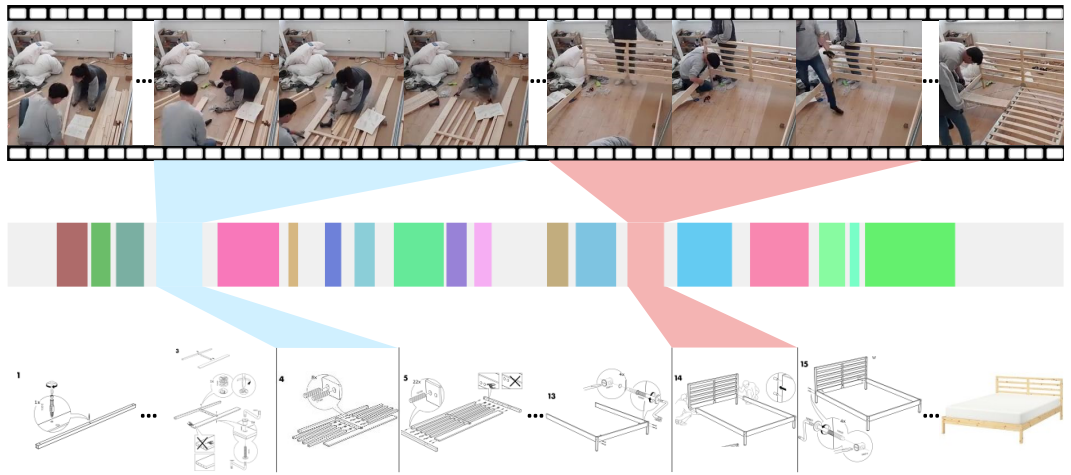


Figure 1. An illustration of video-diagram alignment between a YouTube video (top) He0pCeCTJQM and an Ikea furniture manual (bottom) s49069795.

Overview - Dataset

Figure 2. Website background of Ikea Assembly in the Wild Dataset: https://iaw.davidz.cn

Ikea Assembly in the Wild (IAW) Dataset

- **420** Ikea furniture
- **1005** videos
- **8263** assembly step diagrams
- **15649** pairs of video clips and steps

Overview - Method

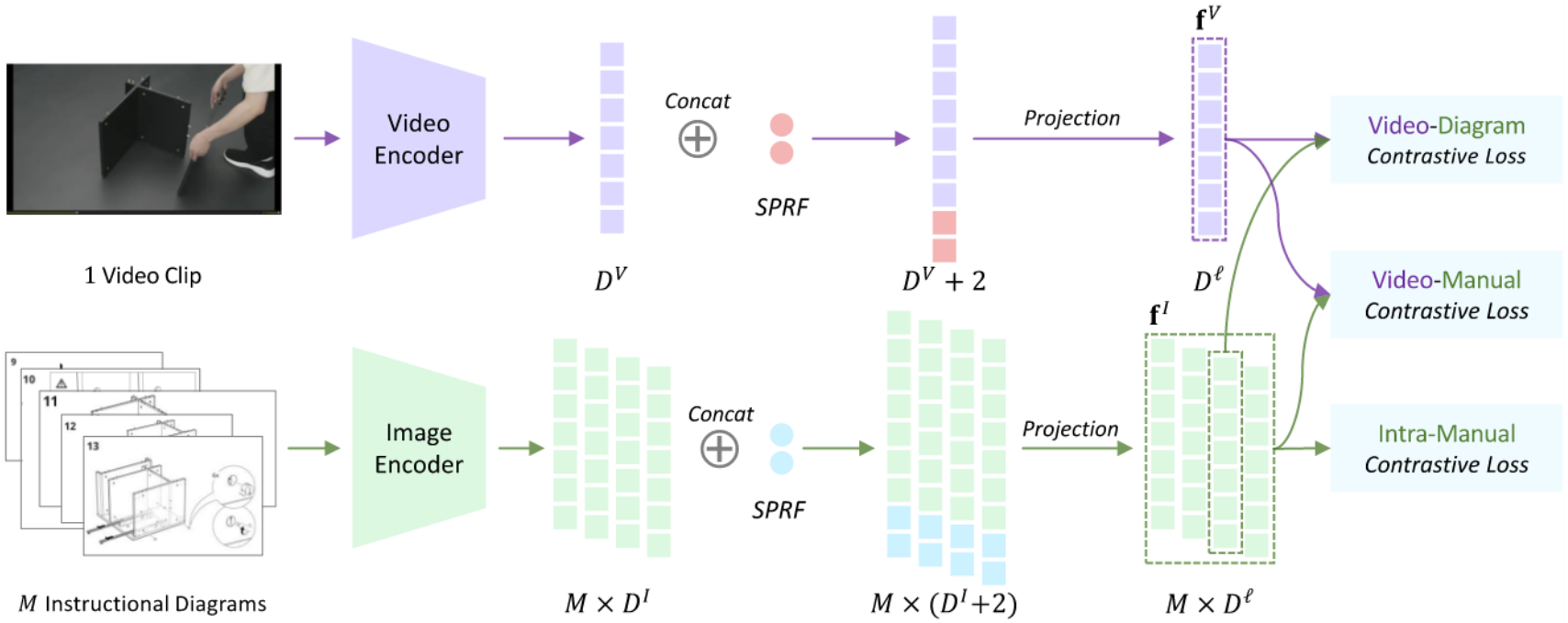


Figure 15(a). Training Stage.

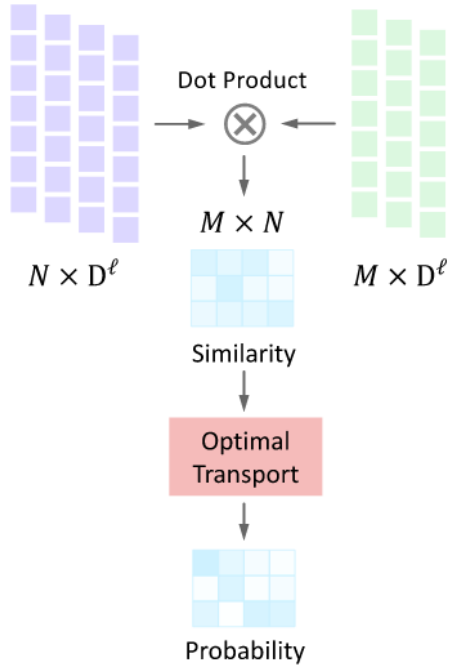


Figure 15(b). Inference Stage.

Motivation

- From application view
  - Help assembler to locate step in online videos.
  - Automatically show corresponding manual page by having a camera in front of a person who is assembling the furniture.

Motivation

- From application view
  - Help assembler to locate step in online videos.
  - Automatically show corresponding manual page by having a camera in front of a person who is assembling the furniture.
- From research view
  - Language video alignment (Temporal Sentence Grounding in Videos)
    - Language can be ambiguous (a picture is worth a thousand words).
    - Most assembly videos do not have narratives during assembling.
  - Dataset
    - No alignment between assembly videos and corresponding manuals.

Ikea Assembly in the Wild (IAW) Dataset

Figure 2. Website background of Ikea Assembly in the Wild Dataset: https://iaw.davidz.cn

- **420** Ikea furniture
  - From Ikea official website
  - 14 categories
    - Sofa
    - Bed
    - Wardrobe
    - Table
    - etc.
- **1005** videos
  - From YouTube
  - Total duration ≈ **183** hours
  - Average duration ≈ 11 min (min 1, max 74)
  - 4 extra attributes
    - First- or third-person view
    - Moving camera or not
    - Indoor or outdoor
    - Number of assemblers

Ikea Assembly in the Wild (IAW) Dataset

- **8263** assembly step diagrams
  - manually cropped from 461 manuals
  - 8568 manual page images

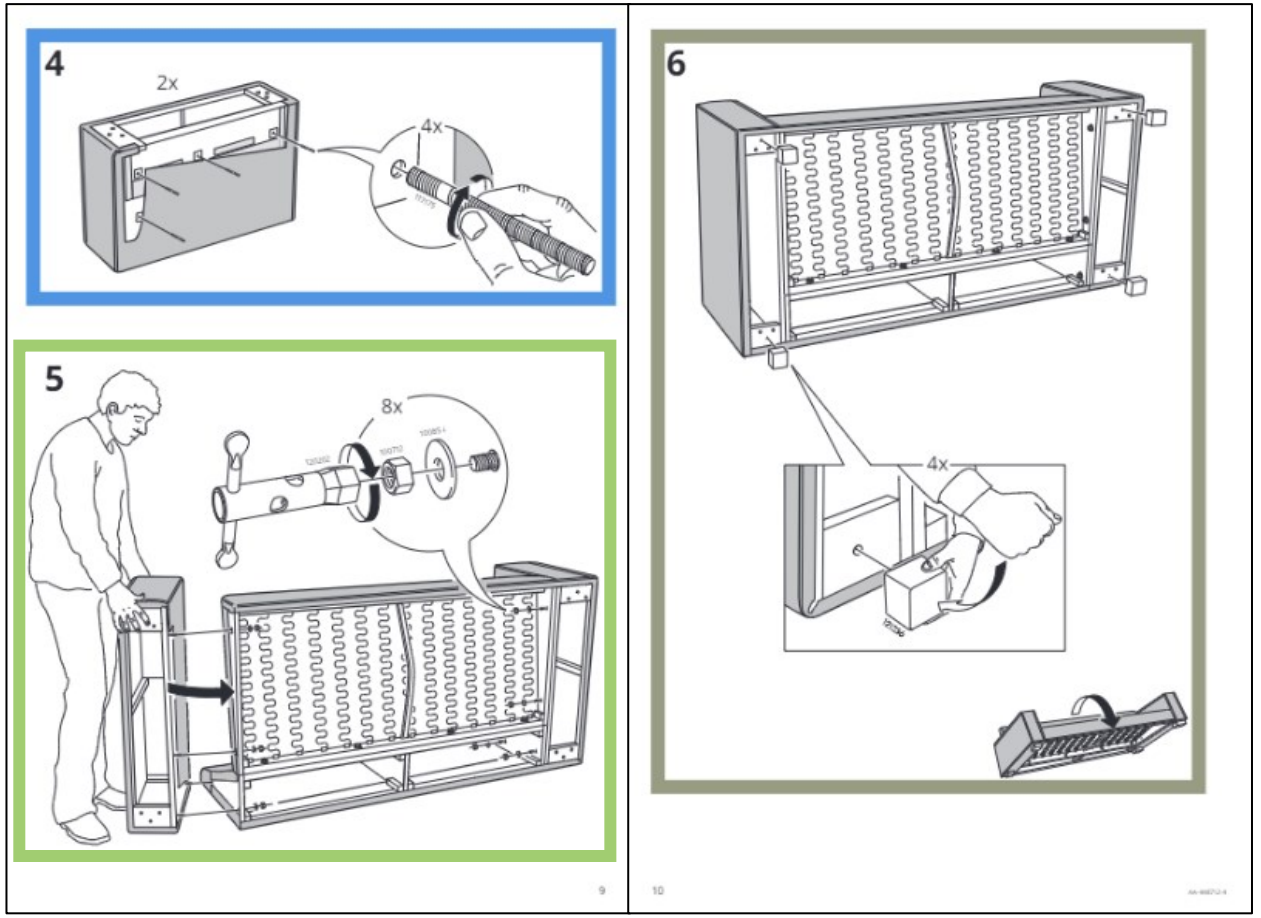


Figure 3. Ikea furniture s79011430, manual page 9 and 10.

Ikea Assembly in the Wild (IAW) Dataset

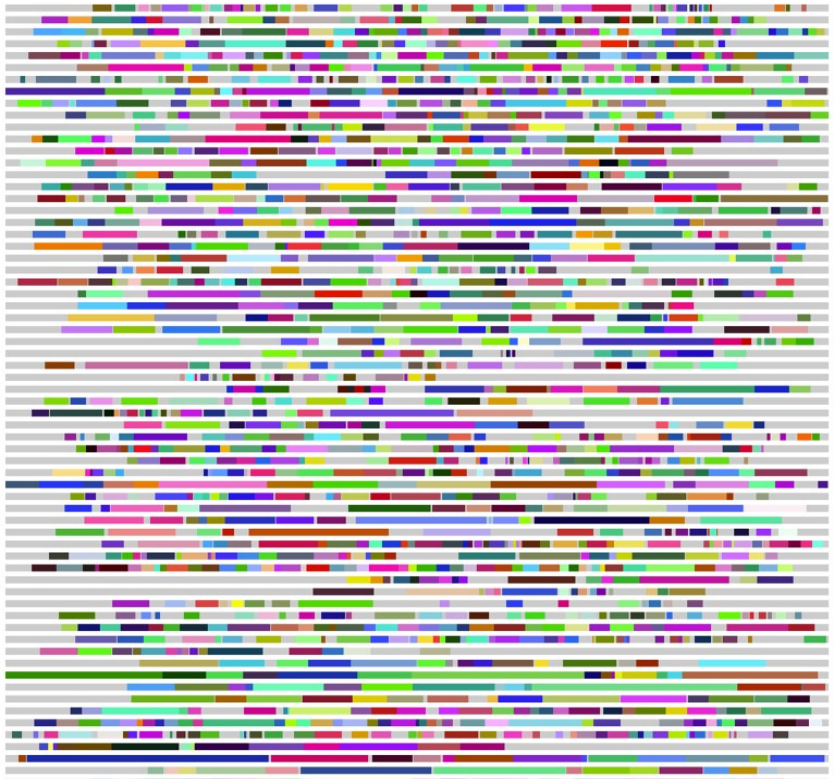


Figure 4. Visualization of video segment annotations (partial). Each color strip corresponds to a step diagram.

- **15649** pairs of video clips and steps
  - ≈ **114** hours of video (≈ 61%)
  - Amazon Mechanical Turk
  - ANU CVML Video Annotation Tool (Vidat)

Method – CLIP

Figure 8. Visualization of CLIP.

Method – CLIP

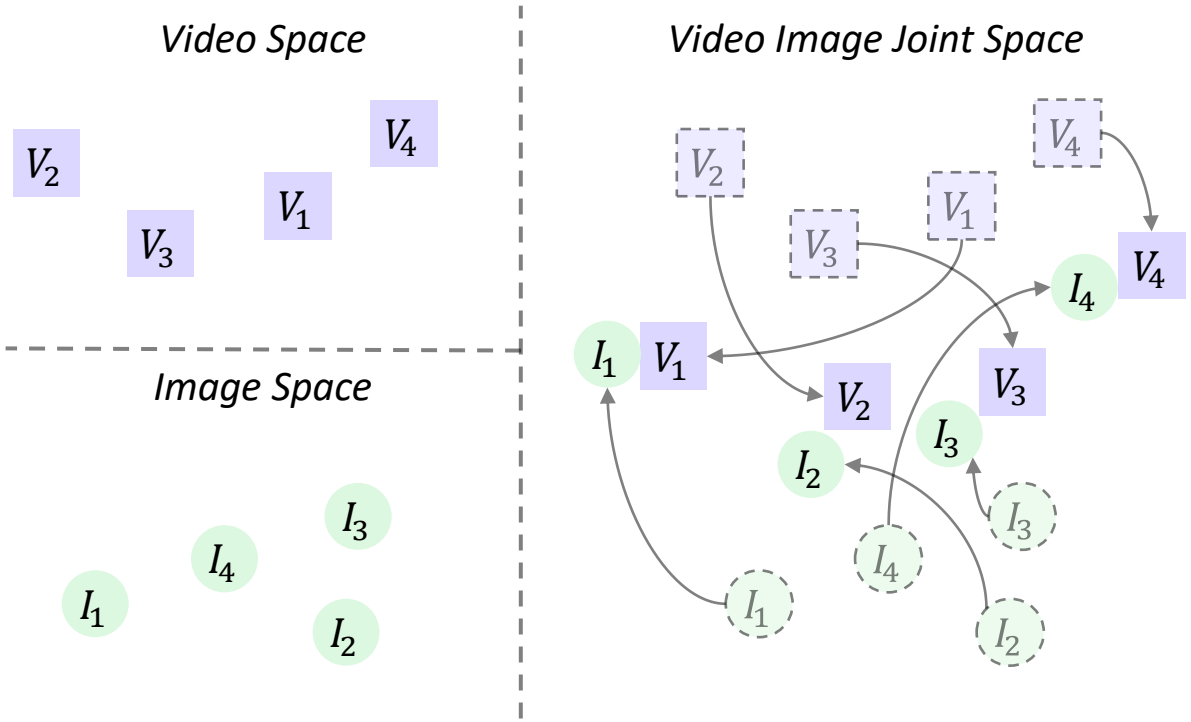


Figure 8. Visualization of CLIP.

Method – CLIP

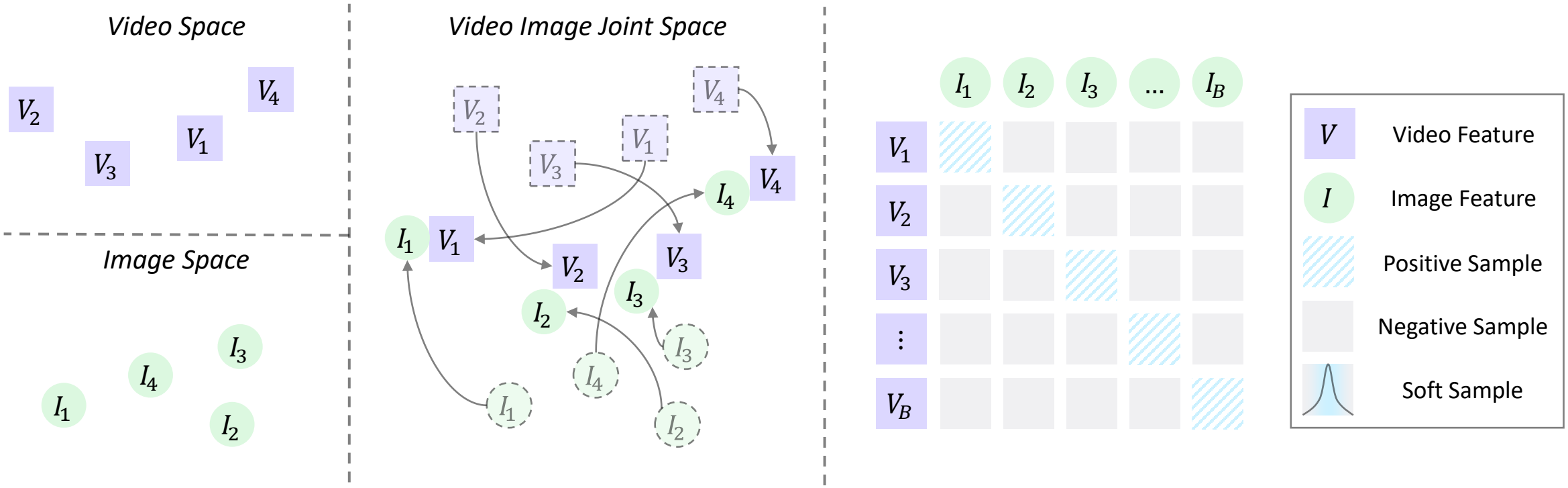


Figure 8. Visualization of CLIP.

Method – Issue: Collision

Note: CLIP uses a batch size of **32,768** pairs.

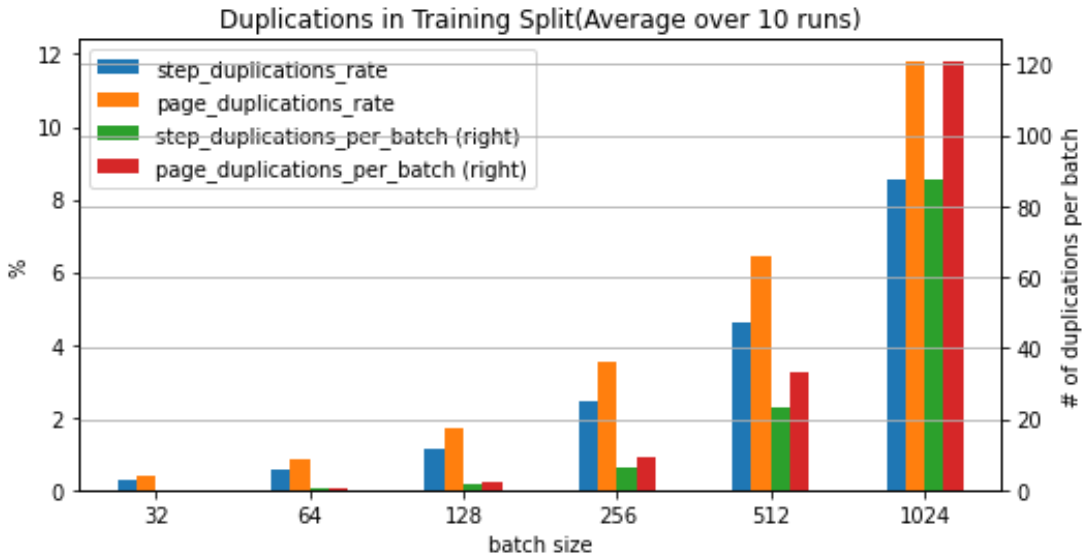


Figure 9. Illustration of collision (duplication) issue with different batch sizes.

Method – Video-Diagram Contrastive Loss

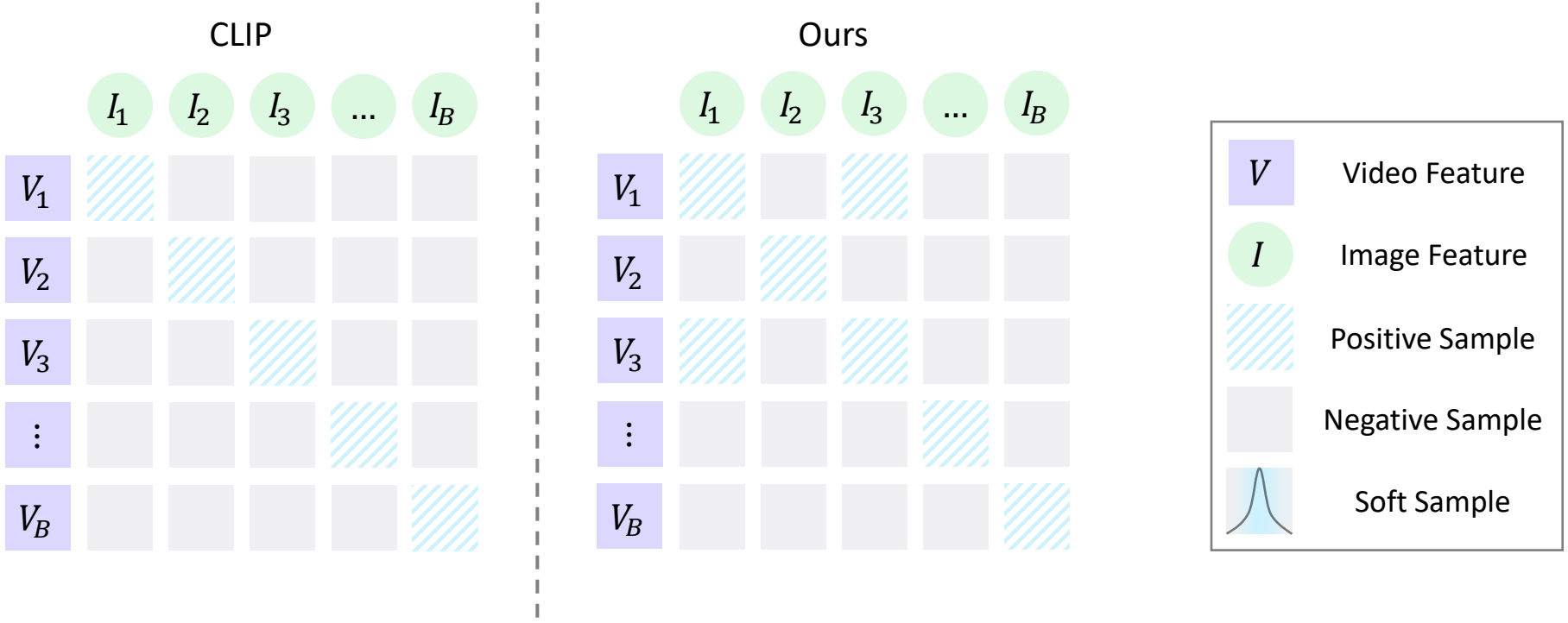


Figure 10. Comparing CLIP with loss A: Video-Diagram Contrastive Loss.

Characteristics:

- Using Jensen-Shannon (JS) divergence as the loss function instead of the InfoNCE to support multilabel multiclass classification.

Method – Video-Diagram Contrastive Loss

Table 1. Performance of difference losses.

Loss	Video to diagram retrieval				Diagram to video retrieval					
	Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	Step	Page	Step	Page	Step	Page	Step	Page	Step	Page
CLIP	19.61	19.05	4.274	4.180	16.94	**10.25**	38.67	23.45	**0.590**	**0.373**
VD	**20.58**	**19.34**	**4.036**	**4.090**	**17.08**	10.13	**39.89**	**24.64**	0.583	0.371

Method – Video-~~Diagram~~ Manual Contrastive Loss

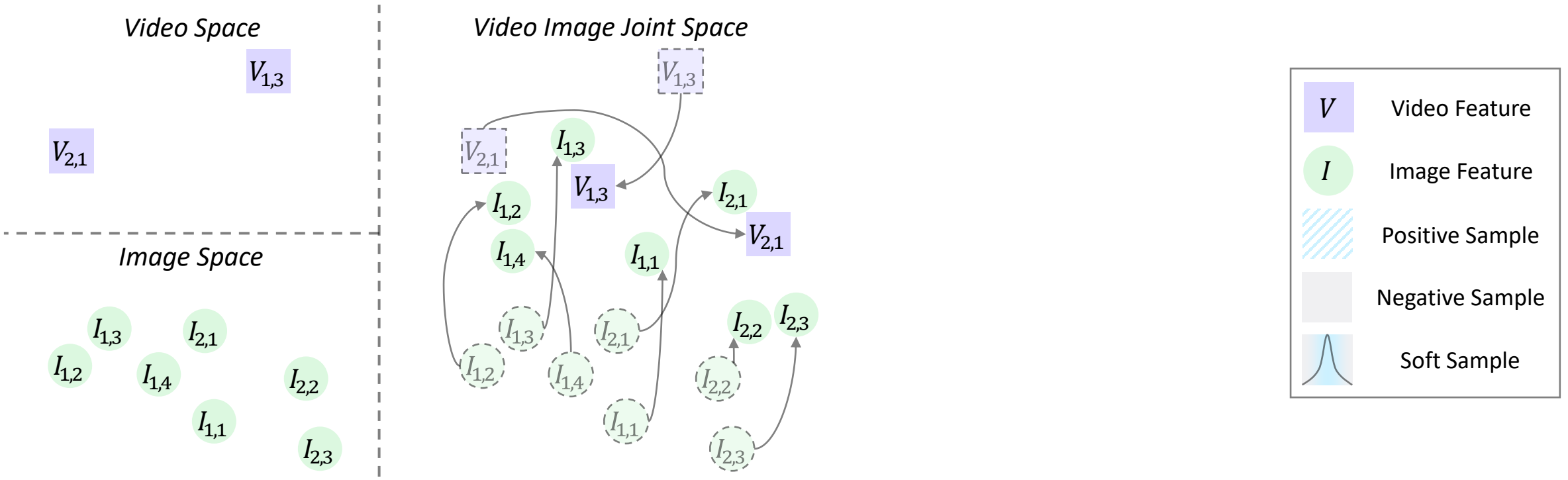


Figure 11. Visualization of loss B: Video-Manual Contrastive Loss.

Characteristics:

- Use an important prior that we already know the correspondence between video and manuals, no need to align video and diagram pairs globally across the entire dataset.
- Use the Cross Entropy loss.

Method – Video-~~Diagram~~ Manual Contrastive Loss

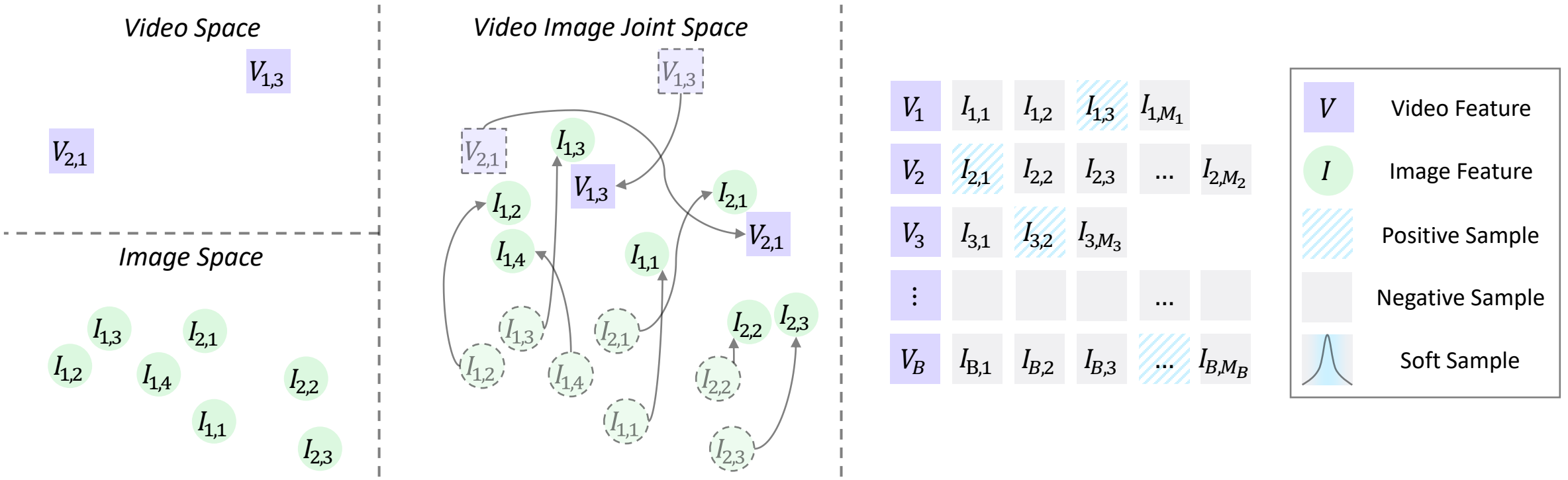


Figure 11. Visualization of loss B: Video-Manual Contrastive Loss.

Characteristics:

- Use an important prior that we already know the correspondence between video and manuals, no need to align video and diagram pairs globally across the entire dataset.
- Use the Cross Entropy loss.

Method – Video-~~Diagram~~ Manual Contrastive Loss

Table 2. Performance of difference losses.

Loss	Video to diagram retrieval				Diagram to video retrieval					
	Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	Step	Page	Step	Page	Step	Page	Step	Page	Step	Page
VD	20.58	19.34	4.036	4.090	17.08	10.13	39.89	24.64	0.583	0.371
VM	**28.20**	**34.59**	**3.789**	**2.991**	**21.02**	**16.64**	**44.43**	**31.93**	**0.618**	**0.393**

Method – Intra-Manual Contrastive Loss

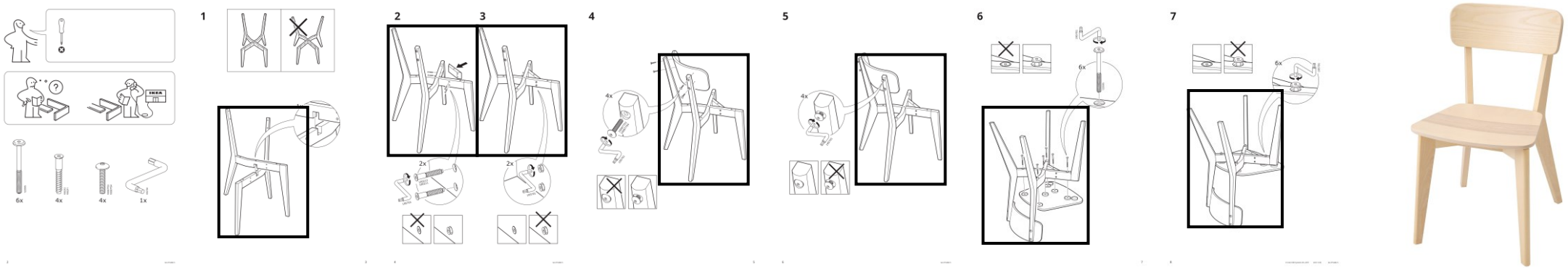


Figure 12. Manual of Ikea furniture [80457236].

Method – Intra-Manual Contrastive Loss

Figure 13. Visualization of loss c: Intra-Manual Contrastive Loss.

Characteristics:

- Enhance the difference between steps within a manual.
- Use a gaussian distribution to model the difference.
- Use the JS divergence loss.

Method – Intra-Manual Contrastive Loss

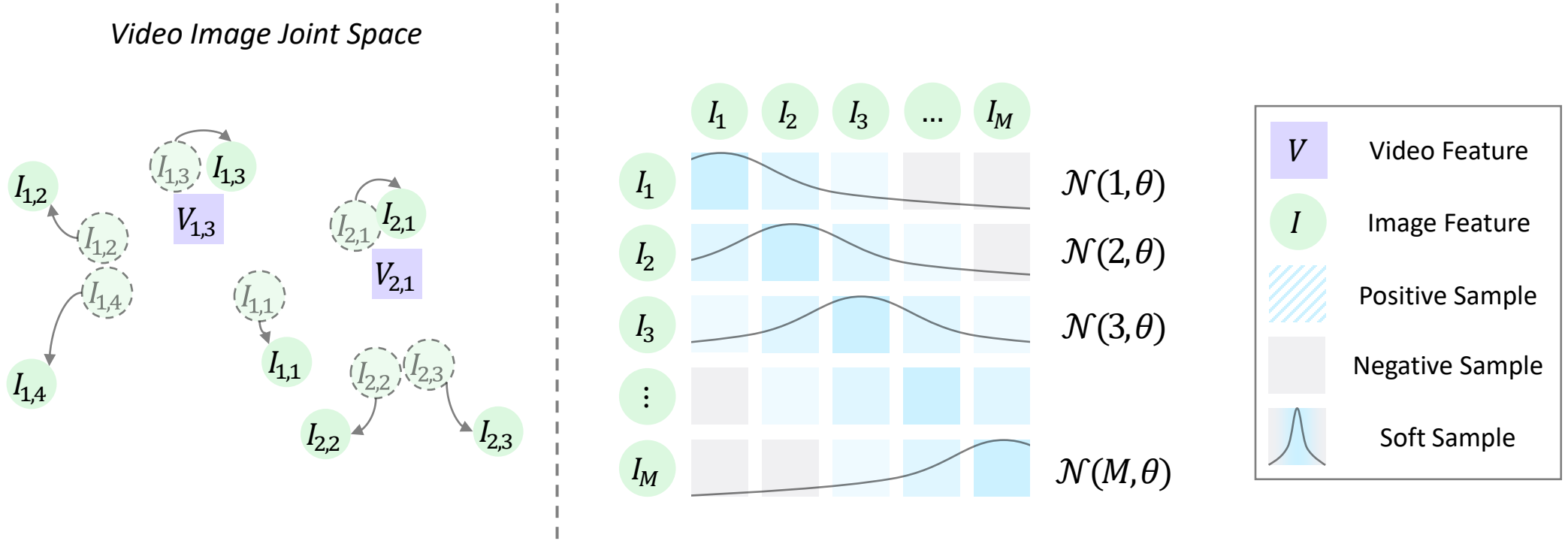


Figure 13. Visualization of loss c: Intra-Manual Contrastive Loss.

Characteristics:

- Enhance the difference between steps within a manual.
- Use a gaussian distribution to model the difference.
- Use the JS divergence loss.

Method – Intra-Manual Contrastive Loss

Table 3. Performance of difference loss combinations.

Loss	Video to diagram retrieval				Diagram to video retrieval					
	Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	Step	Page	Step	Page	Step	Page	Step	Page	Step	Page
VM	28.20	**34.59**	3.789	2.991	21.02	**16.64**	44.43	31.93	**0.618**	**0.393**
VM+IM	**28.62**	34.55	**3.734**	**2.928**	**22.30**	16.48	**45.00**	**32.20**	**0.618**	0.390

Method – Sinusoidal Progress Rate Feature (SPRF)

- There is a positive correlation between the progress of video and step index.
- Start time t_{start} end time t_{end} and video duration $t_{duration}$.
- j-th step from a manual with M steps.

$$r^V = \frac{t_{start} + t_{end}}{2t_{duration}}$$

$$r^I = \frac{j}{M}$$

$$\text{SPRF} = (\sin(\pi r^V), \cos(\pi r^I))$$

Table 4. Performance of difference Progress Rate Features. **PE**: Positional Embedding. **SPRF After**: SPRF locates after the final linear layer and before the loss.

Method	Video to diagram retrieval				Diagram to video retrieval					
	Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	Step	Page	Step	Page	Step	Page	Step	Page	Step	Page
w/o	22.28	27.70	5.983	4.639	16.97	12.95	36.36	27.25	0.548	0.357
PE Add	19.10	25.72	4.317	3.248	14.49	12.71	35.04	26.93	0.544	0.356
PE 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

Method – Optimal Transport (OT) for Post-Processing

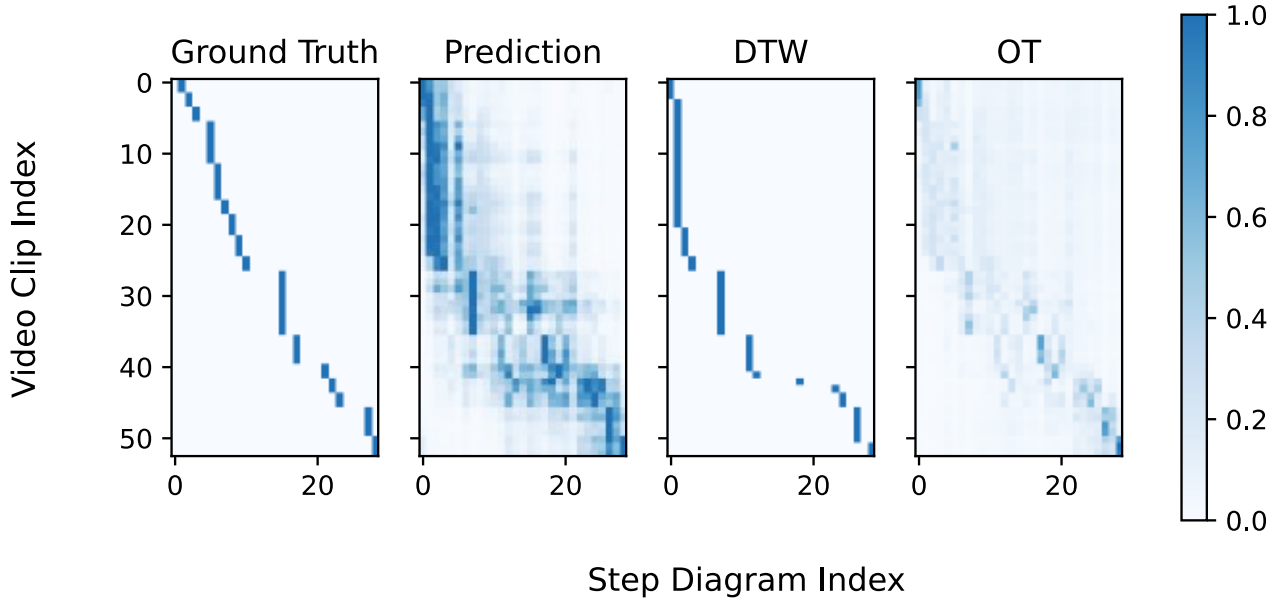


Figure 14. An example of post processing with DTW and OT with furniture 30341149 and video dzLNgz861Hk.

- Cost Matrix C:

$$s_{ij} = f_{sim}(\mathbf{f}_i^V, \mathbf{f}_j^I)$$
$$\bar{s} = \max_{ij} s_{ij}$$
$$\underline{s} = \min_{ij} s_{ij}$$
$$C_{ij} = \frac{s_{ij}^{\alpha} - \underline{s}^{\alpha}}{\bar{s}^{\alpha} - \underline{s}^{\alpha}}$$

- The normalization by $\bar{s}^{\alpha} - \underline{s}^{\alpha}$ restricts the range of C_{ij} to $[0,1]$, and $\alpha > 1$ accentuates the similarity differences.
- The optimal transport plan $T^{\star}$ is obtained by solving the entropy regularized OT problem,

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

- Sinkhorn-Knopp algorithm.

Method – Optimal Transport (OT) for Post-Processing

Table 5. Performance of difference post-process methods.

Method	Video to diagram retrieval				Diagram to video retrieval					
	Top1 Acc.%↑		AIE↓		R@1↑		R@3↑		AUROC↑	
	Step	Page	Step	Page	Step	Page	Step	Page	Step	Page
w/o	28.62	34.55	3.734	2.928	22.30	16.48	45.00	32.20	0.617	0.390
DTW	31.45	36.20	**3.382**	**2.752**	23.20	17.32	32.45	17.55	0.467	0.310
OT	**31.61**	**36.71**	3.458	2.816	**26.62**	**18.28**	**49.11**	**32.28**	**0.626**	**0.401**

Takeaways

- A large Ikea furniture **dataset (IAW)** contains ground truth alignment between assembly YouTube videos and corresponding manuals.

Takeaways

- A large Ikea furniture **dataset (IAW)** contains ground truth alignment between assembly YouTube videos and corresponding manuals.
- A **brand-new task** setting that aligns video demonstrations with instructional diagrams.

Takeaways

- A large Ikea furniture **dataset (IAW)** contains ground truth alignment between assembly YouTube videos and corresponding manuals.
- A **brand-new task** setting that aligns video demonstrations with instructional diagrams.
- Three modified contrastive losses specifically designed for this alignment task.

Takeaways

- A large Ikea furniture **dataset (IAW)** contains ground truth alignment between assembly YouTube videos and corresponding manuals.
- A **brand-new task** setting that aligns video demonstrations with instructional diagrams.
- Three modified contrastive losses specifically designed for this alignment task.
- Optimal Transport can be used as a post-process for the alignment task.

Thanks!

For more details:

https://academic.davidz.cn/en/publication/zhang-cvpr-2023/

Project